Enhanced Privacy and Efficiency in Machine Learning Through Innovative Paradigms

WAN, FAN

Enhanced Privacy and Efficiency in Machine Learning Through Innovative Paradigms

Fan Wan

Department of Computer Science
Durham University

This dissertation is submitted for the degree of
Doctor of Philosophy

Ustinov College October 2024

I would like to dedicate this thesis to my loving family and myself for striving forward.

Declaration

The research presented in this thesis was conducted at the Department of Computer Science, Durham University, United Kingdom. This thesis is original and has not been submitted for any other degree or qualification. Except where explicitly acknowledged in the text, the work is entirely my own.

Note on publications Included in This Thesis: As of the submission date, four chapters of this thesis significantly draw upon papers that have been submitted to or published in academic conferences and journals:

- ***Fan Wan****, Rui Sun, Haoran Duan, Xueqi Qiu, Xingyu Miao, Yang Long.* "Asynchronous Personalized Federated Learning through Global Memorization", submitted to IEEE Transactions on Image Processing (TIP), 2024. (Chapter 3)

- ***Fan Wan****, Junyan Wang, Haoran Duan, Yang Song, Maurice Pagnucco, Yang Long.* "Community-Aware Federated Video Summarization", published in IEEE International Joint Conference on Neural Networks (IJCNN), 2023 (Chapter 4).

- *Rui Gao***,* ***Fan Wan*****, Daniel Organisciak, Jiyao Pu, Haoran Duan, Peng Zhang, Xingsong Hou, Yang Long.* "Privacy-Enhanced Zero-Shot Learning via Data-Free Knowledge Transfer", published in IEEE International Conference on Multimedia & Expo(ICME), 2023 (Chapter 5).

- ***Fan Wan****, Xingyu Miao, Haoran Duan, Jingjing Deng, Rui Gao, Yang Long.* "Sentinel-Guided Zero-Shot Learning: A Collaborative Paradigm without Real Data Exposure", published in IEEE Transactions on Circuits and Systems for Video Technology(TCSVT), 2024 (Chapter 6).

Note on Publications Not Included in This Thesis: In addition to the papers mentioned, I have published several other works during the research phase of this thesis. These publications have enriched my understanding and contributed significantly to the foundational knowledge underpinning this thesis. Despite their value, these publications do not fit into the narrative of this thesis and have not been included in the text.

- *Yuchen Li**, ***Fan Wan****, *Yang Long.* Sid-NeRF: Few-shot NeRF Based On Scene Information Distribution, published in IEEE International Conference on Multimedia & Expo(ICME), 2024.
- *Xingyu Miao, Yang Bai, Haoran Duan, Yawen Huang,* ***Fan Wan****, Xinxing Xu, Yang Long* DS-Depth: Dynamic and Static Depth Estimation via a Fusion Cost Volume, published in IEEE Transactions on Circuits and Systems for Video Technology(TCSVT), 2024.
- *Xingyu Miao, Yang Bai, Haoran Duan,* ***Fan Wan****, Yawen Huang, Yang Long, Yefeng Zheng.* "CTNeRF: Cross-time Transformer for dynamic neural radiance field from monocular video", published in Pattern Recognition, 2024.

Fan Wan
October 2024

Acknowledgements

I've always considered myself incredibly fortunate. Though I grew up in modest circumstances, I never felt like I was missing out on anything, thanks to the unwavering support of my family. While my parents were away building a life for us through hard work, I spent fourteen important years with my grandfather, Jiegen Wan. His wisdom, shaped by the rare opportunity to attend a private school during World War II, left a deep impression on me. He cherished education, and it was through his patient teaching that I learned to appreciate it too. He, along with my parents, set me on the path that has led me here today.

Along this journey, I've been lucky enough to have not only a supportive family but also incredible mentors and friends who've guided me along the way.

First and foremost, I owe a huge thank you to my supervisor, Professor Yang Long. His steady guidance gave me the freedom to explore the fascinating world of computer vision while encouraging me to think deeply about the challenges I faced. His advice helped me through the highs and lows of both the research process and my personal journey through this PhD.

I also want to thank Professor Jingjing Deng, whose constant support and contributions made my daily research life so much richer. His dedication was invaluable to me.

I'm also incredibly grateful to the university that's been my academic home and to all the colleagues who've accompanied me along the way. Senior members like Junyan Wang, Haoran Duan, Yang Bai, and Peng Zhang offered me so much support. Working with Rui Gao, Xinyu Miao, Xueqi Qiu, Yuchen Li, Leyuan Zhang, and Tianyu Zhang was a real pleasure. I also want to extend my thanks to friends like Jiyao Pu and Minye Shao for their friendship and encouragement during this journey.

Of course, my family has been my foundation through it all. My parents, Zhusheng Wan and Jieyun Chen, worked tirelessly to give me the opportunities they didn't have. Even though they didn't receive much formal education themselves, they always believed in the power of learning. Their sacrifices allowed me to study abroad and eventually pursue a PhD. My sister, Fang Wan, has been a constant source of encouragement, always standing by me through every step. And my wife, Meng Li, has been my greatest support. Her selfless care

for our son, Leo Wan (Zhijin Wan), gave me the space and time to focus on my work. None of this would have been possible without her love and unwavering support.

I'm also deeply thankful for the friends I made abroad. My neighbors in Newcastle, Andrew and Claire Wilkin, have been a source of steadfast support since 2017. Their kindness has been a light in my life during my time in the UK. I also want to thank my English teacher, Catherine Strydom from South Africa, whose patience and dedication helped me become more confident in expressing my ideas. I wish her all the happiness and success she deserves.

This isn't the end—it's just the beginning. There are more challenges ahead, but I look forward to facing them, knowing they're just part of the journey.

Abstract

The rapid evolution of communication technology and the widespread use of Internet of Things (IoT) devices have led to an unprecedented increase in data generation. This surge of data, from sources like smartphones, sensors, and networks, drives innovation across multiple sectors, from healthcare to urban planning. However, this rapid growth also introduces significant challenges, particularly in ensuring privacy as personal information is increasingly collected and shared across digital platforms. Balancing privacy protection with effective data utilization has become a key issue in modern machine learning applications.

This thesis addresses these challenges by exploring the potential of Federated Learning (FL) and Zero-Shot Learning (ZSL) as solutions for privacy-preserving data use. Although promising, these techniques still face gaps in safeguarding user privacy while maximizing data utility. The research presented here aims to bridge this gap, developing new methodologies that protect privacy while enabling efficient exploitation of large datasets.

To address the issue of statistical and system heterogeneity in FL, the thesis introduces an Asynchronous Personalized Federated Learning framework (AP-FL), which incorporates model interpolation and a data-free knowledge transfer method to enhance robustness and efficiency. In the context of Video Summarization, it proposes a frame-based aggregation method and a Community-Aware Clustering Federated Framework (CFed-VS), designed to address privacy concerns and manage the complexity of video data.

Further, the research explores Privacy-Enhanced Zero-Shot Learning (PE-ZSL) and Sentinel-Guided Zero-Shot Learning (SG-ZSL), which offer novel approaches for zero-shot classification without direct access to real data. These frameworks protect sensitive data while ensuring effective knowledge transfer, marking significant advancements in secure AI learning environments.

Through these contributions, this thesis advances the state of machine learning by addressing key issues related to data privacy, heterogeneity, and efficiency. The findings presented here not only improve the robustness of FL and ZSL frameworks but also pave the way for future research into privacy-preserving AI technologies.

Table of contents

List of figures

List of tables

List of symbols

AI Artificial Intelligence

CFL Clustered-Based Federated Learning

$DFKD$ Data-Free Knowledge Distillation

DP Differential Privacy

FL Federated Learning

$GZSL$ Generalized Zero-Shot Learning

IID Independently and Identically Distributed

IoT the Internet of Things

KD Knowledge Distillation

$LLMs$ Large Language Models

ML Machine Learning

$Non-IID$ Non-Independent and Identically Distributed

PFL Personalized Federated Learning

VS Video Summarization

ZSL Zero-Shot Learning

Chapter 1

Introduction

1.1 Background

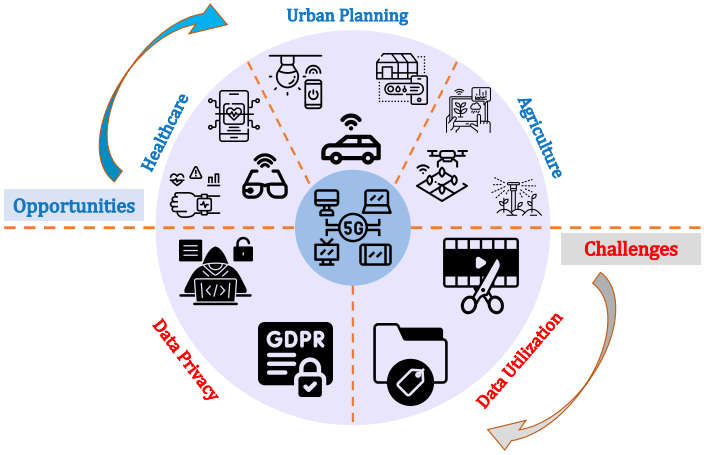


Fig. 1.1 5G, IoT, and AI: Navigating Opportunities and Challenges in the Data Revolution

The advent of communication technology and the widespread adoption of the Internet of Things (IoT) devices have precipitated an unprecedented surge in data generation from diverse sources such as smartphones, sensors, and interconnected devices. This digital proliferation is not just a statistic; it fuels transformations across industries as shown in Fig 1.1: from healthcare, where wearable devices monitor patient health in real-time, to urban planning, where IoT sensors enable smart cities to optimize traffic flow and energy consumption. Furthermore, in the agricultural sector, data collected from IoT devices aids in precision farming techniques, significantly increasing crop yields while conserving resources. Such real-world applications underscore the potential of these technologies to revolutionize how we live, work, and interact with our environment. However, alongside these opportunities,

the burgeoning data landscape presents challenges in processing and analyzing vast volumes of information, and notably, significant concerns regarding the safeguarding of individual privacy within this data-driven revolution.

Data Explosion and Privacy Concerns: The anticipated exponential growth in data volume, with projections reaching 175 zettabytes by 2025 [2], amplifies the critical urgency to address privacy concerns. This vast amount of data, often generated from personal devices and shared across multiple platforms, increases the risk of data breaches and unintended exposure. For instance, data can be exposed through unsecured networks, inadequate encryption, or through attacks targeting centralized servers where data is aggregated. The intertwinement of personal identity with data accentuates the demand for robust privacy protection mechanisms. Legislative measures such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have established new benchmarks in data privacy, imposing rigorous protocols for data handling and protection. These regulations highlight the primacy of individual rights over their data and challenge researchers and developers to pioneer privacy-preserving data processing methodologies.

The Intersection of Machine Learning and Data Privacy: The dilemma of data privacy is intimately connected with the advancements in machine learning (ML). Conventional ML methodologies, which often depend on the aggregation of data in centralized repositories for model training, inherently risk user privacy and data security. This centralization of sensitive data renders it an attractive target for malicious actors, thereby amplifying concerns regarding data breaches and unauthorized access.

In response, McMahan et al. [3] proposed Federated Learning (FL), a powerful paradigm for privacy-preserving machine learning. FL decentralizes the training process by allowing each participating device (client) to train its model locally, ensuring that raw data never leaves the device. Instead of sharing the data, clients periodically send their model updates to a central server, which aggregates these updates to create a global model. This iterative process continues over several rounds, enabling the global model to improve without centralizing any sensitive data. By keeping data on the clients' devices and only sharing model parameters, FL significantly reduces the privacy risks associated with traditional centralized data storage, while still enabling collaborative learning across distributed networks.

Despite the advantages of privacy enhancement, FL is confronted with challenges such as communication overhead and statistical heterogeneity among distributed datasets, which may compromise model performance and convergence. For example, in a typical FL setting, the need to frequently communicate model updates between the central server and distributed devices can result in high communication costs, particularly in environments with limited

bandwidth. Additionally, statistical heterogeneity refers to the variation in data distribution across different devices; for instance, in a healthcare scenario, patient data collected from different hospitals may vary greatly due to demographic differences, which makes it difficult for the global model to generalize effectively, potentially leading to slower convergence and reduced model accuracy.

The Imperative of Data Utilization: The effective utilization of data is crucial for fostering innovation and extracting actionable insights. However, the task of annotating large datasets for supervised learning constitutes a significant bottleneck, demanding extensive human effort and time. Within this context, Zero-Shot Learning (ZSL) [4] emerges as a promising solution, enabling models to generalize to unseen classes without the need for exhaustive annotation. By exploiting semantic relationships between known and unknown classes, ZSL facilitates the categorization of new instances without requiring additional labeling. Despite its potential, ZSL faces challenges such as domain shift and constrained generalization capabilities in complex real-world scenarios. Importantly, ZSL approaches frequently overlook privacy considerations, potentially exposing sensitive data during model training and deployment.

As data continues to grow at an unprecedented rate, particularly with the proliferation of multimedia content, effectively extracting insights from such data becomes increasingly challenging. Video content, in particular, stands out due to its sheer volume and complexity. Videos are now a primary medium for communication and information sharing across sectors such as entertainment, education, and surveillance, contributing to the exponential growth of unstructured data. Unlike text or static images, video data presents additional challenges due to its temporal dimension, high information density, and variation in format and length. Processing and analyzing this type of data requires novel approaches that can handle its distinct characteristics efficiently.

In this context, Video Summarization (VS) [5] emerges as a critical tool to address these challenges by distilling lengthy videos into concise summaries that capture the core information. By doing so, VS helps streamline information retrieval, accelerate decision-making processes, and enhance user engagement by reducing the cognitive load of sifting through large volumes of video content. However, despite the clear benefits, video summarization technologies face significant obstacles, including accurately capturing the essence of complex content, ensuring scalability across different video types, and addressing generalization challenges.

Moreover, privacy concerns often take a back seat in video summarization research, despite the sensitive nature of many video datasets. For instance, videos used in healthcare, surveillance, or personal communications often contain identifiable personal information or confidential content. The process of summarizing such videos could inadvertently expose

sensitive details if privacy-preserving measures are not in place. This increases the risk of unauthorized access, data breaches, or misuse during data processing and sharing. Therefore, it is crucial to develop video summarization methodologies that not only effectively condense video content but also incorporate robust privacy-preserving protocols, ensuring that sensitive information is protected throughout the process.

1.2 Motivation

In the modern digital landscape, the vast amounts of data generated from IoT devices and other technologies offer immense potential for machine learning applications across various sectors. However, alongside this potential comes significant challenges, particularly in safeguarding user privacy and efficiently utilizing these large datasets. Despite the promising approaches offered by FL, ZSL, and VS, they still fall short in addressing both privacy concerns and data efficiency simultaneously. This gap underscores the need for a secure and efficient paradigm that tackles these challenges head-on. The motivation for this research is to develop methodologies that protect user privacy while maximizing data utility, pushing the boundaries of current ML capabilities.

1.3 Research Challenges and Objectives

Research Challenge 1: Statistical and System Heterogeneity in Federated Learning.

Challenge Overview: Federated learning has emerged as a transformative solution for enabling machine learning across distributed data sources while preserving user privacy. However, the foundational assumptions of FL often clash with real-world complexities. FL methodologies traditionally assume that data from different devices are uniform, or independently and identically distributed (IID). In practice, however, the data across devices is often heterogeneous due to user preferences, geographical diversity, and other socio-economic factors, leading to statistical heterogeneity. This discrepancy creates challenges for model training, as the aggregated global model may struggle to capture the true underlying distribution, reducing both its accuracy and its ability to generalize across diverse data environments.

In addition to statistical challenges, FL systems face significant system heterogeneity, where differences in device capabilities, network connectivity, and operational conditions create further complications. A key manifestation of system heterogeneity is the frequent oc-

currence of client dropouts. Due to varying network conditions, device failures, or differences in available computational resources, participant devices in an FL system may intermittently drop out during the training process. These dropouts not only reduce the volume of data available for training but can also result in certain data categories being underrepresented or missing entirely in the global model. This can lead to biased models that perform suboptimally in scenarios requiring comprehensive coverage of all client data, thereby limiting the generalization and fairness of the resulting models across intended applications.

Expanded Technological Innovation and Application Scenarios: Consider a global health monitoring system where wearable devices across different geographic regions collect health measurements. The diversity in user demographics and environmental factors results in highly non-IID data distributions. An advanced FL mechanism could dynamically adjust its training strategy based on localized data characteristics, significantly improving predictive accuracy for disease outbreaks or health trends. Similarly, in a decentralized social media analysis tool, user-generated content's variability necessitates robust models capable of understanding nuanced user engagements across different cultures and regions. Addressing client dropouts is critical in such applications to ensure comprehensive data analysis, especially during crucial events like elections or natural disasters, where real-time, reliable insights are indispensable.

Research Objective 1: Enhance Framework Effectiveness and Robustness with Federated Learning.

Address the critical challenges posed by non-IID data distributions and frequent client dropouts, aiming to fortify the FL framework's robustness and operational efficiency. Innovating FL with advanced mechanisms will mitigate these issues, enhance privacy preservation, and build a more resilient and efficient federated learning model. We aim to operationalize this objective by addressing the following Research Questions (RQ):
RQ 1.1: ***How can we refine federated learning architectures to manage non-IID data distributions and improve the overall model robustness and learning efficiency?***

- To address this, we will develop novel strategies and algorithms that can accurately represent and integrate non-IID data distributions during FL model training.
- Additionally, we will adapt the model aggregation process to reduce biases caused by imbalanced data distributions, ensuring equitable learning across diverse client datasets.

RQ 1.2: ***What innovative solutions can minimize the detrimental effects of client dropouts on the federated learning process, ensuring continuity and completeness of learning?***

- To mitigate the impact of client dropouts, we will implement mechanisms that compensate for data or category loss in FL environments.
- Furthermore, we will devise strategies that allow dropped clients to rejoin and participate effectively in subsequent training rounds.

Research Challenge 2: Security and Privacy in Video Summarization

Challenge Overview: Extracting meaningful information from vast datasets is crucial for the advancement of AI technologies. Among various types of data, video content stands out due to its richness and complexity, making it an invaluable resource for many applications. However, video data, with its temporal dimension and large file sizes, poses significant challenges for efficient processing and privacy protection.

To address these challenges, video summarization offers a practical solution by distilling lengthy videos into concise summaries that retain the essential information. This approach not only reduces the computational burden but also facilitates faster information retrieval and decision-making. Despite its potential, the integration of robust privacy-preserving mechanisms in video summarization is still lacking, especially when handling sensitive video content such as surveillance footage or personal recordings.

Expanded Technological Innovation and Application Scenarios: Imagine a scenario where city-wide surveillance systems aim to enhance public safety while protecting individual privacy. Advanced video summarization techniques (*e.g.*, privacy-preserving or decentralized VS methods) could generate concise reports of unusual activities or gatherings, ensuring that sensitive information is processed securely and only relevant data is transmitted. This approach not only preserves privacy but also reduces bandwidth usage significantly. Another application is in personalized content delivery services, where platforms can offer tailored video summaries based on user interests directly on their devices, minimizing the exposure of viewing habits to external servers.

Research Objective 2: Addressing Privacy and Efficiency in Video Summarization

This objective focuses on developing advanced methodologies for video summarization that not only protect user privacy but also effectively handle the inherent complexity and size of video data. We aim to address the current research gap in secure and efficient video

processing by crafting solutions that can balance the need for privacy with the practical requirements of managing large, information-rich video content. To achieve this objective, the following research questions are proposed:
RQ 2.1: ***How can video summarization techniques be adapted to preserve data privacy while capitalizing on the value of video content?***

- To address this, we will develop novel methods that enable the local processing of video data in FL frameworks, minimizing privacy risks while efficiently summarizing complex video content.
- Additionally, we will design algorithms that handle the unique temporal and structural characteristics of video data, ensuring effective model performance and privacy preservation across diverse video datasets.

RQ 2.2: ***How can novel Federated Learning approaches be designed to address both privacy protection and computational efficiency in video summarization tasks, considering the unique challenges of video data?***

- To address this, we will conduct an in-depth investigation to ensure that the integration of FL and VS adheres to the fundamental principles of FL, enabling secure and efficient processing of video data.
- Additionally, we will leverage cluster-based FL to manage the high heterogeneity of video data, capitalizing on the community-like structure of video datasets to improve model accuracy and performance while preserving privacy across diverse video sources.

By addressing these questions, we aim to develop practical solutions for integrating FL into VS, effectively tackling challenges related to data privacy and security. This work will enhance the use of video content across various applications, contributing to advancements in privacy-preserving machine learning technologies.

Research Challenge 3: Data Annotation, Privacy, and Copyright Protection.

Challenge Overview: In today's era of data explosion, AI technologies rely on vast datasets to fuel innovation. However, the manual process of data annotation remains a significant bottleneck. ZSL provides a promising solution by allowing models to recognize unseen classes without requiring extensive labeled datasets. While this technique reduces the need for data annotation, it often overlooks critical security and privacy concerns. Specifically,

ZSL relies on knowledge transfer between seen and unseen classes, which could inadvertently expose sensitive user data or violate copyright protections. Combining ZSL with FL offers a promising approach to addressing these concerns by enabling decentralized model training without centralizing sensitive data. However, FL itself faces challenges such as model misuse, potential copyright infringement, and ensuring equitable data contribution. Therefore, there is a pressing need for a unified framework that addresses data privacy, security, and copyright protection within AI development.

Expanded Technological Innovation and Application Scenarios: Consider the development of an AI-driven content creation tool that automatically generates articles or videos on emerging topics. In such a scenario, ZSL could enable the AI to understand and create content on subjects not present in its training data, drastically reducing the need for constant retraining and data annotation. Addressing privacy and copyright in this context, especially when dealing with sensitive information or creative content, necessitates a framework that allows for the secure transfer of knowledge without direct data sharing. Such a framework would enhance the efficiency of content creation while minimizing risks related to privacy breaches and copyright issues.

Research Objective 3: Innovate Zero-Shot Learning for Privacy and Performance.

The goal of this research is to develop a comprehensive framework that enhances the privacy, security, and efficiency of ZSL, while ensuring copyright protection and minimizing the reliance on data annotation. This will involve the secure transfer of knowledge, collaborative model training, and the integration of privacy-preserving techniques. To achieve this, we propose the following research questions and corresponding steps to address them.
RQ 3.1: ***How can innovative mechanisms enhance the utility and confidentiality of data in machine learning, facilitating the effective recognition of seen and unseen classes without compromising data privacy?***
This research question is addressed by the following steps:

- **Design privacy-preserving methods:** Investigate mechanisms such as differential privacy and secure multiparty computation that allow for the protection of sensitive data during model training, without exposing actual user data.
- **Optimize knowledge transfer in ZSL:** Develop techniques that enable effective knowledge transfer between seen and unseen classes, ensuring that models can generalize to new categories without sharing raw data or violating data privacy.

- **Establish a balance between privacy and performance:** Explore trade-off strategies that allow flexible control over privacy costs while maintaining model accuracy, ensuring that privacy enhancements do not significantly degrade model utility.

RQ 3.2: ***What are the effective approaches for utilizing semantic information in zero-shot learning for knowledge transfer, and how can the trade-off between privacy and performance be optimized?***
This research question is addressed by the following steps:

- **Identify optimal semantic representations:** Research which types of semantic information (e.g., attribute-based, textual descriptions) are most effective for knowledge transfer in ZSL, and evaluate their impact on model performance and privacy.
- **Address bias in ZSL:** Investigate strategies to reduce the inherent bias towards seen classes in traditional ZSL methods, ensuring that the model can fairly recognize both seen and unseen categories without over-relying on the seen class data.
- **Integrate privacy-preserving techniques into training:** Incorporate privacy-preserving algorithms into the ZSL training process to safeguard both data privacy and intellectual property, ensuring secure model deployment in sensitive domains.

By addressing these questions, our research aims to strike a balance between extensive data utilization and the critical requirements of data privacy, copyright protection, and ethical AI use. Through this work, we seek to improve the development and deployment of AI models, ensuring that innovation is aligned with security and respect for intellectual property.

1.4 Main Contributions

This thesis presents a suite of innovative approaches to addressing contemporary challenges in federated learning, video summarization, and zero-shot learning, areas critical to the advancement of machine learning and artificial intelligence. The contributions are systematically outlined below, reflecting the depth and breadth of the research conducted.

- **Innovative Framework for Non-IID Data and Dropout Mitigation:** We introduce an Asynchronous Personalized Federated Learning Framework (AP-FL), a pioneering solution that addresses both statistical and system heterogeneity in federated learning. To manage non-IID data distributions, AP-FL incorporates model interpolation techniques that enhance the robustness and generalizability of models across diverse data sources. Additionally, AP-FL tackles client dropouts by leveraging a data-free

knowledge transfer approach inspired by ZSL, which generates synthetic samples to compensate for missing data. This dual mechanism ensures continuous learning and high model performance, even in the presence of significant data absenteeism and distributional challenges.

- **Frame-Based Aggregation for Video Tasks:** We propose a novel frame-based FedAvg aggregation method tailored for VS tasks. This method systematically considers the video length in model contributions, enhancing the accuracy and relevance of summarized content. This approach addresses the unique challenges of video data in federated learning, ensuring that the summaries generated are both comprehensive and contextually rich.

- **Clustering Federated Framework for Heterogeneous Data:** The introduction of a Community-Aware Clustering Federated Framework (CFed-VS) marks a significant advancement in handling data heterogeneity within video summarization. By clustering clients based on data distribution similarity, CFed-VS efficiently manages diverse video content, reducing computational costs and improving the global model's training efficiency.

- **Mixture Transformer for Enhanced Model Generalization:** With the Mixture Transformer, we offer an innovative solution to improve model generalization in non-IID settings. This development ensures that federated learning models can effectively learn from time-series data, demonstrating superior performance on the SumMe[6] and TVSum[7] datasets.

- **Privacy-Enhanced Zero-Shot Learning (PE-ZSL) for Data Copyright and Sensitivity Protection:** We propose a novel PE-ZSL framework that enables zero-shot classification without exposing real data, addressing the critical need for data copyright and sensitivity protection. This data-free knowledge transfer framework ensures privacy during model training, offering robust solutions for AI applications in sensitive domains, without sharing any real data.

- **Sentinel-Guided Zero-Shot Learning (SG-ZSL) for Enhanced Privacy and Knowledge Transfer:** Building on PE-ZSL, SG-ZSL introduces a more refined approach to model training and knowledge transfer by incorporating a detailed comparison of traditional ZSL methods—transductive and inductive. This framework leverages omniscient and quasi-omniscient teacher models, enhanced with differential privacy techniques, ensuring secure learning without exposing sensitive data. SG-ZSL offers

a deeper analysis of privacy-preserving learning processes while maintaining model effectiveness, providing a flexible and secure solution for diverse AI applications.

Each of these contributions addresses challenges related to data heterogeneity, privacy, and efficient learning, offering advancements in federated learning, video summarization, and zero-shot learning. The methodologies proposed provide practical insights that can be applied across various fields, from healthcare to content creation, and contribute to the ongoing development of privacy-preserving AI technologies.

1.5 Thesis Structure

This thesis is organized into the following chapters, each contributing significantly to the overarching goal of advancing machine learning methodologies in the context of privacy preservation, data efficiency:

- **Chapter 1** provides an overview of the research background, motivation, research questions, and corresponding objectives. It delineates the main contributions of this thesis and outlines its structural composition.
- **Chapter 2** delves into a comprehensive review of the relevant literature in the domains of Federated Learning, Video Summarization, Zero-Shot Learning, and Knowledge Distillation.
- **Chapter 3** details the development and evaluation of the AP-FL framework. It addresses the challenges of statistical and systems heterogeneity inherent in federated learning environments, particularly focusing on non-IID data distributions and client dropouts. This chapter presents a novel personalized learning approach leveraging model interpolation and introduces a data-free knowledge transfer mechanism, culminating in state-of-the-art performance on standard benchmarks like CIFAR10 [8], CIFAR100 [8], EMNIST [9], and Fashion MNIST [10].
- **Chapter 4** explores the novel domain of the CFed-VS, tackling the dual challenges of data privacy and heterogeneity in video data. It proposes a frame-based aggregation method of FedAvg tailored for video-related tasks, a novel clustering federated framework to handle heterogeneous data efficiently, and a Mixture Transformer model to enhance generalization in non-IID settings, demonstrating superior performance on SumMe[6] and TVSum[7] datasets.

- **Chapter 5** introduces the concept of the PE-ZSL via a data-free knowledge transfer framework. It addresses the critical need for privacy-preserving zero-shot learning, proposing innovative solutions for data copyright protection and sensitivity elimination without real data sharing. This chapter also elaborates on 'black-box' and 'white-box' scenarios for model sharing, alongside an analysis of teacher models in both omniscient and quasi-omniscient settings, showcasing promising results in both conventional and generalized ZSL tasks.

- **Chapter 6** presents SG-ZSL, a collaborative paradigm that eschews real data exposure. SG-ZSL represents a significant leap in zero-shot learning, addressing key concerns of data privacy and model copyright through a novel collaborative training framework.

- **Chapter 7** presents a comprehensive summary of the key findings and contributions of the thesis. It reflects on the practical implications of the proposed methodologies for privacy-preserving machine learning and evaluates their potential applications. Furthermore, the chapter discusses limitations encountered in the research, such as computational costs and real-world scalability, and suggests future research directions to address these challenges and improve the presented frameworks.

Chapter 2

Background

In the digital age, the importance of data security and the efficient utilization of large datasets is paramount in machine learning. This chapter explores key developments in the field, with a focus on foundational work in machine learning and privacy-preserving techniques. We review Federated Learning as a significant advancement for decentralized, privacy-preserving model training, while also examining Video Summarization and Zero-Shot Learning as essential methods for improving data efficiency and utility. Additionally, Knowledge Distillation is discussed for its role in enhancing model transferability and learning efficiency. These topics provide a comprehensive background for the challenges and solutions addressed in later sections, highlighting the intersection of data privacy and machine learning advancements.

2.1 Machine Learning

The evolution of machine learning (ML) from its inception to its current rapid advancement offers a fascinating look into the broader development of artificial intelligence (AI). At its core, ML is based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention [11]. The foundational principles of ML are drawn from multiple disciplines, including statistics, computer science, and information theory [12], reflecting a multidisciplinary approach to solving complex problems.

Early milestones in ML were characterized by the development of basic algorithms such as decision trees and linear regression, which provided the foundation for more advanced models [13]. Rosenblatt's perceptron [14], introduced in the late 1950s, was one of the earliest neural network models and hinted at the potential for systems to 'learn' from their environment.

The resurgence of neural networks, particularly in the form of deep learning, has been a driving force behind recent advances in ML. The ability of deep learning algorithms to learn high-level abstractions from data, combined with increasing computational power and the availability of large datasets, has enabled breakthroughs in areas such as image recognition, natural language processing, and autonomous systems [15]. Key innovations, such as the backpropagation algorithm [16] and the development of convolutional neural networks [17, 18], have played a pivotal role in pushing the field forward.

However, with these advances come challenges, particularly around model interpretability. Traditional ML models, like decision trees or linear regression, are generally more interpretable, offering clear relationships between inputs and outputs. In contrast, deep learning models, such as deep neural networks, are often viewed as "black boxes" due to their complex architectures, making it difficult to understand how they arrive at decisions. This lack of interpretability is especially problematic in critical applications such as healthcare and finance, where understanding a model's decision-making process is vital for ensuring trust and accountability [19].

Moreover, as machine learning models become increasingly integrated into societal infrastructure, additional issues such as data privacy and the ethical implications of automated decision-making also arise. A particularly pressing challenge is data privacy, especially with the advent of large language models (LLMs) [20, 21], which raise concerns over their training on potentially sensitive information. These challenges necessitate innovative solutions like federated learning and differential privacy, which offer ways to train powerful models without compromising individual privacy.

2.2 Data Privacy and Security in Machine Learning

Data privacy and security have become critical concerns in the field of machine learning and artificial intelligence. As ML models grow more capable of extracting insights from vast datasets, the need to protect sensitive information becomes increasingly important. This is not just an academic concern; it reflects a societal demand for technologies that respect privacy and ensure data security.

A key issue is the dual role of data: while it drives the development of ML technologies, it also presents privacy risks if not handled properly. Traditional ML methods often rely on centralizing data from various sources, creating a single point of vulnerability that can lead to breaches and unauthorized access. A notable example is the 2017 Equifax data breach [22], which exposed the personal information of over 147 million individuals, illustrating the

risks associated with centralized data storage. Such breaches not only damage public trust but can also lead to serious legal and financial consequences.

To address these challenges, researchers are developing techniques to ensure privacy-preserving data use. One promising approach is Federated Learning [23, 24], which allows for collaborative model training without requiring the sharing of raw data. By keeping data on users' devices and only exchanging model updates, FL reduces the risk of data exposure. However, FL presents new challenges, such as securing model updates and protecting against inference attacks that could reveal sensitive information from shared gradients.

Advances in cryptographic techniques, such as homomorphic encryption [25] and secure multi-party computation [26], provide methods for processing data while keeping it confidential. These techniques allow data to be analyzed without revealing its contents. Despite their effectiveness, cryptographic methods often require significant computational resources, which can limit their use in environments with constrained resources.

Another important concept in data privacy is differential privacy [27], which quantifies privacy risks associated with including an individual's data in a dataset. By adding noise to data or model outputs, differential privacy ensures that the inclusion of a single data point cannot be easily detected. This method has been successfully applied in real-world systems, including those used by Google and Apple [28], showing its potential as a practical privacy-preserving solution.

As ML continues to evolve, balancing privacy, security, and data utility remains a key research focus. Emerging approaches like zero-shot learning [29] and knowledge distillation [30] offer ways to reduce dependence on sensitive data. For example, zero-shot learning generalizes to unseen categories without needing data from every class, and knowledge distillation transfers knowledge from a more complex model to a simpler one, reducing the risk of exposing detailed data patterns.

In summary, addressing data privacy and security in machine learning is an ongoing challenge. As new risks emerge, so do innovative solutions. The development of privacy-preserving techniques will be essential to ensure that AI technologies can be trusted to protect individual privacy.

2.3 Federated Learning

The emergence of Federated Learning stems from the need to leverage the power of machine learning while addressing the growing concerns around data privacy and security in the digital age. FL represents a significant shift from traditional centralized ML approaches, where data is aggregated in a central repository for model training, to a decentralized framework. In FL,

learning occurs directly on users' devices, allowing sensitive information to remain local and thereby significantly enhancing privacy.

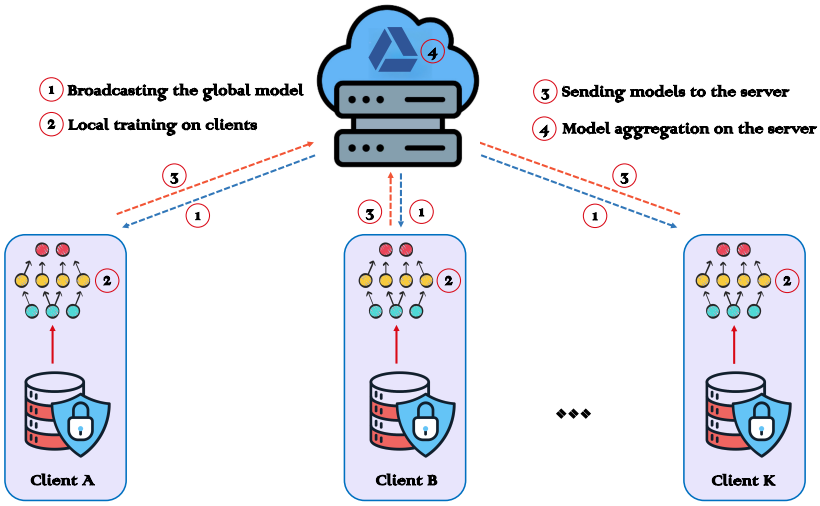


Fig. 2.1 Illustration of the training process involved in Federated Learning.

FL was popularized by McMahan et al. [3], who coined the term to describe a process in which a global model is collaboratively trained across many participating devices (or "clients") that keep their data locally. The typical FL process, as shown in Fig. 2.1, consists of the following main steps:

1. **Broadcasting the Global Model:** The central server broadcasts the initial global model w^t to all participating clients. Each client receives this model and uses it as the starting point for local training.

2. **Local Training on Clients:** Each participating client k trains a local model on its own dataset D_k, using the current global model w^t provided by the server. The client performs local updates using gradient descent as follows:

$$w_k^{t+1} = w_k^t - \eta \nabla \ell(w_k^t; D_k), \tag{2.1}$$

   where w_k^{t+1} represents the updated model parameters for client k after the $t+1$ round, η is the learning rate, and $\nabla \ell(w_k^t; D_k)$ is the gradient of the loss function ℓ with respect to the local data D_k on client k at time t.

3. **Sending Models to the Server:** After completing local training, each client sends its updated model parameters w_k^{t+1} to the central server. This step ensures that no raw data is shared—only the model updates are communicated, preserving privacy.

4. **Model Aggregation on the Server:** The server aggregates the updated models from all participating clients to form a new global model. A common aggregation method is a weighted average based on the size of each client's dataset:

$$w^{t+1} = \sum_{k=1}^{K} \frac{|D_k|}{\sum_{i=1}^{K} |D_i|} w_k^{t+1}, \tag{2.2}$$

where K is the number of participating clients, and $|D_k|$ is the size of client k's dataset.

Steps 1 through 4 are repeated over several rounds until the global model converges.

One of the key advantages of FL is its privacy-preserving nature. Since raw data never leaves the user's device, the risks associated with data breaches and unauthorized access are greatly reduced. This feature is especially valuable in privacy-sensitive industries such as healthcare and finance, where data protection is critical. In addition, FL is well-suited for environments with limited bandwidth, as it minimizes the need for extensive data transmission. Only model updates—much smaller in size than the raw data—are transmitted, making FL an efficient option in areas with restricted connectivity. This opens new possibilities for ML applications in regions where network resources are constrained. FL also addresses the challenge of fragmented data, where data is distributed across different organizations or jurisdictions, often restricted by privacy regulations. Instead of requiring data to be pooled into a central location, FL enables collective model training while keeping the data within its original source, thereby respecting privacy laws and organizational policies.

Despite its advantages, FL is not without challenges. Early implementations of FL encountered issues related to communication efficiency [3, 31], heterogeneity in client data and device capabilities [32, 33], and maintaining model performance in decentralized settings [34, 35]. Solutions such as structured updates and model compression techniques [31] have been developed to mitigate these challenges, allowing FL systems to evolve and improve.

The introduction of Federated Learning marks a pivotal moment in the development of privacy-preserving machine learning. By decentralizing data processing and keeping personal data on the user's device, FL not only enhances privacy but also unlocks the potential of distributed data sources, offering a promising path forward in machine learning as data privacy concerns continue to grow.

Personalized Federated Learning: Personalized Federated Learning (PFL) represents an important advancement within the federated learning framework, designed to address the challenges posed by the diverse nature of client data. While traditional FL approaches are effective in preserving privacy and utilizing distributed data, they often assume that a single global model can serve all clients equally. However, in real-world scenarios, this assumption is rarely valid due to the significant variation in data distributions across different devices or

users. This variation, known as statistical heterogeneity, can reduce the performance of a global model when applied to local contexts.

PFL addresses this challenge by shifting the focus from developing a single shared model to creating customized models tailored to individual clients [36] or groups of clients [24]. Specifically, each client k learns a personalized model w_k based on the global model w_g, but incorporates local data D_k to adjust the model. The local training process can be formulated as:

$$w_k^{t+1} = w_g^t - \eta \nabla \ell(w_g^t; D_k), \tag{2.3}$$

where w_k^{t+1} represents the updated local model for client k after the $t+1$ round, η is the learning rate, and $\nabla \ell(w_g^t; D_k)$ is the gradient of the loss function ℓ with respect to the global model w_g^t and local data D_k. This allows each client to adapt the global model based on its own data distribution.

The key advantage of PFL lies in its ability to adapt the learning process to each client's unique data characteristics, enhancing model performance and relevance at the local level. One commonly used method to personalize models is by introducing a mixture of global and local updates, such that:

$$w_k^{t+1} = \lambda w_g^{t+1} + (1-\lambda) w_k^t, \tag{2.4}$$

where λ controls the trade-off between the global model w_g^{t+1} and the client-specific model w_k^t. A higher λ encourages more reliance on the global model, while a lower λ emphasizes personalization through the local model.

By emphasizing personalization, PFL aims not only to improve user satisfaction and engagement but also to tackle the issue of model fairness [37], ensuring that minority or underrepresented groups are better served.

Various PFL approaches have been explored, ranging from techniques that customize the global model based on local data [38], to more complex frameworks such as multi-task learning [39], which allow for the development of personalized models for each client. These methods often include mechanisms to capture and leverage relationships between clients' data distributions, using techniques like meta-learning and transfer learning. By recognizing and utilizing these relationships, PFL can balance the need for both personalization and generalization, ensuring that personalized models benefit from collective learning while retaining their local specificity.

The impact of PFL goes beyond technical improvements in accuracy or performance. It embodies a more user-focused approach to federated learning, where the unique needs and preferences of individual clients are considered. This aligns with the broader goals of federated learning, which include enhancing privacy and data security while promoting

more inclusive and fair machine learning. As federated learning continues to evolve, the development of PFL techniques underscores the commitment to addressing the complex and diverse needs of users, paving the way for a future where machine learning is not only more private but also more personalized.

Clustered-Based Federated Learning: Clustered-Based Federated Learning (CFL) [40–42] offers an innovative approach within the federated learning framework, specifically designed to address the challenges posed by statistical heterogeneity across distributed clients. The core idea behind CFL is to group clients into clusters based on the similarity of their data distributions [24] or other relevant features [40–42]. By doing so, CFL acknowledges the inherent diversity in client data and aims to improve both the efficiency and effectiveness of federated learning by tailoring the learning process to more homogeneous subsets of clients.

In CFL, clients are clustered based on a similarity metric $d(i, j)$, which measures the distance or similarity between the data distributions of clients i and j. For example, this similarity could be defined based on the distance between the empirical distributions of the datasets D_i and D_j:

$$d(i, j) = \text{distance}(P_{D_i}, P_{D_j}), \tag{2.5}$$

where P_{D_i} and P_{D_j} represent the empirical distributions of data on clients i and j, respectively. Once the similarity is computed, clustering algorithms such as k-means or hierarchical clustering are applied to form clusters of clients with similar data characteristics.

Once the clusters are formed, federated learning is conducted within each cluster. For cluster c, let w_c represent the model parameters for that cluster. The update rule for the model within cluster c follows the standard federated learning approach, where local models on clients $k \in c$ are trained and their updates are aggregated:

$$w_c^{t+1} = \sum_{k \in c} \frac{|D_k|}{\sum_{i \in c} |D_i|} w_k^{t+1}, \tag{2.6}$$

here, w_c^{t+1} is the updated model for cluster c at time step $t+1$, and $|D_k|$ is the size of the dataset on client k. This weighted aggregation allows the cluster model to reflect the collective information from the local models of the clients in that cluster.

The motivation for CFL is both simple and powerful: by grouping clients with similar data characteristics, CFL enables the training of specialized models that are better suited to the specific data profiles of each cluster. This approach allows for more precise model optimization, reducing the negative effects of statistical heterogeneity that can impair the performance of a single global model applied across all clients. CFL strikes a balance

between personalized learning and the collaborative nature of federated learning, enabling the creation of models that are locally relevant while still benefiting from global insights.

However, CFL is not without its challenges. One of the primary difficulties is the complexity of clustering clients in a dynamic and distributed environment [43], where data distributions may evolve over time and clients' operational conditions can vary significantly. Additionally, ensuring the privacy and security of client data during the clustering process is essential, as is maintaining the scalability of the federated learning system as the number of clients and clusters grows.

Despite these challenges, CFL presents a promising approach for federated learning, particularly in environments where client data is highly diverse. By organizing and leveraging this diversity, CFL enhances the potential for creating more effective and nuanced machine learning models, pushing the field closer to its goal of developing learning systems that are both deeply personalized and broadly collaborative.

Challenges in FL: Federated Learning faces a wide range of technical, operational, and ethical challenges that must be addressed to unlock its full potential. As FL continues to evolve, tackling these challenges is essential for its successful deployment across diverse domains. This section highlights two of the most pressing challenges in FL: Statistical Heterogeneity and System Heterogeneity, each presenting unique obstacles for FL systems.

Statistical Heterogeneity arises from the non-identically distributed nature of data across different clients in a federated learning environment. In real-world scenarios, data collected by various clients often reflect diverse patterns, preferences, and behaviors unique to each client. This diversity can hinder the learning process, resulting in models that perform well for some clients but poorly for others. The challenge lies in developing FL algorithms that can learn from such disparate data sources without compromising overall model performance.

To address Statistical Heterogeneity, innovative approaches are required that can accommodate these varied data distributions while ensuring that the federated model remains generalizable across all clients. Strategies such as Personalized Federated Learning [44, 34, 45–48] and Cluster-Based Federated Learning [40–42] aim to mitigate the impact of statistical heterogeneity. These approaches tailor the learning process to better fit the specific data characteristics of individual clients or clusters, enhancing the relevance and performance of the federated models.

System Heterogeneity refers to the variability in computational power, storage capacity, network connectivity, and battery life among the devices participating in FL. These differences can lead to disparities in the speed and efficiency with which clients contribute to the learning process. For example, clients with limited resources or unstable network

connections may lag behind more capable devices, causing delays and inefficiencies in model training and updates.

Additionally, System Heterogeneity poses the risk of client dropouts, where clients may become unavailable or leave the FL process due to device malfunctions, connectivity issues, or user decisions. Such dropouts reduce the available training data and can introduce biases into the model if departing clients represent critical or underrepresented data segments.

To address System Heterogeneity, FL algorithms and infrastructures must be designed to accommodate a wide range of device capabilities and operating conditions. Solutions like adaptive learning rates [49, 50], model compression [51, 52], and asynchronous communication protocols [53, 54] can mitigate the effects of system heterogeneity, ensuring that clients with varying capacities can participate effectively in the federated learning process. Table 2.1 provides a comparison of key literature addressing the challenges of Statistical and System Heterogeneity in Federated Learning. This comparison highlights the proposed solutions and the trade-offs associated with each approach, aiding in the understanding of the current state-of-the-art methods.

Table 2.1 Comparison of Literature Addressing Challenges in Federated Learning

Reference	Problem Addressed	Proposed Solution	Advantages and Limitations
Mansour et al. [44]	Statistical Heterogeneity	Personalized FL	High adaptability, but computationally expensive
Duan et al. [40]	Statistical Heterogeneity	Cluster-Based FL	Better model generalization, but requires effective clustering
Wu et al. [49]	System Heterogeneity	Adaptive Learning Rates	Improves performance for resource-limited clients, but may cause slower convergence
Shah et al. [51]	System Heterogeneity	Model Compression	Reduces communication costs, but can degrade model accuracy
Chen et al. [53]	System Heterogeneity	Asynchronous Communication	Allows for flexible client participation, but may introduce staleness in updates

In conclusion, overcoming Statistical Heterogeneity and System Heterogeneity is critical for the continued development and widespread adoption of Federated Learning. By creating more adaptable, efficient, and inclusive FL algorithms, researchers can foster robust and equitable machine learning models capable of leveraging the diverse and distributed datasets of the modern world.

2.4 Zero-Shot Learning

Zero-shot learning (ZSL) [4] stands as a remarkable endeavor within the broader landscape of machine learning, particularly for its ambitious goal to bridge the gap between seen and unseen classes without direct exposure to instances of the latter. This endeavor is driven by the exponential growth in the variety of data, where it becomes impractical to have labeled examples for every possible category. The journey of ZSL from its conceptual inception to its current state reveals a trajectory of evolving methodologies and expanding applications, underscoring the dynamic nature of this research area.

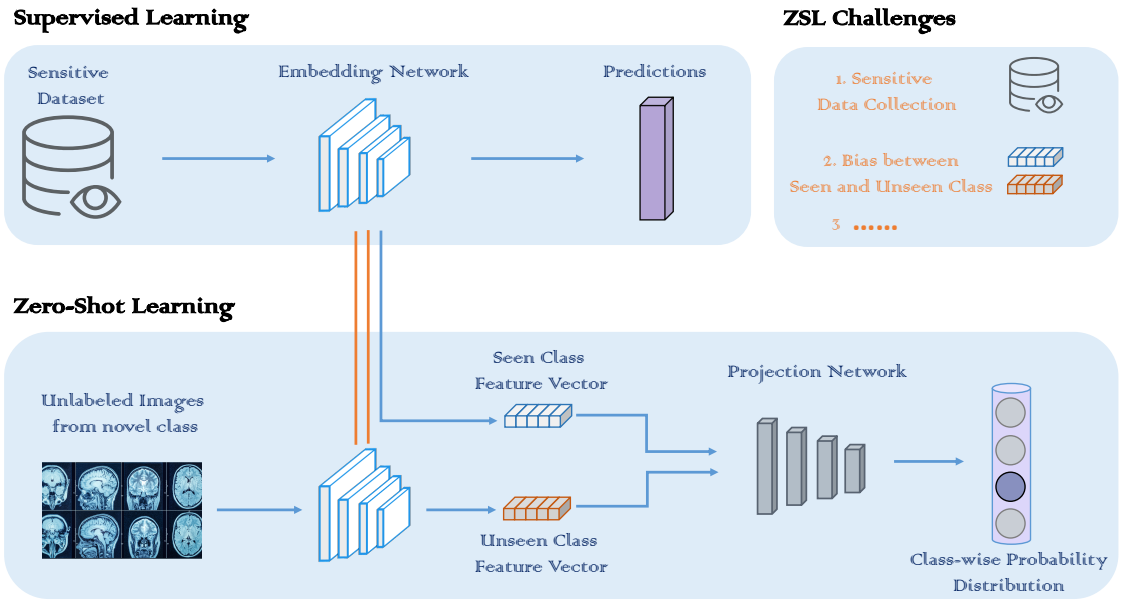


Fig. 2.2 The illustration of the Zero-Shot Learning

The foundational idea of ZSL germinated from the recognition of the limitations inherent in traditional supervised learning, where the performance is bounded by the scope of seen data during training. As shown in Fig. 2.2, ZSL aims to tackle this issue by using semantic embeddings to map the unseen classes without the need for labeled examples from those classes. In contrast, traditional supervised learning depends on having labeled data for each class during the training process.

Early approaches [55, 56] to ZSL focused on attribute-based methods, where classes (both seen and unseen) were described in terms of high-level attributes. This allowed models trained on seen classes to make inferences about unseen classes by leveraging shared attributes. Mathematically, ZSL can be formulated as follows:

Let $X \in \mathbb{R}^d$ be the feature space, and Y_s and Y_u represent the set of seen and unseen class labels, respectively. The goal of ZSL is to learn a classifier $f : X \to Y_u$ for unseen classes Y_u without having any training examples for these classes. Instead, the relationship between seen and unseen classes is defined through a semantic embedding space S, where each class $y \in Y_s \cup Y_u$ is associated with a semantic representation $\phi(y) \in S$. The objective is to transfer knowledge from Y_s to Y_u by mapping visual features $x \in X$ into the same semantic space S.

$$f(x) = \arg\max_{y \in Y_u} \text{sim}(g(x), \phi(y)), \tag{2.7}$$

where $g(x)$ is a function that maps visual features to the semantic space, and $\text{sim}(\cdot,\cdot)$ is a similarity measure (e.g., cosine similarity) between the embedded features and the class semantic representation.

As the field matured, the focus shifted towards improving the representations of classes and instances. The introduction of embedding spaces, where visual features and semantic labels are projected into a common space, marked a significant advancement. Techniques such as Semantic Output Codes [57] and Label Embedding [58] have been pivotal, enabling more nuanced mappings between visual features and class semantics. This period also saw the integration of external knowledge bases like WordNet [59] to enrich the semantic space and provide more context for the relationships between classes.

The advent of deep learning brought transformative changes to ZSL. Deep neural networks, particularly CNNs [60] for visual feature extraction and Graph Convolutional Networks [61] for exploiting structural relationships in data, have significantly enhanced the ability to capture complex patterns. Recent studies have employed GANs [62, 63] and Variational Autoencoders [64, 65] to generate synthetic examples of unseen classes, addressing the data scarcity issue head-on and improving the robustness of ZSL models.

While initial ZSL models [66, 67] were evaluated in a strict setting where only unseen classes were considered during testing, the realization of the practical limitations of this approach led to the emergence of Generalized Zero-Shot Learning (GZSL) [68–70]. In GZSL, the classifier must handle both seen and unseen classes simultaneously. Mathematically, GZSL can be framed as follows:

$$f(x) = \arg\max_{y \in Y_s \cup Y_u} \text{sim}(g(x), \phi(y)), \tag{2.8}$$

here, the challenge is to balance the performance on seen classes Y_s and unseen classes Y_u, ensuring that the model does not overfit to seen classes at the expense of unseen ones.

Table 2.2 below compares the characteristics and methods of Standard ZSL, Generalized ZSL, and Transductive ZSL, highlighting their differences in approach and real-world applicability:

Table 2.2 Comparison of ZSL, GZSL, and Transductive ZSL Approaches

Approach	Scope	Key Challenge Addressed	Main Methodology
Standard ZSL	Unseen classes only	Knowledge transfer from seen to unseen classes	Attribute-based or embedding techniques
Generalized ZSL	Seen and unseen classes	Coexistence of seen and unseen classes	Embedding techniques, external knowledge bases
Transductive ZSL	Seen and unseen classes with unlabeled data	Domain shift between seen and unseen	Incorporation of unlabeled data, domain adaptation

Amidst the rapid developments in ZSL, one aspect that has not been sufficiently addressed is data privacy and security. As ZSL models increasingly rely on sophisticated algorithms and large-scale data from diverse sources, the implications for data privacy cannot be

overstated. The generation of synthetic data for unseen classes, while innovative, introduces potential vulnerabilities and privacy concerns. These arise from the possibility of sensitive data being inadvertently learned or inferred during model training or the generation of synthetic instances, potentially exposing private information through the model's outputs. This highlights the need for privacy-preserving ZSL frameworks, especially as ZSL models become more integrated into applications involving personal or sensitive data. Therefore, future research must focus on ensuring that ZSL techniques are developed with privacy in mind, safeguarding user data while advancing the field.

In addition, ZSL faces challenges regarding its scalability and interpretability, particularly when applied to large-scale datasets with diverse and fine-grained categories. The ability to generalize effectively across vast and heterogeneous data sources is crucial, but it remains difficult for many ZSL models. Furthermore, the interpretability of ZSL models is becoming increasingly important, as stakeholders need to trust and understand the decision-making process. Although ZSL and FL are fundamentally different in their learning paradigms—ZSL focuses on knowledge transfer to unseen classes, while FL emphasizes decentralized model training—both approaches share common challenges in balancing model performance with data privacy concerns. In FL, privacy is maintained by ensuring data remains on the client side, while in ZSL, privacy-preserving techniques are needed to ensure that knowledge transfer does not expose sensitive information, particularly when synthetic data generation is involved. Therefore, advances in privacy-preserving techniques and scalable models in ZSL could offer valuable insights for FL, especially in environments where heterogeneous and sensitive data are critical factors.

As ZSL continues to evolve and intersect with other paradigms, including few-shot learning [71] and self-supervised learning [72], it moves towards the goal of enabling models to learn from limited or no direct examples. The convergence of these paradigms offers exciting possibilities for building machine learning systems that can operate in increasingly complex and dynamic environments while respecting privacy and ensuring trustworthiness.

2.5 Data-Free Knowledge Distillation

Data-Free Knowledge Distillation (DFKD) [73] represents a pivotal advancement in machine learning, particularly in addressing privacy concerns and optimizing model training efficiency. As models grow more complex and data privacy becomes a key concern, DFKD emerges as a solution that enables knowledge transfer from teacher models to student models without requiring access to the original training data. This section outlines the evolution of DFKD, key milestones, recent innovations, and its crucial role in enhancing data security.

The concept of knowledge distillation [30] revolves around transferring knowledge from a large, complex teacher model T to a smaller, more efficient student model S. Traditional distillation relies on both teacher and student models having access to the same dataset D, where the student model learns by matching the softened outputs (or logits) of the teacher model. Formally, the distillation loss function L_{KD} is defined as:

$$L_{\text{KD}} = \alpha L_{\text{hard}}(S(x), y) + (1-\alpha) L_{\text{soft}}(S(x), T(x)), \tag{2.9}$$

where L_{hard} is the standard cross-entropy loss between the student predictions $S(x)$ and the true labels y, and L_{soft} is the loss between the softened outputs of the teacher $T(x)$ and the student model. The hyperparameter α controls the balance between these two losses.

However, when privacy concerns or data availability constraints arise, the need for the original dataset D becomes problematic. DFKD addresses this issue by eliminating the dependence on real data. Early approaches [74, 75] focused on generating synthetic data that mimics the original data distribution, allowing the student model to learn from the teacher without directly accessing the true dataset.

A significant milestone in DFKD was the introduction of techniques to generate synthetic data directly from the teacher model. This process involves generating inputs x' that maximize the activation of certain neurons or layers in the teacher model, ensuring that the synthetic data captures the decision boundaries learned by the teacher. Mathematically, this can be formulated as an optimization problem where the synthetic input x' is found by solving:

$$x' = \arg\max_{x} \sum_{i} \text{Activation}(T(x)_i), \tag{2.10}$$

where $\text{Activation}(T(x)_i)$ represents the activation of the i-th neuron in the teacher model when given input x. By maximizing these activations, the generated data approximates the input space that the teacher model was trained on, enabling the student model to learn from the teacher effectively.

Recent advancements [76, 77] in DFKD have improved the fidelity of synthetic data through techniques such as activation maximization and the incorporation of prior knowledge. Activation maximization, where inputs are optimized to maximize the response of specific neurons in the teacher model, has become a powerful tool in generating meaningful synthetic data. In addition, regularization techniques [78] have been introduced to ensure that the generated data does not overfit the teacher model's idiosyncrasies.

DFKD directly addresses critical privacy concerns in machine learning by removing the need for real data during knowledge transfer. This capability is especially important in domains where data privacy is paramount, such as healthcare and finance. For instance,

DFKD allows organizations to deploy student models derived from powerful teacher models without exposing sensitive or proprietary data, ensuring that both knowledge and data remain secure.

Despite the privacy advantages of DFKD, potential vulnerabilities still exist. Recent studies [79, 80] have highlighted risks such as inversion attacks, where adversaries attempt to reconstruct sensitive information from synthetic data. Addressing these challenges requires more robust synthetic data generation methods, potentially incorporating privacy-preserving techniques such as differential privacy, where carefully calibrated noise is added to the synthetic data to prevent information leakage. The modified optimization process for synthetic data generation can be expressed as:

$$x' = \arg\max_{x} \left(\sum_{i} \text{Activation}(T(x)_i) + \mathcal{N}(0, \sigma^2) \right), \tag{2.11}$$

where $\mathcal{N}(0, \sigma^2)$ is Gaussian noise with variance σ^2, introduced to protect privacy while maintaining the quality of the generated data.

In addition to its privacy-preserving properties, DFKD faces challenges in generating high-quality synthetic data that fully encapsulates the teacher model's knowledge across various domains. Striking the balance between data quality and computational efficiency remains a major area of research. Furthermore, integrating DFKD with emerging paradigms such as federated learning and encrypted computation could push the boundaries of secure, efficient, and high-performance model training while maintaining data privacy.

In conclusion, DFKD represents a significant shift towards more privacy-conscious machine learning methodologies. As research continues, further improvements in synthetic data generation techniques and the integration of robust security measures will solidify DFKD's role in facilitating secure and efficient knowledge transfer across various AI applications.

2.6 Image Classification and Video Summarization Tasks

In this thesis, two primary tasks are explored: Image Classification and Video Summarization. These tasks were chosen due to their relevance in evaluating the proposed machine learning methodologies, specifically within the contexts of Federated Learning (FL), and Zero-Shot Learning (ZSL). This section introduces each task, describes the datasets used, and outlines the key metrics employed to gauge the performance of the models.

Image Classification

Image classification is one of the most fundamental and widely studied tasks in machine learning. The objective is to assign a label to an image based on its visual content, a task that is essential for various applications, including facial recognition, medical imaging, and autonomous driving. In the context of Federated Learning, image classification provides a straightforward yet effective benchmark to evaluate the performance of models trained across distributed clients with non-IID data.

Datasets

For image classification, this research primarily utilizes the following datasets:

- **CIFAR-10** [8]: A widely used dataset consisting of 60,000 32x32 color images across 10 classes, with 50,000 images for training and 10,000 for testing. Each class contains 6,000 images, ensuring a balanced distribution across categories.
- **CIFAR-100** [8]: Similar to CIFAR-10, but with 100 classes, each containing 600 images (500 training and 100 test images per class). The 100 classes are grouped into 20 superclasses, providing a more fine-grained classification challenge compared to CIFAR-10.
- **EMNIST** [9]: An extension of the MNIST dataset, EMNIST contains 814,255 handwritten character images from 62 classes (10 digits and 52 uppercase and lowercase letters). The dataset is divided into several subsets, allowing for the evaluation of models on digit and character recognition tasks. EMNIST is particularly relevant in Federated Learning experiments, as it simulates real-world data heterogeneity, making it a robust benchmark for evaluating model performance in non-IID scenarios.
- **Fashion-MNIST** [10]: A more complex variant of MNIST, this dataset contains 70,000 images of fashion products, split into 60,000 training images and 10,000 test images across 10 classes.

Evaluation Metrics

The performance of image classification models is evaluated using the following metrics:

- **Accuracy**: The proportion of correctly classified images out of the total number of images.

- **Precision and Recall**: Precision measures the fraction of true positive predictions among all positive predictions, while recall quantifies the fraction of true positives identified among all actual positives.
- **F1 Score**: The harmonic mean of precision and recall, providing a balanced measure of performance.

These datasets and metrics allow for the rigorous testing of models developed under Federated Learning and Zero-Shot Learning frameworks, particularly when training occurs over non-IID distributed data.

Video Summarization

Video Summarization (VS) aims to distill the essential content of a video into a more concise format, typically by selecting keyframes or segments that represent the video's main events. This task is crucial in environments where video content is generated in vast quantities, such as surveillance, entertainment, and social media. Video Summarization aligns well with machine learning tasks as it requires models to capture both spatial and temporal information, making it a more complex and informative benchmark for evaluating Federated Learning.

As shown in Fig. 2.3, the process of video summarization involves inputting a sequence of raw video frames into an AI model, which then selects the most representative frames (or shots) to produce a condensed summary of the original video. This summarized output allows for more efficient consumption of video content by retaining key moments while discarding redundant or less relevant parts.

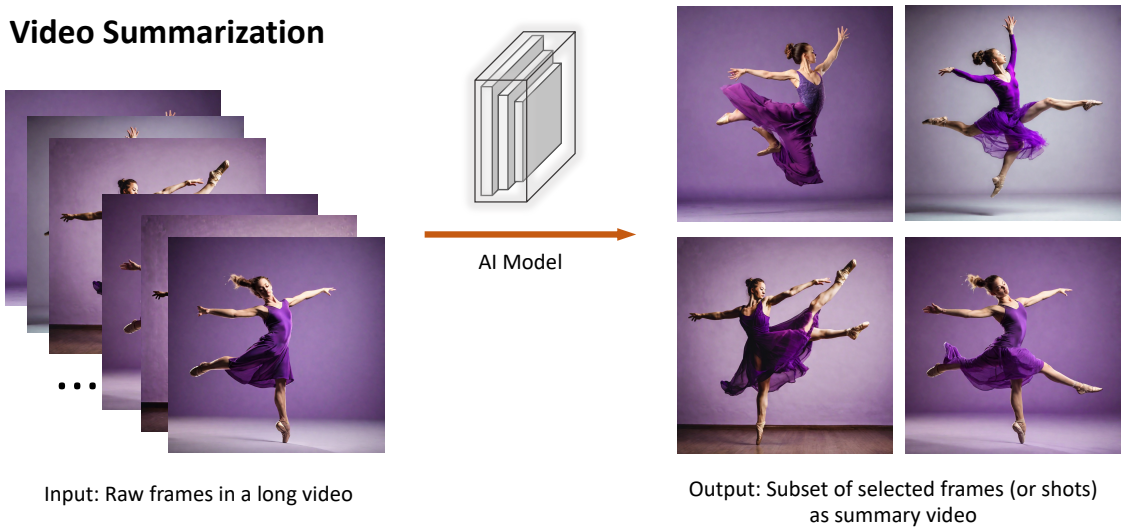


Fig. 2.3 The illustration of the Video Summarization

Datasets

The following datasets were used to evaluate video summarization models in this research:

- **TVSum** [7]: This dataset consists of 50 videos from a wide range of genres (e.g., news, documentaries, sports) and provides human-annotated importance scores for video frames, facilitating both supervised and unsupervised video summarization tasks.
- **SumMe** [6]: Comprising 25 videos, SumMe focuses on personal and realistic videos, offering diverse content with human-annotated summaries that serve as a benchmark for evaluating the quality of video summaries.

Evaluation Metrics

Video summarization is evaluated using the following key metrics:

- **F1-Score**: The F1-Score is widely used to compare the overlap between machine-generated summaries and human-annotated ground-truth summaries.
- **Precision and Recall**: Similar to image classification, these metrics evaluate the proportion of selected frames or segments that are relevant (precision) and the proportion of relevant frames or segments that are successfully selected (recall).
- **Coverage**: Coverage measures the percentage of the original video content that is retained in the summary, ensuring that the generated summaries are both concise and representative of the full content.

Relevance of Image Classification and Video Summarization to This Research

Both image classification and video summarization serve as critical tasks for evaluating the proposed machine-learning models. While image classification allows for the testing of models on simpler, static datasets, video summarization provides a more dynamic and complex task that challenges models to handle temporal dependencies and diverse content. These tasks are complementary and highlight the versatility and robustness of the models developed in this thesis. In particular, video summarization provides a valuable testbed for Federated Learning due to the decentralized nature of video data sources, and Data-Free Knowledge Distillation offers an innovative solution for maintaining privacy in such settings.

Chapter 3

Asynchronous Personalized Federated Learning Through Global Memorization

Prologue

In the era of modern computing, the vast amount of data generated by Internet of Things devices and communication networks brings with it an increasing need for privacy-preserving machine learning solutions. Traditional approaches, which rely on centralized data collection and processing, expose significant vulnerabilities in terms of user privacy and data security, making it essential to explore new methods that can protect data without compromising model performance.

Asynchronous Personalized Federated Learning (AP-FL) addresses these critical concerns by focusing on two key challenges in federated learning systems: non-IID data distributions and client dropouts. These issues are common in real-world scenarios, where data generated by different clients often vary significantly, and where clients may intermittently leave the learning process due to connectivity issues or other limitations. AP-FL introduces a personalized learning approach combined with model interpolation, allowing each client to develop tailored models that reflect their unique data characteristics.

Furthermore, the framework incorporates a data-free knowledge transfer mechanism to handle client dropouts. This ensures that even when clients leave the network temporarily, the overall learning process remains robust, preserving model integrity and continuity. By addressing these challenges, AP-FL not only improves the predictive accuracy of federated models but also enhances learning efficiency, making it a practical solution for real-world federated learning applications.

Declaration: This chapter is a modified version of **"Asynchronous Personalized Federated Learning through Global Memorization"**, submitted to IEEE Transactions on Image Processing (TIP), 2024.

3.1 Introduction

The rapid proliferation of Internet of Things devices, from home automation systems and wearable health monitors to smart city sensors, coupled with the advances in communication technology, has led to an explosion of data generation in our daily lives. This vast expanse of data spans intricate applications such as facial recognition systems, detailed health data from fitness trackers, and extensive urban data from smart infrastructure, all of which harbor the potential to significantly advance the field of artificial intelligence. However, the deeply personal nature of such data, combined with an alarming escalation in privacy breaches, has intensified global scrutiny over data privacy. Legislative milestones like the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) in the United States have underscored the imperative for robust data protection measures. These developments compel the AI community and scholars to innovate a framework that not only ensures rigorous protection of privacy but also enables efficient utilization of the burgeoning data, striking a crucial balance between utility and confidentiality.

In response to the urgent need for innovative solutions that preserve privacy while leveraging vast datasets, Federated Learning [3] has emerged as a groundbreaking paradigm. FL facilitates the collaborative training of a global model across multiple devices or clients without the necessity of centralizing local data. This decentralized approach involves each participant training models on their own devices, followed by the aggregation of these models into a cohesive global model, which is then updated and redistributed to all participants. By enabling data to remain securely on local devices, FL adeptly addresses the critical balance between data privacy and utility. Its application spans diverse sectors, from enhancing privacy in smart cities [81, 82] and improving diagnostic accuracy in healthcare [83, 84] to personalizing user experiences in digital services [85, 86], thereby illustrating its transformative potential in securely and efficiently harnessing data across industries.

Federated Learning, while promising, grapples with significant challenges such as statistical and systems heterogeneity. Statistical heterogeneity arises when data from diverse user devices vary widely due to factors such as geographical distribution, differing time zones, or unique user behaviors, leading to client drift. This phenomenon can degrade performance and slow the convergence of the global model, as seen in [87]. Furthermore, system heterogeneity compounds these issues, with disparities in device capabilities—like network

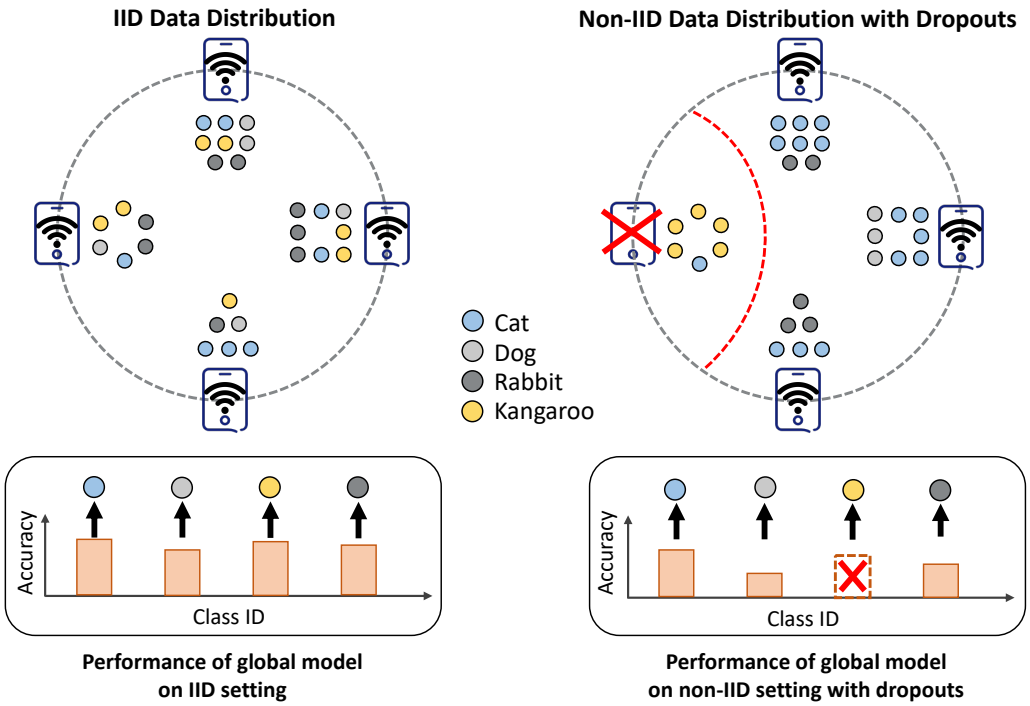


Fig. 3.1 The illustration of the impact of non-IID data distribution and dropout clients with monopoly classes on global performance.

bandwidth or battery life—affecting timely updates and further destabilizing training [35]. These heterogeneities not only challenge model training but also heighten the risk of creating monopolistic classes where single participants or groups disproportionately influence the model due to their unique data contributions.

The severe implications of monopolistic class dropouts, particularly within contexts of statistical and systems heterogeneity, are vividly illustrated in the healthcare sector. For example, if a healthcare provider uniquely treating a rare medical condition exits a federated network due to regulatory changes or technical failures, the global model instantly loses critical diagnostic data. This sudden dropout not only degrades the model's accuracy but also exposes the inherent vulnerabilities of relying on limited data sources. As depicted in Figure 4.1, while the model under idealized IID conditions might perform well, it encounters significant challenges in real-world settings marked by non-IID data distributions, especially when essential data sources vanish. This necessitates the development of innovative methods that effectively manage such dropouts, addressing both system heterogeneity and ensuring robust performance across diverse and realistic conditions.

Research on the impact of challenges from statistics and system heterogeneities have been extensive yet fragmented [44, 34, 45]. Previous studies [88–92] have addressed various dropout scenarios on the performance of global models, primarily under the assumption of

independent and identically distributed (IID) conditions, where the impact of dropouts is minimal [88, 89, 93]. However, the real-world applicability of these findings is limited as they often overlook the complexities introduced by non-IID data distributions. Recent attempts to explore these issues in more realistic settings [92, 90, 91] have revealed significant gaps in existing methodologies, particularly in their ability to handle unpredicted dropouts and maintain data diversity without compromising the model's integrity.

In this research, we introduce the Asynchronous Personalized Federated Learning Framework (AP-FL) as a new approach to tackle the challenges of statistical and system heterogeneity. AP-FL employs a data-free knowledge transfer method to train a generator on the server side. With the aid of semantic information from Zero-Shot Learning and supervision from the received global model, the generator can generate seen samples from non-dropout clients and unseen samples from dropout clients to facilitate client model training. However, synthetic samples generated by the generator heavily rely on global model performance, which poses a risk when global model performance is suboptimal. To address this risk, we propose a decoupled model interpolation algorithm to mitigate the negative impact of synthetic data on Personalized model training.

The main contributions of this work are summarized as follows:

- In order to address the non-IID challenge, we propose a novel personalized federated learning framework leveraging model interpolation.
- A novel FL framework to solve the class missing due to dropouts via data-free knowledge transfer and ZSL mechanism.

3.2 Related Work

Statistic Heterogeneity presents a major challenge in Federated Learning (FL) setups. Conventional FL approaches frequently experience client drift issues [87] in the presence of highly heterogeneous statistics (non-IID), resulting in diminished global model performance and suboptimal generalization across numerous clients. To address this challenge, several existing works [44, 34, 45–48] have started to research Personalized Federated Learning (PFL) which has recently gained considerable attention for its ability to adapt the global model to better fit each client's local data distribution. One of the research methodologies focuses on personalized a single global model by introducing techniques, such as Data Augmentation[36], Client Selection[94], Regularization[35], and Meta-Learning[95].

Data augmentation, such as FAug [36], promotes statistical homogeneity by generating new data or using proxy data for clients, enabling the satisfaction of the IID assumption and benefiting the training of a unified global model through server-side generative adversarial

networks trained with limited client-side samples for IID dataset generation. Client selection, such as the adaptive reinforcement learning algorithm proposed by [94], identifies representative client subsets to capture the global data distribution, mitigating non-IID data impact and improving the performance and communication efficiency of the trained model. Model regularization, exemplified by Fedprox[35], introduces a regularization term in the loss function to constrain personalized models from deviating significantly from the global model, effectively limiting the impact of irregular client updates. Meta-learning, inspired by local fine-tuning from the global model, was introduced into PFL, building initial meta-models for clients to fine-tune after one model gradient descent step, as exemplified by [95], which combines meta-learning and reinforcement learning to adaptively optimize the federated learning process.

Systems Heterogeneity as another crucial factor beyond statistical heterogeneity that should be considered in the federated network, since interplay exists between them in federated learning [32]. In a real-world federated training task, thousands of devices possibly participate, with diverse system-level attributes, hardware configuration(CPU, Memory), network connectivities (wired and wireless network), and battery capability [96, 32]. Such characteristics substantially heighten the uncertainty within a federated network, giving rise to challenges such as misleading optimization direction, straggler issues, and client dropout problems. To tackle systems and statistical heterogeneity problems, [96] proposed an adaptive client sampling algorithm that reduces convergence duration by determining the relationship between overall learning time and sampling probabilities. In addition, [32] proposed a novel federated optimization algorithm, widely known as FedProx. FedProx alleviates the impact of systems and statistical heterogeneity on convergence behavior by introducing a proximal term to the objective, thereby increasing stability. This addition offers a principled approach to handling heterogeneity associated with partial information, allowing for convergence guarantees and an analysis of the effects of heterogeneity. While these approaches have effectively mitigated the impact of system and statistical heterogeneity issues broadly, their efficiency remains limited in specific situations, such as client dropout problems. Most recently, very limited studies have started focusing on client drop problems. Wang and Xu [97] propose the concept of "friendship" between clients, wherein clients with similar data distributions and local model updates are considered friends. This approach seeks to alleviate the impact of client dropout by substituting a friend client's local model update for the dropout client's update when computing the next round global model, resulting in minimal substitution error. However, while this method mitigates the negative impact on global model performance, it does not enhance the global model's effectiveness on the dropped client's local dataset.

Asynchronous Personalized Federated Learning. Building on the advancements in Personalized Federated Learning (PFL), our proposed Asynchronous Personalized Federated Learning (AP-FL) framework is designed to address both statistical and system heterogeneity. In PFL, the challenge of non-IID data is tackled by allowing each client to develop a model that is personalized to its local data distribution. In AP-FL, this personalization is achieved by maintaining the same model architecture across clients but allowing for distinct model weights that adapt to the specific data characteristics of each client. This approach ensures that while all clients benefit from shared global knowledge, their individual models are fine-tuned to address local data heterogeneity. To mitigate the effects of client dropouts, AP-FL incorporates a novel data-free knowledge transfer mechanism, allowing the generation of synthetic samples that aid in the continuous training of client models even when clients are temporarily offline. This strategy effectively handles both asynchronous updates and the challenges posed by varying client availability, leading to a more robust and efficient federated learning process.

3.3 Methodology

This section presents the proposed AP-FL. We first describe the problem statement, followed by AP-FL framework design. Several key modules of AP-FL are detailed at both the server and client sides.

3.3.1 Problem Statement

Conventional federated learning approaches, such as FedAvg [3], address C-class classification problems across K clients. For each client $k \in \{1,2,\ldots,K\}$, its private local dataset $\mathscr{D}_k$ is drawn from the local data distribution $p_k(x,y)$, where $x \in \mathscr{X}$ is the input feature, and $y \in \mathscr{Y}$ denotes the corresponding label. The goal of FL is to enable clients to jointly train a global model with parameters θ^* over the combined global dataset $\mathscr{D} = \bigcup_k \mathscr{D}_k$.

The global objective is to find the optimal model parameters θ^* that minimize the global loss $\mathscr{L}(\theta)$, which can be formulated as:

$$\theta^* = \arg\min_{\theta} \mathscr{L}(\theta), \tag{3.1}$$

where $\mathscr{L}(\theta)$ represents the empirical loss over the entire global dataset $\mathscr{D}$. This global loss is computed as the weighted sum of the local losses $\mathscr{L}_k(\theta)$ from each client k, with the weights proportional to the size of each local dataset $\mathscr{D}_k$:

$$\mathscr{L}(\theta) = \sum_{k=1}^{K} \frac{|\mathscr{D}_k|}{|\mathscr{D}|} \mathscr{L}_k(\theta), \tag{3.2}$$

here, $\mathscr{L}_k(\theta)$ denotes the local loss for client k, and $\frac{|\mathscr{D}_k|}{|\mathscr{D}|}$ is the weighting factor based on the proportion of data that client k contributes to the global dataset.

All clients aim to optimize the global model θ by minimizing their local expected risk:

$$\mathscr{L}_k(\theta) = \mathrm{E}_{(x,y)\in\mathscr{D}_k} \mathscr{L}(\theta;(x,y)). \tag{3.3}$$

The key steps involved in a complete FL training process are outlined below: (i) At communication round t, the aggregator server randomly selects K clients available for training and sends the global model θ^* to the selected clients, which they deploy as a local model, θ_k^t. (ii) Each selected client trains its local model θ_k^t using its dataset $\mathscr{D}_k$ for E local epochs. (iii) Once the aggregator server collects local model updates from enough participants, θ_k^{t+1}, the server aggregates all updates based on Equation 3.2. (iv) Repeat steps (i)$\sim$(iii) until the model reaches convergence.

Client drift issues posed a serious challenge when implementing FL in the real world. The performance and efficacy of the vanilla FedAvg algorithm have been demonstrated in Independent and Identically Distributed (IID) settings, where each client has similar data distribution, and samples are identically distributed among clients. However, it fails in Non-IID settings, where data distribution between clients can be highly skewed, and sample distribution may differ significantly. This can lead the locally trained model to be optimized in a direction that deviates significantly from its trained in an IID dataset.

Figure 3.2 illustrates how FedAvg performs in both IID and non-IID settings. The average model θ^{t+1} is equidistant to both local optima θ_1^* and θ_2^* in an IID setting, which brings it closer to the global optimum θ^*. However, in Non-IID settings, the resulting average model θ^{t+1} may not be close to the global optimum θ^*, causing the global model not to converge to its true global optimum. In these scenarios, the single global model is difficult to generalize well to all clients, and the performance of the global model may not even exceed the local model where the client does not participate in FL training[98]. This is contrary to the original intention of the client to participate in FL.

Analogous to the straggler issue in distributed systems, client dropout is a prevalent phenomenon in federated networks with system heterogeneity. In certain non-IID scenarios, such as those characterized by extreme shifts in data quantity and class categories, client dropout can amplify the adverse effects on global model optimization. An existing study [98] indicates that, given a sufficient number of clients continuously participating in federated learning training under IID data settings, the accuracy of the global model remains unim-

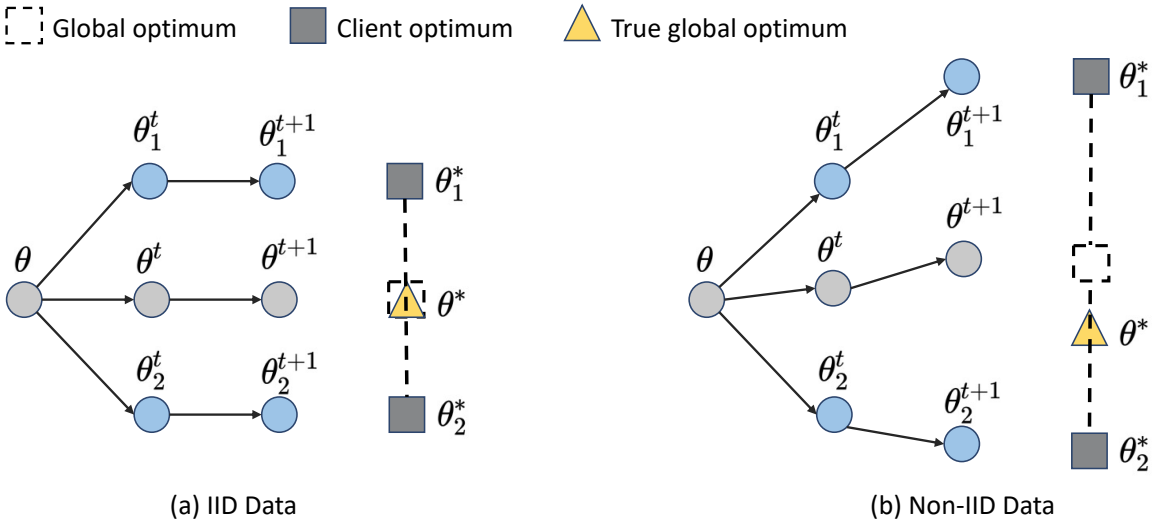


Fig. 3.2 Illustration of client drift in FedAvg in Dirichlet non-IID settings.

paired, even if permanent dropouts among some clients. However, in non-IID scenarios, where some clients have unique or minority classes that are not present in the datasets of other clients, the dropout of those clients can significantly negatively impact the performance of the global model. This is because the performance of the global model relies on contributions from all participating clients to learn a representative model. When a dropout client has unique class category that is not represented by other clients, as the training continues, the global model will be fitted to the optimal of the other available classes, resulting in an extremely rapid decline in the global model's ability to identify the missing class data.

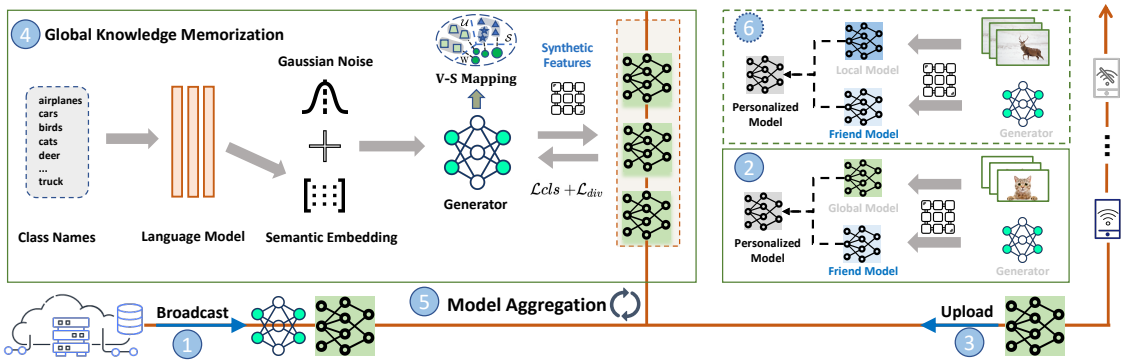


Fig. 3.3 Overview of the Asynchronous Personalized Federated Learning.

Our work is motivated by recent advances in PFL[50], but it goes beyond it by addressing system heterogeneity, specifically the challenge of client dropout, in addition to the problem of statistical heterogeneity. We aim to develop a global knowledge that can help non-dropout and dropout clients to build a personalized model that can tackle client local drift issues,

even when the data on dropout clients are distinct from those on all non-dropout clients. To achieve this goal, we propose to train a personalized supervised classification model for a group of non-dropout clients S_n and dropout clients S_d, where $S_n, S_d \in K$.

$$\theta_k^p = \arg \min_{\theta_1,\ldots,\theta_K} \sum_{k \in S_n \cup S_d} \frac{|\mathscr{D}_k|}{|\mathscr{D}|} \mathscr{L}_k(\theta_k^p), \tag{3.4}$$

where θ_k^p represents the personalized model residing on client k. Our approach is distinct from other methods that aim to mitigate client dropout, as our focus is not only on dropout clients but on all clients. By enabling dropout clients to benefit from global knowledge and establish their personalized models, our method can address the challenge of statistical heterogeneity while also tackling the issue of client dropout.

3.3.2 Proposed Framework: AP-FL

Numerous studies in recent years have focused on addressing statistic and systems heterogeneity by capturing global knowledge, such as GAN-based approaches [99–103]. However, most of them require the generator access to clients' raw data, contradicting the original principles of federated learning. Alternatively, knowledge distillation-based methods [104–106] rely on a proxy dataset and tackle client drift issues by leveraging disagreement between global and client models. Nevertheless, the availability of a proxy dataset in real-world federated learning scenarios cannot always be guaranteed.

To tackle these challenges posed by client drift and client dropout in above non-IID scenarios, we introduce a novel federated learning framework termed AP-FL, illustrated in Figure 3.3. AP-FL is a plugin that could cap into most widely use neural network, and features a lightweight semantic generator, maintained by the central server, which captures global knowledge through data-free knowledge transfer from the global model. This semantic generator is disseminated to non-dropout clients to support the development of personalized models tailored to their data distribution. Considering the likelihood of a single client dominating minority classes in real applications, we adopt the Zero-Shot learning paradigm, enabling the semantic generator to create synthetic data for minority classes present on dropout clients. This is achieved by establishing a mapping between semantic information and features, even without direct access to the dropout client data by the global model. Consequently, this approach facilitates asynchronous training of personalized models by dropout clients based on their unique data distribution, supported by the semantic generator.

Global Knowledge Memorization. As previously discussed, non-IID scenarios can result in client drift issues, adversely affecting model performance. Therefore, it is essential to

devise a conditional generator, denoted as G, maintained on the central server side to capture the global perspective of data distribution. This generator aims to assist each client in developing a personalized model θ_k while preserving user privacy. Specifically, the server broadcasts G to support non-dropout (non-dropout) clients S_n in training personalized models by generating synthetic samples that enhance the diversity of client data distribution. The completed process of global knowledge memorization could be summarized as follows: Firstly, the generator is initialized on the server-side as follows:

$$\hat{x} = G(z, y; \omega), \tag{3.5}$$

here, ω denotes the parameters of G, and $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ represents the standard Gaussian noise, which is introduced to increase the diversity of the generated data and reduce overfitting. The variable y is the label representing the desired output class, while $\hat{x}$ is the synthetic sample corresponding to the input noise z and label y.

Due to the scarcity of resources for training G, only the global model θ^* and the client local models θ_k are accessible. Therefore, it is imperative to ensure that the synthetic samples $\hat{x}$ generated by the G are compatible with the input space of client local models θ_k. This can be formulated as follows:

$$\mathscr{L}_{ce} = -\sum_{i=1}^{C} y_i \log\left(\sigma\left(D\left(\hat{x}; \theta_k\right)_i\right)\right), \tag{3.6}$$

here, i indexes the classes, and C represents the total number of classes. The softmax function $\sigma(\cdot)$ outputs a probability distribution over the C classes, and $D(\hat{x}; \theta_k)$ denotes the output of the client model θ_k when given the synthetic sample $\hat{x}$. The term y_i is the ground truth label for class i, and the cross-entropy loss $\mathscr{L}_{ce}$ measures how well the model's predicted distribution aligns with the true labels. To well fit the synthetic samples effectively with each client model's data distribution, we incorporate a weighted average of the loss function, considering the distribution of distinct categories for each user. Consequently, the weighted average cross-entropy loss is defined as follows:

$$\mathscr{L}_{cls} = \sum_{k \in S_o} \alpha_k^y \mathscr{L}_{ce}^k, \tag{3.7}$$

where α_k^y represents the proportion of samples in class y of the k-th non-dropout client in the entire global training set, and $\mathscr{L}_{ce}^k$ represents the cross-entropy loss produced by k-th non-dropout client.

Employing only the $\mathscr{L}_{cls}$ may result in the generator's model collapse [107], causing G to output identical data for every class. To motivate G to enhance the diversity of synthetic samples, we incorporate a regularization term into the loss function. Specifically, we introduce a diversity loss term, which encourages the generator to generate varied samples. The diversity loss is defined as follows:

$$\mathscr{L}_{diversity} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1,j\neq i}^{n_s}\frac{\left|\hat{x}_i - \hat{x}_j\right|_2}{n_s - 1}, \tag{3.8}$$

where n_s is the number of synthetic samples, and $\hat{x}_i$ and $\hat{x}_j$ are two different synthetic samples of same classes generated by the generator G. The term $n_s - 1$ in the denominator is a normalization factor to appropriately scale the diversity loss. This loss term encourages the generator to produce diverse synthetic samples by minimizing the Euclidean distance between any two different synthetic samples. The overall loss function for the generator is then defined as follows:

$$\mathscr{L}_G = \lambda\mathscr{L}_{cls} + (1-\lambda)\mathscr{L}_{diversity}, \tag{3.9}$$

where λ is the hyper-parameters that control the relative importance of the two loss terms. By minimizing this loss function, the generator G is encouraged to produce diverse synthetic samples that better capture the underlying data distribution of the client models.

PFL via Decoupled Model Interpolation. The majority of existing studies addressing non-IID problems concentrate on data generation-based methods, such as mixing up non-IID real and synthetic data into a unified IID training set for each client's local model or utilizing fake data to capture the disagreement between global and local models for bi-level knowledge distillation, ultimately enhancing the performance of global or local models. However, for both approaches, the data generation capability of the generator heavily relies on the global model's accuracy. Consequently, the training quality of the generator cannot be guaranteed in this manner, a limitation similar to that encountered in our solution.

To overcome this challenge, we propose a decoupled model interpolation method that modulates the impact of synthetic samples within personalized federated learning. In this approach, users utilize the trained generator to generate synthetic samples $\hat{x}$ conforming to their local data distribution $\mathscr{P}_k$. These synthetic samples are subsequently employed to train a classifier, referred to as the friend model. Finally, we combine the client model and friend model to create a personalized model that more effectively adapts to the user's local data. The following equation illustrates the decoupled model interpolation method:

$$\theta_k^p = \beta\theta_k + (1-\beta)\theta_k^f, \tag{3.10}$$

where θ_k^f and θ_k^p represent the model parameters of the friend model and personalized model separately in k-client, and β represents the confidence coefficient for the friend model.

PFL for the dropouts. In addition to tackling statistic heterogeneity, AP-FL also handles client dropout issues lead by systems heterogeneous. In real-world FL training, which may involve thousands of clients, communication bandwidth constraints within a distributed system necessitate the selection of only a limited number of clients to participate in each training round. This situation can result in clients possessing all data of minority classes not engaging in FL training from start to finish before dropping out. This implies that those data categories are never present in any non-dropout client, and we refer to them as unseen classes for the global model. Sending the global model to these dropouts would be unproductive, as the global model has not seen data from these dropouts' categories and, consequently, cannot identify the data for these categories.

Inspired by the works in Zero-Shot Learning [108], we distinguish data from non-dropouts and dropouts as seen data $\mathscr{D}_s$ and unseen data $\mathscr{D}_u$, respectively. The relationship between their data categories is disjoint and can be formulated as $\mathscr{Y}_s \cap \mathscr{Y}_u = \emptyset$. The main challenge lies in obtaining informative semantic information that allows the generator to establish the mapping between features and semantic information, enabling the generator to synthesize unseen data from dropout clients. Conventional ZSL approaches benefit from auxiliary semantic embedding information, such as attributes annotated by experts in relevant fields. However, traditional FL datasets lack such auxiliary information. To tackle this issue, we employ foundation models like BERT [109] and CLIP [110], which are pre-trained on extensive data and can predict underlying properties, such as attributes.

While large foundation models such as BERT and CLIP have been leveraged to aid the semantic embedding process, it is important to acknowledge that these models introduce strong priors due to their extensive pre-training on large-scale datasets. This could potentially raise concerns about the fairness of comparison with other approaches that do not utilize such pre-trained models. To mitigate this, we carefully ensure that the usage of BERT and CLIP is balanced with techniques that limit their overwhelming influence on final model performance. Additionally, their role is primarily to provide semantic structure in the absence of labeled data, rather than directly contributing to model learning in traditional ways. By using BERT and CLIP in a controlled manner, we aim to ensure that the comparison with other FL systems remains as fair and unbiased as possible, focusing on the federated learning performance rather than overreliance on pre-trained representations.

To develop the mapping between features and semantic information, we support the Generator on the central server side, which can generate pseudo data from $\mathscr{D}_s$ to $\mathscr{D}_u$. So the

input of Generator in Eq. (3.5) should become the following format:

$$\hat{x} = G(z, A(y); \theta), \tag{3.11}$$

where the $A(\cdot)$ represents auxiliary semantic embedding. Finally, the personalized model in all clients can be formulated as follows:

$$\theta_k^p = \begin{cases} \beta\,\theta_k + (1-\beta)\theta_k^f & \text{non-dropout clients;} \\ \beta\,\theta_k^l + (1-\beta)\theta_k^f & \text{dropout clients.} \end{cases} \tag{3.12}$$

Where θ_k^l represents the localized global model for the dropout in k-client, and θ_k^f and θ_k^p denote the model parameters of the friend model and personalized model separately in k-client.

The global knowledge model captures rare class distributions through the ZSL approach. When dropout clients contain rare or unique class data, the generator synthesizes samples for these classes using semantic embeddings. This approach helps maintain model performance in non-IID settings by generating synthetic data for unseen or underrepresented classes. By mapping semantic features to the output space, the model can generalize across diverse data distributions, ensuring that rare classes are effectively learned. As a result, both dropout and non-dropout clients benefit from accurate and robust personalized models, ensuring comprehensive generalization across all classes.

Discussion. Our proposed PFL approach is essentially an implementation of the clustering-based Federated Learning (CFL) method on the client side. CFL aims to group clients with similar data distributions to help clients train a pair-wise group model with a friend model. However, the training process of CFL can be affected by the dynamic grouping of clients due to the emergence of new data samples. In contrast, our method offers a more flexible strategy by generating synthetic data. Users can continuously generate synthetic data based on their changing data distribution and train a personalized model using a local-side clustering method. This personalized model is better suited to the user's data distribution, leading to improved generalization performance.

Moreover, our approach utilizes asynchronous aggregation to enhance the robustness of the training process, particularly in scenarios where client dropouts or system heterogeneity are present. Asynchronous aggregation allows the global model to be updated as soon as updates from any client are received, thus reducing waiting time and improving training efficiency. However, we acknowledge that asynchronous aggregation can introduce consistency challenges, as updates may be based on stale or outdated models. This trade-off contrasts with synchronous aggregation, which waits for updates from all clients, ensuring

that the global model is consistently updated but potentially introducing delays due to slower or offline clients. In our framework, asynchronous aggregation helps mitigate the impact of client dropouts and system variability, while still maintaining high model performance through careful model update mechanisms.

3.4 Experiments

In this section, we present the evaluation of the effectiveness of our proposed method, AP-FL, and compare it with several advanced methods in different datasets and settings. The evaluation focuses on two key aspects: (1) personalized model accuracy in non-dropout clients, and (2) the improvement in model accuracy for dropout clients with the assistance of global knowledge.

3.4.1 Basic Set

Dataset: This study presents experimental results on four diverse image datasets: CIFAR10 [8], CIFAR100 [8], EMNIST [9], and Fashion MNIST [10]. The CIFAR10 dataset contains 60,000 32x32 color images divided into ten classes, which has been widely used for image classification tasks. CIFAR100 is a more challenging variant of CIFAR10, consisting of 100 classes. The EMNIST dataset is a collection of over 800,000 images of 26 handwritten letters, while Fashion MNIST comprises 70,000 grayscale images of 28x28 pixels, representing ten different clothing categories. To maintain consistency in image resolution, we resized all images to 32x32 pixels. To evaluate our model, we set aside 10% of the data for testing purposes, and we distributed the test data among the clients while ensuring that the test data had the same label distribution as the training data on each client's side.

Heterogeneity Settings: The performance of the proposed AP-FL framework is evaluated in two distinct heterogeneity settings to analyze its efficacy under varying degrees of heterogeneity. (1) **Full Participated Setting**, solely accounts for statistical heterogeneity and considers an ideal FL scenario where all clients are available and selected randomly by the server without dropped calls. Similar to [111, 112], we adopt the Dirichlet Distribution $Dir(\alpha)$ to control the degree of non-IID distribution. Specifically, we set α to three different values, namely 0.1, 0.05, and 0.01, across three image datasets - CIFAR10, CIFAR100, and EMNIST. Since FEMNIST already considers various kinds of imbalances, such as data heterogeneity, data imbalance, and class imbalance, we did not apply the Dirichlet distribution to FEMNIST. Furthermore, we varied the number of clients to five and ten to simulate different levels of non-IID data. (2) **Dropout Setting**, a dropout factor is introduced to simulate more practical

scenarios where FL training encounters both statistical and system heterogeneity. In this setting, we adopt the Pathological non-IID [113] approach, where only certain classes of data are assigned to each client. We simulate ten clients to jointly train a global model in all datasets, and then we use the hyper-parameter γ to control the number of classes on each client. As shown in Table 3.1, when $\gamma = 2$ means that there are two classes of data on each client. We assume some rare clients with monopoly classes will drop out to verify the effectiveness of the proposed personalized model in dropout clients.

Baselines: This study presents a comprehensive comparison of the proposed AP-FL framework with several baseline algorithms in two distinct settings. **In the Full Participated Setting**, we compare AP-FL against FedAvg [3], FedProx [35], SCAFFOLD [87], FedGen [107], and FedDF [114]. In addition, we evaluate the performance of AP-FL against local training, which involves training a local model without the use of federated learning. **In the Dropout Setting**, we compare FedAvg [3] and local training as the baseline approaches. For FedAvg, we conduct one-off fine-tuning training for the global model trained by the non-dropout client in the dropout client with its monopoly classes and then test its global model performance in monopoly classes. For local training, we send the initial global model to dropout clients and train the local model without federated learning.

Table 3.1 Data Partitioning for $\gamma = 2$ Pathological Non-IID on CIFAR10 dataset, in the Dropout Setting. The classes $[8,9]$ denote the minority classes monopolized by rare clients.

Device No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Classes	0, 1	2, 3	6, 7	4, 5	2, 4	2, 3	6, 7	4, 5	[8, 9]	0, 1

Implement Details: We implement all experiments of AP-FL in PyTorch, where the classifier in all experiments is a standard CNN model, which consisting of two 5×5 convolution layers (the first with 32 channels, the second with 64 channels, each followed with 2×2 max polling), two fully connected layers each with 1600, 512 units and ReLU activation. For semantic embedding, we use 512-dimensional word embedding generated by CLIP [110]. Our generator network architecture is borrowed from [115], but we replace the input of an original one-hot label with the semantic embedding generated from various models. All methods were trained with a batch size of 50 and optimized using the Adam optimizer with an initial learning rate of 0.0002, for a total of 20 local training epochs. During the generator training stage, synthetic samples of size 600 for each class were fed into each non-dropout client model to supervise the generator training. The hyper-parameter λ in Eq. (3.9) was set to 0.5 for each dataset, and the server aggregated the loss from different client models based on the proportion of samples in the classes of each client.Finally, the trained generator

and the aggregated global model were broadcasted to each client to complete personalized model training. The hyperparameter β in Eq. (3.12) was set to 0.01 for the CIFAR-10 and CIFAR-100 datasets, and 0.1 for the EMNIST and Fashion MNIST datasets.

Table 3.2 Comparison with SOTA FL algorithms in Full Participation settings

Dataset	Client Num	Heteroge. Setting	Test Accuracy(%)						
			Local	FedAvg	FedProx	SCAFFOLD	FedGen	FedDF	AP-FL
CIFAR10	5	α =0.01	15.71 ± 0.39	43.83 ± 0.90	51.48 ± 1.21	54.47 ± 0.99	28.66 ± 1.19	44.66 ± 1.40	**61.84 ± 1.75**
		α =0.05	28.72 ± 0.32	61.61 ± 1.36	60.08 ± 3.19	64.28 ± 1.43	41.86 ± 0.47	60.27 ± 0.39	**65.14 ± 0.32**
		α =0.1	33.00 ± 1.16	65.77 ± 1.77	65.07 ± 0.40	67.37 ± 1.02	46.61 ± 2.88	64.58 ± 0.95	**69.46 ± 0.18**
	10	α =0.01	15.76 ± 0.04	38.79 ± 4.97	45.98 ± 0.58	46.09 ± 2.50	26.67 ± 2.50	37.06 ± 1.26	**56.28 ± 0.51**
		α =0.05	24.95 ± 0.87	52.96 ± 0.24	51.68 ± 0.32	53.01 ± 0.74	27.51 ± 1.76	52.07 ± 1.97	**58.73 ± 1.75**
		α =0.1	35.04 ± 1.54	58.15 ± 0.94	56.36 ± 0.26	60.04 ± 1.08	43.08 ± 0.55	57.89 ± 1.00	**61.39 ± 0.28**
CIFAR100	5	α =0.01	13.89 ± 0.34	30.16 ± 0.42	29.28 ± 0.13	33.80 ± 1.19	30.04 ± 2.14	30.47 ± 1.43	**35.28 ± 4.21**
		α =0.05	24.53 ± 0.44	32.19 ± 2.13	34.58 ± 1.05	36.74 ± 0.41	32.17 ± 1.21	35.34 ± 1.32	**38.47 ± 0.42**
		α =0.1	25.23 ± 0.38	34.63 ± 0.32	34.89 ± 0.49	37.18 ± 1.73	34.93 ± 1.03	36.84 ± 2.41	**39.95 ± 1.45**
	10	α =0.01	14.47 ± 1.53	28.37 ± 1.10	28.11 ± 1.03	30.32 ± 1.05	28.18 ± 0.58	28.39 ± 2.65	**31.74 ± 1.52**
		α =0.05	23.40 ± 0.28	30.01 ± 0.56	32.16 ± 0.50	33.49 ± 0.73	29.55 ± 0.41	33.12 ± 1.74	**35.86 ± 0.47**
		α =0.1	24.09 ± 1.53	32.34 ± 0.65	32.78 ± 0.13	34.95 ± 0.58	31.88 ± 0.65	33.51 ± 1.24	**36.74 ± 0.44**
EMNIST	5	α =0.01	24.36 ± 0.23	86.56 ± 0.95	85.43 ± 0.61	85.30 ± 0.37	82.41 ± 2.34	88.06 ± 0.37	**89.07 ± 1.26**
		α =0.05	33.20 ± 0.29	89.33 ± 0.16	87.97 ± 0.40	89.22 ± 0.21	86.86 ± 0.89	89.27 ± 0.27	**91.24 ± 0.52**
		α =0.1	36.86 ± 0.26	90.85 ± 0.31	89.36 ± 0.55	**91.88 ± 0.46**	90.12 ± 0.63	90.32 ± 0.26	91.60 ± 0.16
	10	α =0.01	13.38 ± 0.26	65.98 ± 3.95	77.09 ± 1.49	69.23 ± 1.47	66.74 ± 8.45	65.72 ± 1.33	**82.48 ± 0.43**
		α =0.05	19.03 ± 0.03	82.32 ± 0.35	83.23 ± 0.71	84.06 ± 1.24	81.05 ± 1.69	83.19 ± 1.27	**85.27 ± 0.16**
		α =0.1	32.22 ± 0.02	88.69 ± 0.47	87.68 ± 0.47	87.88 ± 0.81	88.45 ± 0.49	**89.12 ± 0.16**	88.94 ± 1.20
Fashion MNIST	5	-	49.15 ± 0.19	88.28 ± 0.89	87.68 ± 0.89	88.60 ± 1.20	87.05 ± 2.21	88.79 ± 0.95	**89.36 ± 0.58**
	10	-	41.61 ± 0.73	85.94 ± 1.51	85.74 ± 0.16	85.50 ± 0.45	85.23 ± 2.44	83.97 ± 3.62	**87.04 ± 0.17**

3.4.2 Experimental Results

Comparison with SOTA in Full Participated Settings: Table 3.2 presents a comprehensive evaluation of the accuracy of various algorithms on different Dirichlet non-IID distributions, demonstrating that our proposed AP-FL framework surpasses most state-of-the-art (SOTA) methods, particularly in highly heterogeneous scenarios, such as alpha = 0.01 or 0.05. Furthermore, we increased the skewness of label distribution between clients by expanding the number of clients. As Table 1 demonstrates, even in this 10-client scenario, AP-FL maintains superior performance over other algorithms. Compared to FedGen, which directly feeds synthetic data into the global model, our approach can effectively alleviate the impact of spurious data on model performance through the decoupled model interpolation technique. Additionally, Figure 4.5 shows the comparative performance of all algorithms at varying degrees of label distribution skewness, with AP-FL demonstrating more consistent and stable performance as data heterogeneity increases.

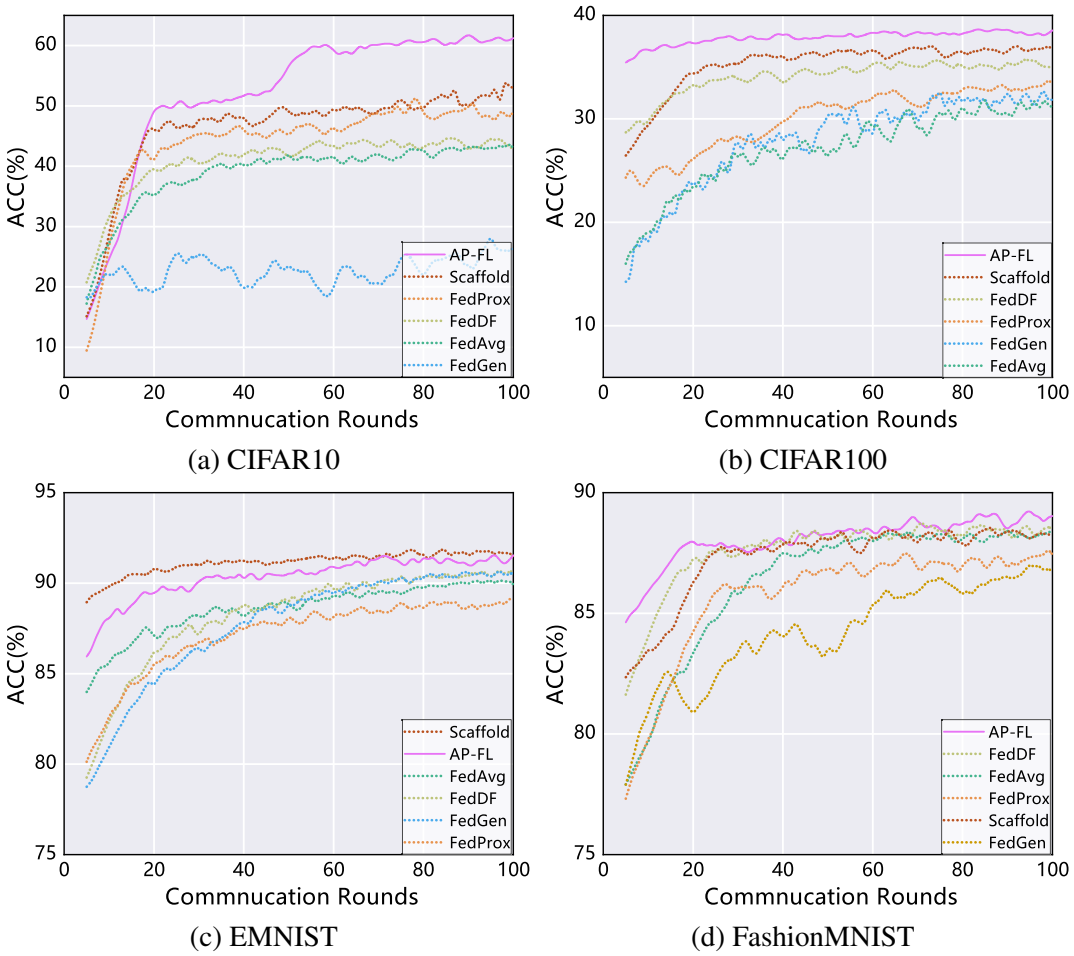


Fig. 3.4 Evaluation of model performance on four datasets and five clients, the α set to 0.01, 0.05, 0.1 and $-$, respectively, for CIFAR10, CIFAR100, EMNIST, and FashionMNIST.

Comparison with Existing Works in Dropout Settings. Table 3 presents the results of the comparison between our proposed AP-FL framework and the Local and FedAvg-FT baselines. Our Personalized Model trained with AP-FL on CIFAR10, EMNIST, and Fashion MNIST datasets outperforms the Local model and FedAvg-FT in most cases, indicating the effectiveness of our approach in assisting dropout clients to train their own Personalized model. However, the performance of AP-FL on CIFAR100 is slightly behind FedAvg-FT. We attribute this to the fine-grained nature of the dataset, which poses a challenge for language models like CLIP/BERT to generate semantically distinctive information for subclasses under certain categories, leading to poor quality of generated unseen synthetic samples. In summary, our findings suggest that when clients with monopolistic categories drop out, AP-FL presents a more competitive alternative to training a local model or fine-tuning a global model for those dropout clients.

Table 3.3 Comparison with FedAvg in Dropout settings. 'MC' represent the missing classes due to dropout client with minority classes.

Dataset	CIFAR10		CIFAR100		EMNIST		Fashion MNIST	
MC(%)	10%	20%	10%	20%	10%	20%	10%	20%
Local	29.47 ± 1.69	26.74 ± 0.49	22.61 ± 1.52	21.97 ± 1.10	30.15 ± 2.14	29.73 ± 1.19	47.51 ± 1.06	47.63 ± 0.98
FedAvg-FT	31.43 ± 0.58	29.82 ± 1.19	$\mathbf{23.15 \pm 1.32}$	$\mathbf{24.73 \pm 1.43}$	34.81 ± 2.41	34.05 ± 0.56	51.78 ± 0.37	51.96 ± 0.74
AP-FL	$\mathbf{34.18 \pm 0.49}$	$\mathbf{32.97 \pm 0.26}$	23.12 ± 1.55	24.65 ± 1.08	$\mathbf{37.91 \pm 0.71}$	$\mathbf{36.29 \pm 0.28}$	$\mathbf{58.97 \pm 0.58}$	$\mathbf{56.83 \pm 0.32}$

3.4.3 Ablation Study

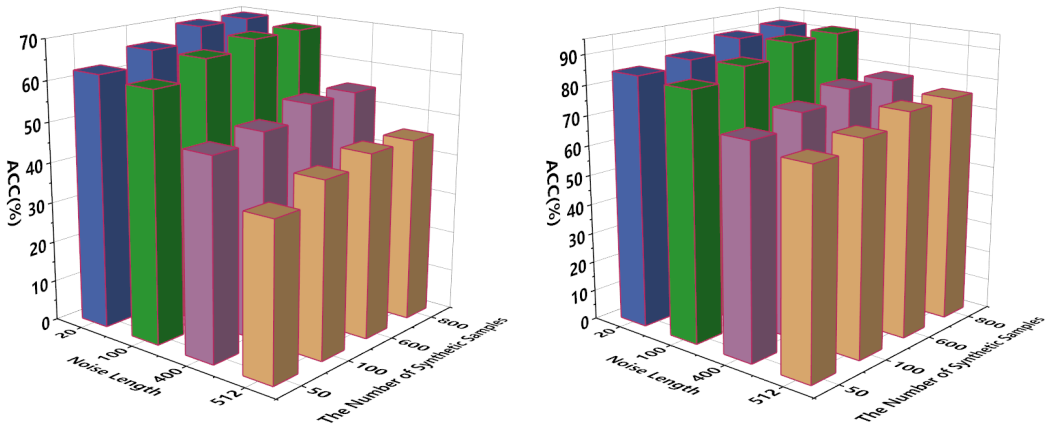
Effect on Different Semantic Information. In our ablation study, we investigated the impact of using different semantic embeddings in the dropout settings. Specifically, we evaluated our model with three types of semantics, namely word2vec (W2V), BERT, and CLIP. As shown in Table 3.4, the results with all three types of semantics are comparable, indicating the robustness of our model to different semantic embeddings. However, we observed that our model achieved the best performance with CLIP representation, suggesting the effectiveness of using CLIP as the semantic embedding.

Table 3.4 Analysis of synthetic features on different types of semantic embedding in the dropout settings, where $\mathscr{A}_n$ corresponds to the accuracy of the friend model tested on non-dropout clients, and $\mathscr{A}_d$ corresponds to the accuracy of the friend model tested on dropout clients.

Dataset	CIFAR10		CIFAR100		EMINIST		FashionMNIST	
Domain	$\mathscr{A}_n$	$\mathscr{A}_d$	$\mathscr{A}_n$	$\mathscr{A}_d$	$\mathscr{A}_n$	$\mathscr{A}_d$	$\mathscr{A}_n$	$\mathscr{A}_d$
W2V	50.74	41.32	18.92	15.49	59.86	44.25	62.14	50.43
BERT	55.92	51.63	21.76	22.05	65.14	51.27	73.31	52.84
CLIP	58.21	49.79	25.62	26.43	70.72	54.60	74.16	58.63

Effect on the Hyper-Parameters. We performed two ablation studies on the CIFAR10 and EMNIST datasets to investigate the impact of two hyper-parameters, namely noise dimension and the number of synthetic samples, on the performance of the friend model in the Full Participation Setting. The results are presented in Figure 6.5. Four different noise dimensions, i.e., 20, 100, 400, and 512, were chosen to illustrate the relationship with the performance of the friend model. We observed that the performance decreases with increasing noise dimension on both datasets, indicating that high-dimensional noise may lead to significant interference. Regarding the number of synthetic samples, we varied the number of synthetic samples from 50 to 1000 in the experiments. As shown in Figure 6.5, the accuracy of the friend model on both datasets remains stable once the number of samples exceeds 600. We attribute this phenomenon to the fact that a lack of false data results in poor performance of

the friend model due to the insufficient number of samples, whereas an excessive amount of false data can lead to a limited diversity of false data, which can be a bottleneck for the performance of the friend model.



(a) CIFAR100 (b) FashionMNIST

Fig. 3.5 The impact of noise dimension and the number of synthetic samples on the performance of the friend model with $\alpha = 0.1$.

3.5 Conclusion

In this work, we introduce the Asynchronous Personalized FL framework (AP-FL), which addresses the non-IID and dropout issues in FL by training a semantic generator to capture the global data distribution from non-dropout clients. This generator is then used to generate synthetic samples for each non-dropout client, aiding in the establishment of a personalized model to mitigate the client drift issue. Additionally, AP-FL leverages semantic information and the Zero-Shot learning paradigm, allowing the generator to generate previously unseen samples for dropout clients with monopoly classes and enhance data diversity for training personalized models in dropout clients. Our experiments demonstrate that AP-FL outperforms state-of-the-art methods for addressing non-IID and dropout issues in FL.

Epilogue

In addressing Research Question 1.1, our work presents a novel Personalized Federated Learning framework grounded in model interpolation to effectively tackle the challenges associated with non-IID data distributions. By introducing advanced strategies that leverage

generated data, our approach mitigates the adverse effects of non-IID data, enhancing both the robustness and learning efficiency of personalized models across heterogeneous client datasets. This method ensures that models trained within federated environments are better adapted to the unique data distributions of individual clients, promoting equitable learning outcomes and reducing biases caused by skewed data distributions.

Regarding Research Question 1.2, we propose innovative solutions inspired by zero-shot learning to address client dropout issues in federated learning. By treating the data from dropout clients with unique categories as unseen classes, we utilize semantic embeddings generated by models like CLIP to create a global generator. This generator compensates for missing data caused by client dropouts by synthesizing data for both seen and unseen classes, maintaining the integrity of the global model's training process. Furthermore, the global generator provides essential support for the reintegration of dropout clients, enabling them to continue their training seamlessly when they rejoin the federated learning process. Through these mechanisms, our approach minimizes the negative impact of client dropouts and ensures continuity and completeness in the federated learning workflow.

Chapter 4

Community-Aware Federated Video Summarization

Prologue

In the previous chapter, we introduced Asynchronous Personalized Federated Learning (AP-FL), addressing key challenges such as non-IID data and client dropouts. Building on this, we now explore its application in video summarization, a domain where data privacy and efficiency are paramount.

The rapid growth of video content demands efficient summarization techniques for better user engagement and information retrieval. However, the sensitive and personalized nature of video data presents significant privacy risks, especially in distributed environments like federated learning. This makes video summarization an ideal case to evaluate privacy-preserving frameworks like AP-FL.

Video summarization poses unique challenges—handling large, heterogeneous datasets while ensuring privacy—which align with the core goals of federated learning. To address these, we introduce Community-Aware Federated Video Summarization (CFed-VS). CFed-VS leverages community-aware clustering and frame-based aggregation to adapt to the diverse nature of video data, preserving privacy while delivering efficient summaries.

By applying AP-FL to video summarization, CFed-VS demonstrates how federated learning can manage sensitive and diverse data in real-world applications. This chapter furthers our objective of developing secure, efficient, and user-focused machine learning solutions for privacy-sensitive tasks.

Declaration: This chapter is a modified version of **"Community-Aware Federated Video Summarization"**, published in IEEE International Joint Conference on Neural Networks (IJCNN), 2023.

4.1 Introduction

With the tremendous growth of video material, automatic tools for understanding and analyzing video content have become an increasingly urgent need. Recent statistics have shown that it will take more than 82 years for a person to watch all videos uploaded to YouTube per day [116]. A promising remedy is that automatic video summarization can enable human users to quickly identify the key content of videos and accelerate knowledge gain and information retrieval. Such a technology has been applied in many scenarios, such as fast indexing and online video recommendation. However, to provide information-rich video summarization and satisfy the wide variety of user needs, existing approaches rely on the large-scale collection of video data and important score annotations to train a robust model. Increasing awareness and concerns about privacy restrictions, e.g., the EU's General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA), have become one of the largest challenges in this domain. Moreover, huge communication costs are incurred during data transmission, which also impedes the development of video summarization technologies.

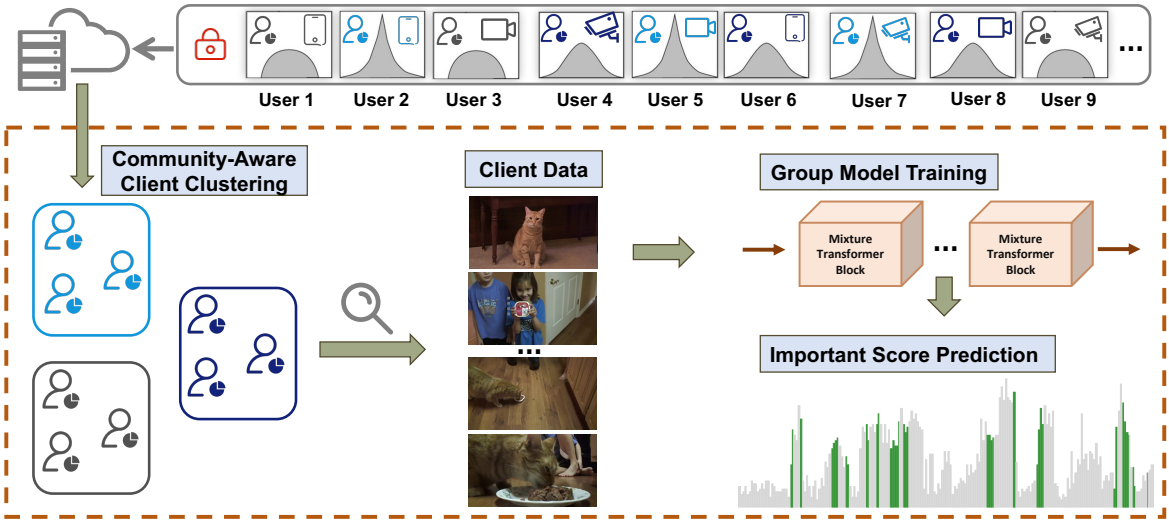


Fig. 4.1 Community-Aware Federated Video Summarization aims to deploy large-scale VS task training when video data are distributed on edge devices. Based on the similarity of data distribution across clients, the server will cluster clients before FL model training, and then maintain multi-group models to address statistical heterogeneity challenges.

As a burgeoning machine learning scheme, Federated Learning [117] aims to tackle the problem of data islands while preserving the privacy of data. The key idea of FL is to jointly train global models across various edge devices without collecting any private raw data from the client, thus effectively alleviating user concerns about data privacy and reducing the high costs associated with transmission. Along with its promising prospect, research on FL faces key challenges from high statistic heterogeneity [34]. Concretely, Federated learning relies on Stochastic Gradient Descent (SGD). Jointly training a global model with the Independent and Identically Distributed (IID) statistics from various clients is equivalent to the IID sampling of training data in a centralized training paradigm, which ensures the stochastic gradient is an unbiased estimate of the total gradient during FL training [118]. FL has been applied in the smart city [81], recommender system [119], and healthcare [83] fields. Despite the successful applications of FL in the above areas, introducing FL to solve data privacy problems and high transmission costs in traditional VS tasks is not trivial. In the real world, data distribution can easily appear in the distribution of non-IID, and video data has an even stronger bias and diversity according to the photographer's preference. Modeling on the non-IID data with FL paradigm will lead to client drift issues [87], which will lead to performance degradation and slow convergence speed of the global model.

A unique property for VS problems is the user community heterogeneity due to the diverse user profile, preferences and behavior. To tackle problems due to statistical heterogeneity in FL, recent attempts [40–42] aim to cluster the clients based on model parameters or gradients and maintain a multi-group model. Nevertheless, the server may be required to wait for extra communication rounds before receiving parameters of the client model with significant changes to calculate the similarity for the clustering procedure, which will lead to the deterioration of model training efficiency and an increase in communication costs. Wang et al. [98] proposed a novel data-driven approach to calculate the similarity of client data distribution, in which clients are grouped based on their similarity in two types of summaries of client data distribution: label distribution, and conditional features distribution. However, it is unrealistic to apply the proposed methods [98] directly to cluster clients in VS tasks, since some video data lack specific categories, using the average feature is also impractical due to the different lengths of the videos.

To our best knowledge, this is the first work to explore the feasibility of the Federated Learning Video Summarization (FLVS) task, and we first established the baseline of FedAvg [117] in FLVS. According to our initial observations and analysis, we proposed a technical roadmap with three key directions for the FLVS problems: 1) In contrast to traditional FL using sample-based aggregation, we explore the **Frame-Based** FedAvg in FLVS tasks so that the length of video is taken into account when assigning weight contributions of client

models. 2) We observe that the community factor is the key impact on the heterogeneous data distribution. A novel **Community-Aware** Clustering FL framework for Federated Video Summarization (CFed-VS) is thus proposed, which *clusters clients based on the relative distance between the data distribution of each client*, as shown in Figure 4.1. It is worth noting that our CFed-VS requires only one-off clustering operation compared to traditional clustering-based FL training, which improves the training efficiency of global models. 3) We then propose the **Mixture Transformer** for obtaining better model generalization in the non-IID setting for learning time-series data. In summary, the key contributions of our work are as follows:

- We propose a more effective frame-based aggregation method of FedAvg for video-related tasks, and systematically analyze the assignment of client model contribution.
- A novel clustering federated framework is proposed to leverage the relative distance between the data distributions of each client, to tackle the challenge of heterogeneous data and reduce the computation cost during the clustering process.
- Mixture Transformer is proposed to enhance model generalization in the non-IID setting, and extensive experiments demonstrate state-of-the-art performance on SumMe and TVSum datasets.

4.2 Related Work

4.2.1 Video Summarization

Video summarization is one of the most important directions in video recognition [120–124] to generate [125, 126] a short video clip while keeping the main content or stories of the original video [127, 120]. Recently, several video summarization approaches have been proposed, and they can fall into two broad categories. One of them refers to unsupervised learning, which uses manually designed criteria to prioritize and select frames or subshots from original videos [127, 128]. Another one is supervised learning, which utilizes human-edited examples to learn how to summarize novel videos [129, 130]. Also, some LSTM-based deep learning approaches have been proposed for both supervised and unsupervised video summarization. Mahasseni et al. [131] specified a generative adversarial framework that consists of the summarizer and discriminator for unsupervised video summarization. Wang *et al*.[132] proposed a novel model named Dual Mixture Attention (DMASum) with meta-learning, which solved the softmax bottleneck problem in video summarization.

4.2.2 Federated with Statistic Heterogeneity

Statistic heterogeneity (also named non-IID) is one of the major challenges in federated learning. The widespread aggregation strategy in federated learning, FedAvg [117] suffers performance deterioration on non-IID due to client drift issues [87]. To address this problem, a line of research focuses on learning a single global model under the non-IID setting [32, 133, 35]. For example, FedProx [32] adds a proximal term to the local objective of the client to effectively limit the impact of abnormal local model updating. Another line of research overcomes this problem via personalized federated learning (PFL) [48, 134, 39], which seeks to personalize the global model for each client. PFL has been adopted in many approaches, including model-agnostic meta-learning [48], model regularization [134], and multi-task learning [39]. Cluster-Based Federated Learning (CFL) [40–42] incorporates personalization at the group level while keeping the benefits of PFL. Prior works cluster clients based on the similarity of client model parameters or gradients. However, obtaining a significant change in the client's parameter or gradient, requiring the server to wait for additional communication rounds, greatly reduces the training efficiency. To this end, we develop a time-series similarity method to generate a summary of data distribution under the privacy specification of FL, then the server clusters the client based on the summary before initiating FL training.

4.2.3 Vision Transformer

Following Transformer in NLPs [135], Vision Transformer (ViT) [136–138] has made great successes in various vision tasks, including object detection [139], semantic segmentation [140, 141], action recognition [142], and so on. For example, Li *et al.*[143] propose a hierarchical Transformer (Swin Transformer) whose representation is computed with Shifted windows that can bring greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. Moreover, the great success of image Transformers has led to the investigation of Transformer-based architectures for video understanding tasks [142]. For instance, UniFormer [144] integrates the merits of 3D convolution and spatio-temporal self-attention in a concise transformer format, and achieves a preferable balance between computation and accuracy.

4.3 Methodology

4.3.1 Rethinking Federated Learning in Video Summarization

Recently, many tasks have successfully adapted federated learning (FL) paradigms, while few studies [145] have examined FL applications on video data, especially on video summarization. To explore the feasibility of federated learning on video summarization, we first investigate the characteristics of FedAvg [117], the widely used federated learning method. Then we provide an in-depth analysis of federated model training on video summarization datasets.

For a learning problem, we define $f(\theta) = \mathscr{L}(x, y, \theta)$ as the loss function representing the error in the model's predictions on input pair (x, y), where θ represents the model parameters. In federated learning, the goal is to train a global model collaboratively across multiple clients, each having its own data distribution $\mathscr{P}_k$. Thus, the global objective is to find the optimal parameters θ^* that minimize the total loss across all clients, represented as follows:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta),\ F_k(\theta) = \frac{1}{n_k} \sum_{i \in \mathscr{P}_k} f_i(\theta)\ , \tag{4.1}$$

where n_k denotes the number of samples on client k, and $F_k(\theta)$ represents the local objective function for client k. The global objective is to minimize the weighted sum of all local objectives, where the weights are determined by the proportion of data on each client. This equation defines the optimization goal for the global model in a traditional federated learning setup. Conventional federated learning methods, like FedAvg, optimize Eq. 4.1 with the following steps. (i) At each communication round t, the server randomly selects K clients available for training, then sends the global model $\hat{\theta}^t$ to the selected clients and deploys it as $\theta_k^{(t)}$ (ii) Each selected client then trains its model $\theta_k^{(t)}$ locally with its own data distribution $\mathscr{P}_k$ for E_{local} epochs. (iii) The server waits until all selected devices have uploaded corresponding parameters $\theta_k^{(t+1)}$ to aggregate the new global model via $\hat{\theta}^{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t+1)}$. The above process will be repeated until the model reaches convergence.

Frame-based FedAvg in Video Summarization. According to the FedAvg algorithm, the weight contribution of the client model is based on the number of samples, *i.e.*$\frac{n_k}{n}$. However, we assume that directly applying this sample-based FedAvg to video summarization is impractical, as the length of the video is also important for model training in video understanding tasks [146]. To this end, we propose to apply video frames $\frac{v_k}{v}$ instead of video

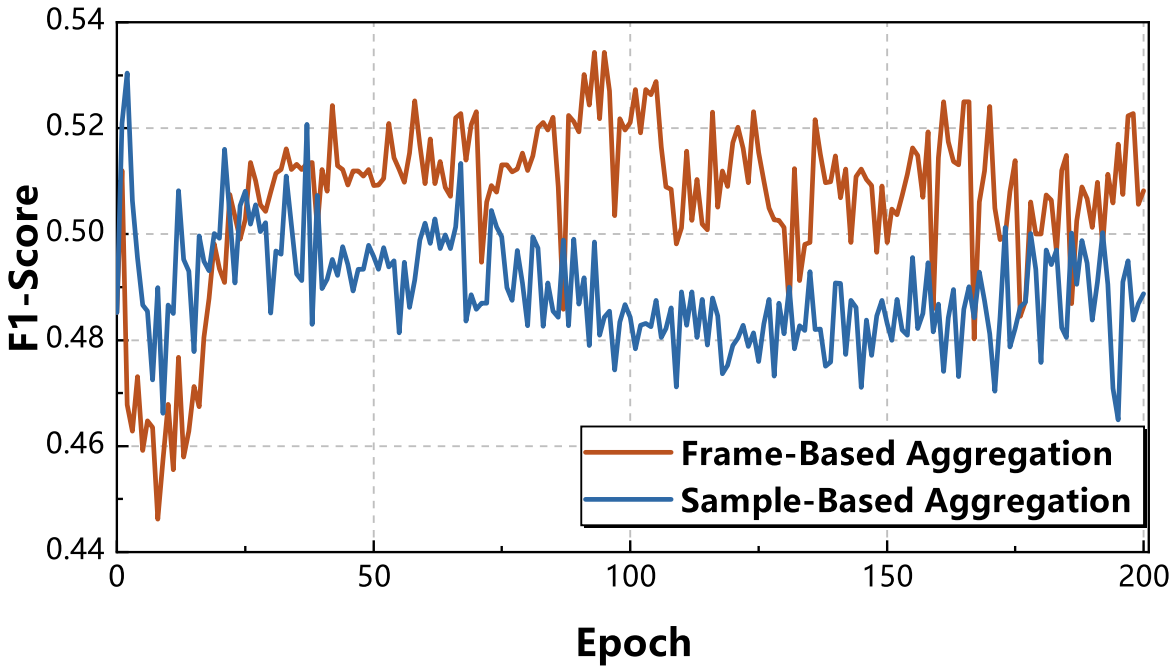


Fig. 4.2 Comparison of two weight aggregation strategies on the TVSum dataset. Experiments are conducted in IID data distribution with the same settings.

samples $\frac{n_k}{n}$ as:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{N} \frac{v_k}{v} F_k(\theta) \,, \tag{4.2}$$

where v_k is the total number of frames that resides in k-client, and v represents the total number of frames across all clients. To verify our assumption, we conducted different aggregation strategy experiments by using vsLSTM [1] on the TVSum dataset, which is shown in Figure 4.2. It can be observed that the frame-based aggregation can quickly outperform the baseline and converge to more reliable performance. In this analysis, the frame-based aggregation strategy is 1.2% higher than the sample-based aggregation strategy (conventional FedAvg aggregation strategy). The proposed frame-based aggregation is particularly useful when the length of the user video is different.

4.3.2 Non-IID Data Distribution Analysis

To analyze the data distribution on video summarization datasets by applying federated learning, we simulate non-IID and IID settings by forcing each client to have limited classes, as shown in Table 4.1. We follow the setting in [40] and simulate ten clients to run the global model within 200 communication rounds. Each participating client runs the global model with 20 local epochs. We then obtain three folds of insights as follows.

Table 4.1 F1-score (%) of different data distribution on both TVSum and SumMe datasets, using vsLSTM [1] baseline. “Max-Min F1-Score" represents the difference between the best and worst results tested by the global model on all clients. “Mean F1-Score" represents the average F1-Score across all clients. “# Round to Reach Target F1-Score" donates the communication rounds of the global model to reach target F1-Score. The “Target F1-Score" was chosen 48% and 30% for TVSum and SumMe datasets separately.

Dataset	**Federated**				**Centerlized**
	# Classes	Max-Min F1-Score	Mean F1-Score	# Round to Reach Target F1-Score	F1-Score
TVSum	1	10.98	49.29	75	54.27
	4	5.21	52.08	46	
	10 (IID)	2.23	53.14	17	
SumMe	1	12.19	33.62	64	37.72
	3 (IID)	4.87	36.48	18	

- Compared to the centralized paradigm, the performance of using FedAvg (frame-based) would not decrease too much, which also verifies the feasibility of using federated learning in video summarization tasks.
- With the decrease of classes in each client, the performance of the global model will decrease from 53.14% to 49.29%, and the performance of “round to reach" will increase from 17 to 75. These results indicate that higher data heterogeneity would affect the model performance and convergence speed.
- The value of “Max-Min F1-Score" increased with the decrease of categories in each client, which indicates that the global model is difficult to generalize on all the clients in non-IID settings.

Therefore, using federated learning in video summarization tasks will face high data heterogeneity (non-IID) challenges, which affects the global model in terms of performance and convergence speed. To address the above challenges, we propose the Community-Aware Clustering Federated Video Summarization strategy by clustering the clients with similar data distribution, and training corresponding group models.

4.3.3 Community-Aware Federated Video Summarization

There has been significant research investigating high data heterogeneity challenges [147, 34, 41] due to the community diversity of users. Specially, cluster-based federated learning clusters the clients based on the similarity of the model parameters [40–42] or data distribution [98] across different clients recently. Parameters-based clustering algorithm requires the server to consume additional communication rounds to obtain significantly varying gradients or parameters, thereby reducing model training efficiency and increasing communication costs. To this end, we propose a data-driven approach via leveraging the relative distance between the data distribution of each client, and clients with a close distance of data distribution can be clustered as a community before FL model training.

Community Distribution Estimation. Due to the data privacy policy in federated learning, it is not allowed to directly calculate the distance between any two private data distributions among clients. We thus set a proxy sample from another public dataset to collect the comparison of distances simultaneously. By calculating the distance between proxy samples and the center of all training samples of each client, we can obtain the essential information on the data distribution in each client, which can be regarded as the summary of data distribution for all clients. Due to the property of video data (time-series), we adopt the widely used Dynamic Time Warping (DTW)[148] as the distance measurement.

Concretely, the centre server first broadcast a single proxy sample $\mathbf{x}_p$ to each available client k. Then the client calculates the pair-wise distance $\mathscr{D}[\mathbf{x}_u, \mathbf{x}_p]$ between each training sample $\mathbf{x}_u$ and the proxy sample $\mathbf{x}_p$ via the DTW distance function, where $\mathbf{x}_u$ is the sample in the local dataset $X_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_u, \ldots, \mathbf{x}_U\}$. Finally, we estimate the distance d^k_{center} from the center of data distribution of k-client to the proxy sample $\mathbf{x}_p$. We regard d^k_{center} as the summary of each client's data distribution, and the detailed implementation can be seen in Algorithm 1.

Clustering Procedure. Once each device sends its summary of data distribution d^k_{center} to the central server, the server calculates the pair-wise distance between any two clients' summary via L1 distance, i.e. $|d^k_{\text{center}} - d^q_{\text{center}}|$, where k, q denotes the different client. We store all pair-wise values of various clients to form the distance matrix M_d. Then we adopt the K-Means algorithm[149] to cluster the client into the m group based on the distance matrix. An overview of the clustering procedure is also described in Algorithm 1 and an illustration of the clustering procedure is shown in Fig 4.3. After completing the clustering operation, we train m cluster-wise models via the proposed Frame-Based FedAvg.

Our proposed CFed-VS strategy allows for a one-off clustering of participating clients into various groups before starting federated training. In the event that new clients join, the server can send only the proxy sample and assign these clients to the appropriate group based

Algorithm 1 Community-Aware Federated Video Summarization

Input: $\mathscr{C} \leftarrow$ Client groups;
$X_k \leftarrow$ Dataset of k-client with U samples;
$M_d \leftarrow$ the distance matrix of the data distribution;
$DTW() \leftarrow$ dynamic time warping function;
Output: Uploaded group model parameters $\theta_{c,t}$

procedure CFED-VS
  Server broadcasts $\mathbf{x}_p$ to all K clients
  for $k \in K$ **do**
    $d^k_{\text{center}} = \frac{1}{U}\sum_{u=1}^{U} \text{DTW}(\mathbf{x}_u, \mathbf{x}_p)$
    Client k uploads d^k_{center} to server
  end for
  $M_d = \{|d^k_{\text{center}} - d^q_{\text{center}}|\} \mid k,q \in K, k \neq q$
  $\mathscr{C} \leftarrow \mathscr{F}_{K-means}(M_d, m)$
  for $c \in \mathscr{C}$ **do**
    $\theta_{c,0} \leftarrow \theta_0$
    for `each round` $t = 1,2,\ldots,T$ **do**
      $\theta_{c,t+1} \leftarrow$ `FEDAVG(`$\theta_{c,t}$`,` K_c`)`
    end for
  end for
end procedure

function DTW($\mathbf{x}_1$, $\mathbf{x}_2$)
  $l_1 \leftarrow$ `length of` $\mathbf{x}_1$; $l_2 \leftarrow$ `length of` $\mathbf{x}_2$
  Initialize $\mathscr{D}$ as a matrix of size (l_1+1, l_2+1) with ∞
  $\mathscr{D}[0,0] \leftarrow 0$
  for $i = 1,2,\ldots,l_1$ **do**
    for $j = 1,2,\ldots,l_2$ **do**
      $cost = d(\mathbf{x}_1[i], \mathbf{x}_2[j])$
      $\mathscr{D}[i,j] = cost + \min(\mathscr{D}[i-1,j], \mathscr{D}[i,j-1], \mathscr{D}[i-1,j-1])$
    end for
  end for
  return $\mathscr{D}[l_1, l_2]$
end function

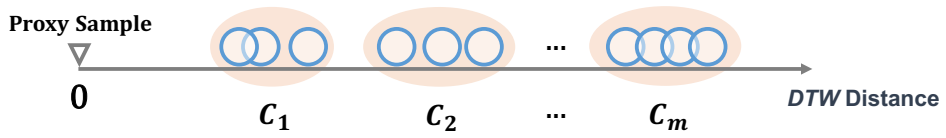


Fig. 4.3 The illustration of clustering client methods in CFed-VS. The blue circle denotes the distance from the proxy sample to the data centre of various clients. The orange oval represents different client groups and indexes by $\mathscr{C}_m$.

on their feedback summary. Compared to other grouping methods based on data distribution, our approach does not reveal the user's data category information and only requires a certain amount of computing resources on the client side.

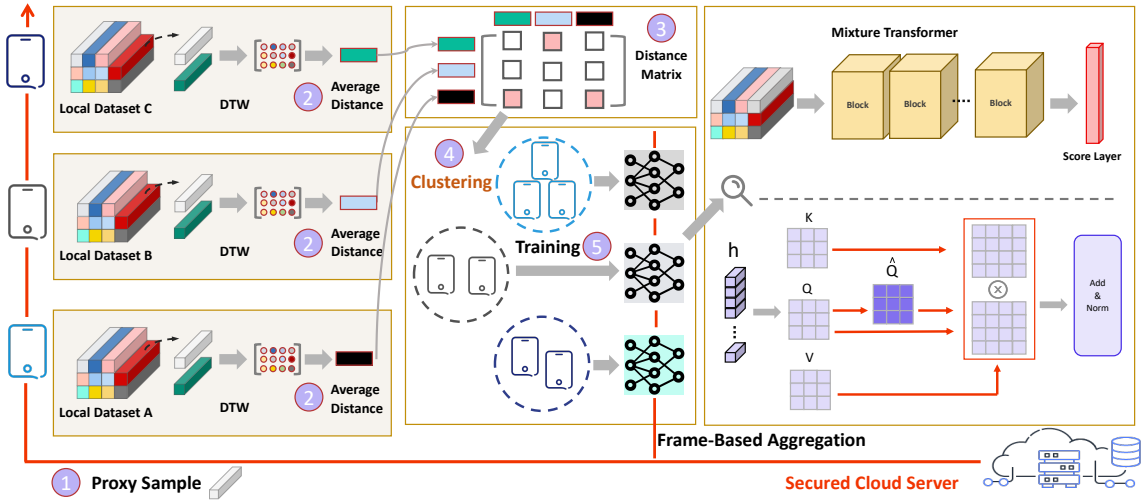


Fig. 4.4 Overview of the Community-Aware Federated Video Summarization system. The secured cloud server firstly clusters clients into multi-groups based on the similarity of data distribution before the FL training procedure. Then the Mixture Transformer model can be deployed to each client to carry out training.

4.3.4 Mixture Transformer

Even though the proposed CFed-VS provides a solution to non-IID performance decline, there is still a performance gap when using conventional video summarization models with a non-IID data setting. We then focus on improving the model generalization in the non-IID setting. Considering the outstanding performance of the Transformer [135] for learning time-series data, the global receptive field of the self-attention mechanism can help the model focus on summarizing global information of a video sequence, which is suitable for the video summarization task. Therefore, we propose the Mixture Transformer to address the above challenges.

Transformer Block. In a basic Transformer block, the queries, keys, and values $Q = HW_q$, $K = HW_k$, and $V = HW_v$ are linear projections of the input frame feature H with $Q, K, V \in \mathbb{R}^{N \times d}$, where N and d denote the frame number and channel dimension. The process is defined as:

$$\mathscr{A}(K,Q,V) = \mathscr{F}_{Softmax}(\frac{QK^T}{\sqrt{D_a}})V \ , \tag{4.3}$$

$$H_a^l = MLP(LN(\mathscr{A}^l) + \mathscr{A}^{l-1}) + H_a^{l-1}, \tag{4.4}$$

where $\mathscr{A}$ denotes the attention map, which is computed as the scaled dot-product function, and l denotes the block number. $\mathscr{A}^{l-1}$ represents the attention map from the previous block (block $l-1$), which is added to the current block's attention map $\mathscr{A}^l$ to incorporate prior attention information. This operation enables the model to progressively refine the attention weights over multiple blocks. $\mathscr{F}_{Softmax}$ denotes the *softmax* function. *LN* and *MLP* represent Layer Normalization and Muti-Layer Perception, respectively. In the context of video summarization, the log probability matrix $\mathscr{A}$ becomes a high-rank matrix when the visual contents are complex and the changes between frames are severe. Such high-rank matrix applied *softmax* function will face the **Softmax bottleneck**, as discussed in [150]. It reflects the circumstance that *softmax* function does not have the capacity to express the true attention distribution when d is smaller than $rank(\mathscr{A}) - 1$. Inspired by the work of [132], we apply the Associated Query $\hat{Q} = tanh(W^{\hat{Q}}Q)$ to capture the second-order changes between queries so that complex video content can be represented in a more smoothed attention representation. Then the attention map in Eq. 4.3 is re-computed as:

$$\hat{\mathscr{A}} = \mathscr{A}(K,Q,V) \cdot \mathscr{A}(K,\hat{Q},V)^T \ , \tag{4.5}$$

where $\hat{\mathscr{A}} \in \mathbb{R}^{T \times T}$, namely mixture attention map. $W^{\hat{Q}}$ is the Associated Query parameter. As $\hat{\mathscr{A}}$ is a non-linear function of the attention distribution, the rank of $\hat{\mathscr{A}}$ can be arbitrarily higher than the standard attention map $\mathscr{A}$, which can be used to alleviate the bottleneck problem.

Overall Framework. The architecture overview is shown in Figure 4.4. Firstly, the secured cloud server sends a proxy sample to each user. According to the received feedback, users with similar distances will be clustered as a community. Each community trains a unique Mixture Transformer model using the proposed Frame-Based Aggregation. New users will be assigned to the correct community by comparing their feedback summary in the distance between the proxy sample and the center of their data distribution.

4.4 Experiment

4.4.1 Experimental Setup

Datasets. We evaluate our model on two public datasets: TVSum [7] and SumMe[6]. TVSum was collected from Youtube, which contains 50 videos in 10 categories. The duration of most videos ranges from 1 to 10 minutes. SumMe includes 25 videos with various holidays, events and sports. The video lengths vary from 1.5 to 6.5 minutes. Both datasets contain annotations labeled for key-frames by 25 human annotators. Considering the non-IID setting is category-based [34], we manually flag each video's category as the absence of detailed categories for each video on two datasets. Furthermore, the limited data provided by TVSum and SumMe cannot meet the splitting method in non-IID setting, thus we used the ball-and-urn technique [151] to split a single video sample into multiple fragments, increase the number of samples in TVSum and SumMe to 150 and 100 respectively, and then we followed Pathological non-IID setting [34] to assign client data.

Evaluation Metrics As for the evaluation metrics for the **VS** task, we used the key-shot-based F-score [130] as the metric, and the converted frame-level importance scores to shot-based summaries for all datasets. The kernel temporal segmentation (KTS) [127], where the method can segment the video into separate intervals in time, was used to change the user annotation from frame to key shot level in our experiment. Then we calculate the harmonic average F-score as the evaluation metric. We also used Kendall's τ [152] and Spearman's ρ[153] correlation coefficients to compare the ordinal correlation between the generated summary and the ground truth. As the metric for **FL**, given numerous devices, we evaluate the corresponding group model based on the client's local test set for the CFL-based framework under the same number of groups. For FedAvg and FedProx, we evaluate the global model on the local test data of all clients.

Implement Details For CFed-VS, we set the total number of clients at ten, and the fraction of clients participating in each round of FL is 0.8. The global communication round T is 200, the learning rate is 0.0001, and the local epoch E_{local}=20. For Mixture Transformer, the 1024 dimensional visual features extracted from the *pool*5 layer of the GoogLeNet [154] are used for training, to be consistent with existing methods. Since cluster structures in the real world may be ambiguous, ignoring the knowledge learned by the group model from other communities will reduce the performance of models trained in a single client cluster[155]. Thus, we also adopt the weight-sharing approach [155]. The proxy sample $\mathbf{x}_p$ is selected by selecting a random test sample in SumMe when performing experiments on TVSum, and vice versa. Besides following previous work, frames feature H are first extracted by I3D [156] whose dimensionality is 1024, and each video is segmented into shots by KTS [157],

which is a widely used video temporal segmentation method in the video summarization task.

4.4.2 Experimental Results

Comparison to FL Methods. As shown in Table 4.2, we compared our proposed CFed-VS with four FL baseline frameworks and found that CFed-VS achieves the best performance both in two datasets. The result shows that IFCA[155], FedGroup[40] and CFed-VS outperform significantly to other frameworks in most non-IID settings. We attribute this to the CFL-based approaches, which group clients with similar data distribution into the same community, so group models can effectively learn common properties from communities to mitigate the challenge of data heterogeneity. As the convergency speed shown in Figure 4.5, our proposed approach has a fast convergency speed compared with IFCA and FedGroup, which shows the efficiency of the CFed-VS approach in FLVS tasks.

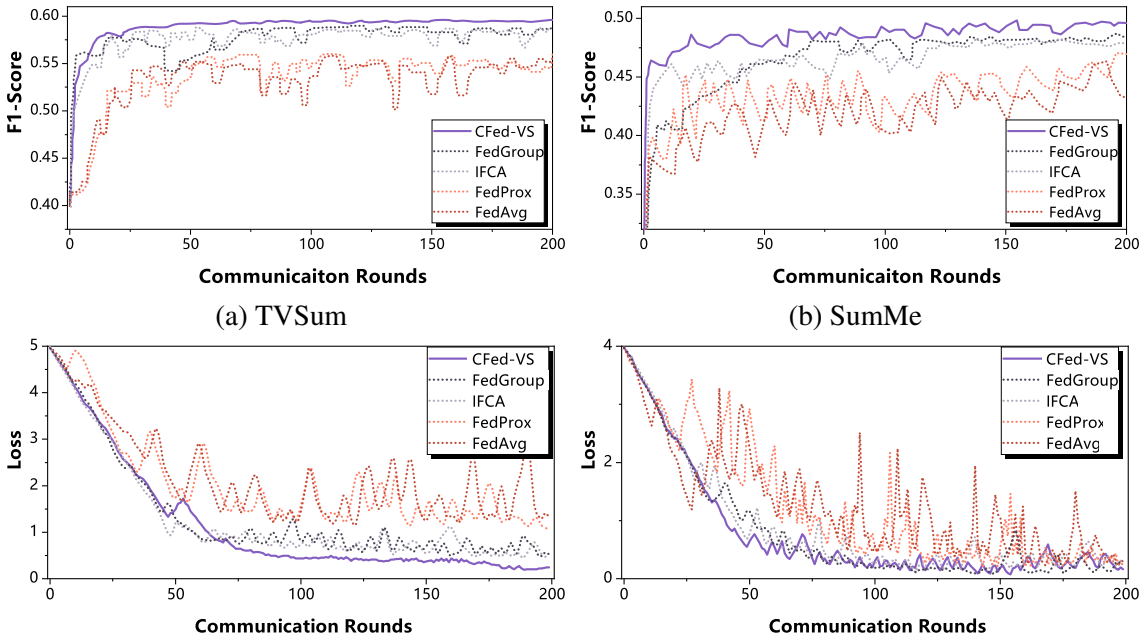
Table 4.2 Comparisons with FedAvg, FedProx, IFCA, FedGroup on TVSum and SumMe dataset.

Dataset	TVSum			SumMe	
# of Classes	1	2	4	1	2
FedAvg[117]	54.23 ± 2.2	54.38 ± 1.5	56.08 ± 0.5	43.78 ± 0.2	49.59 ± 0.5
FedProx[32]	55.65 ± 1.9	54.63 ± 0.6	57.42 ± 1.3	43.62 ± 0.8	50.77 ± 1.2
IFCA[155]	57.29 ± 1.3	57.94 ± 1.7	58.51 ± 1.6	47.23 ± 1.4	51.41 ± 0.5
FedGroup[40]	57.41 ± 1.7	58.39 ± 0.4	58.74 ± 1.9	47.72 ± 1.7	51.59 ± 2.1
CFed-VS	$\mathbf{57.91 \pm 1.4}$	$\mathbf{58.86 \pm 0.4}$	$\mathbf{59.98 \pm 2.7}$	$\mathbf{48.20 \pm 0.8}$	$\mathbf{52.13 \pm 0.5}$

Table 4.3 F1-score (%) of DMASum with state-of-the-art approaches on both SumMe and TVSum dataset.

Method	SumMe	TVSum
DPP-LSTM [1]	38.6	54.7
SUM-GAN [158]	41.7	54.3
Cycle-SUM [159]	41.9	57.6
DMASum [132]	54.3	61.4
SumGraph [160]	51.4	**63.9**
RSGN [161]	45.0	61.0
Mixture Transformer	**55.1**	63.8

Comparison to VS Methods. Our model comparison with state-of-the-art VS methods is summarized in 4.3 and 4.4. From the result of 4.3, it can be seen that Mixture Transformer



(a) TVSum (b) SumMe

(c) TVSum (d) SumMe

Fig. 4.5 Evaluation of model performance on TVSum and SumMe under 2-class non-IID and 1-class non-IID separately.

can achieve competitive performance on both datasets, which indicates that pure Transformer structure can obtain outstanding performance than other models. From 4.4, we can see the correlation coefficients given by DMASum are significantly higher than other state-of-the-art models, which verify that the mixture attention mechanism itself is capable of improving model generalization.

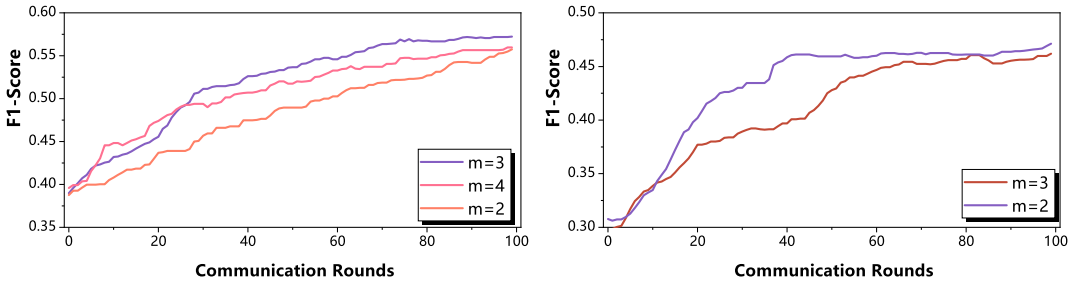
4.4.3 Ablation Study

Effect on Group Number We then adopted the proposed CFed-VS approach to cluster the ten clients into 2 to 4 and 2-3 groups in TVSum and SumMe. The performance of CFed-VS with different group numbers m was conducted in 100 global communication rounds under the 1-class non-IID setting. As shown in Figure 4.6, CFed-VS can efficiently converge on both datasets and achieves the best performance with groups $m = 3$ and $m = 2$ in TVSum and SumMe.

Effect on Client Number We finally examine the effect of device number K on CFed-VS, in which K = 10, 15, 20, 30 under three groups and two groups in TVSum and SumMe separately. Figure 4.7 illustrates the performance of CFed-VS for TVSum and SumMe datasets under the 1-class non-ID setting. We observe that the number of devices does not

Table 4.4 Rank-order correlation coefficients computed between predicted importance scores by different models and human-annotated scores on both SumMe and TVSum datasets using Kendall's τ and Spearman's ρ correlation coefficients.

Method	SumMe		TVSum	
	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
DPP-LSTM [1]	-	-	0.042	0.055
SUM-GAN [158]	0.049	0.066	0.024	0.031
SumGraph [160]	-	-	0.094	0.138
RSGN [161]	0.083	0.085	0.083	0.090
Mixture Transformer	**0.102**	**0.107**	**0.098**	**0.149**

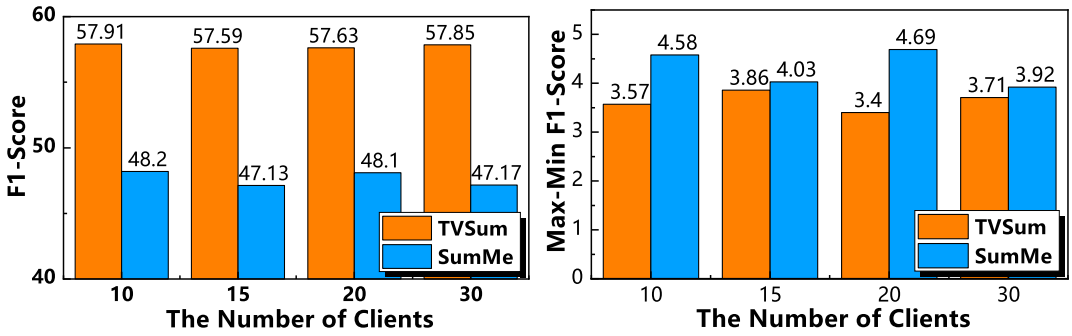


(a) TVSum (b) SumMe

Fig. 4.6 Evaluation of model performance on TVSum and SumMe with different group numbers.

significantly affect the model performance. Meanwhile, the low variance of the Max-Min F1-Score, which measures the difference between the best and worst F1-Score of group models evaluate in all clients, indicates that group models can effectively generalize to the various clients. The results of these experiments indicate that device numbers have a stable impact on the CFed-VS framework.

Discussion on Efficiency and Computational Cost The CFed-VS algorithm, while demonstrating superior performance in terms of the F1 score, introduces additional computational overhead compared to simpler methods like FedAvg. This is primarily due to the inclusion of K-means clustering and the Mixture Transformer backbone, both of which contribute to handling heterogeneous data more effectively but at the cost of higher computational complexity.



(a) F1-Score (b) Max-Min F1-Score

Fig. 4.7 Model performance analysis on TVSum and SumMe with different numbers of clients.

The **FedAvg algorithm** has a computational complexity of $O(N \times d)$, where N represents the number of clients, and d is the dimensionality of the model. FedAvg is well-suited for environments with limited computational resources due to its relatively low overhead.

In contrast, the proposed method involves **K-means clustering** for grouping clients based on data distribution similarities. This clustering step has a complexity of $O(K \times N \times d)$, where K is the number of clusters. The additional computational cost introduced by this step allows the algorithm to improve model generalization in non-IID settings, ensuring that client groups are more homogeneous and thus better suited for personalized model training.

Furthermore, the **Mixture Transformer Backbone** in our method introduces an additional layer of complexity. Transformers generally scale quadratically with the input length, with a computational complexity of $O(L^2 \times d_{\text{model}})$, where L is the sequence length, and d_{model} is the hidden dimension. While this complexity is higher than that of simpler architectures, it allows the model to capture both local and global patterns in video sequences, which is critical for federated video summarization tasks.

Though these additions increase the computational overhead, they result in significantly improved performance, as demonstrated by the results in Fig. 4.5. In future work, optimizations such as model compression techniques or distributed computation strategies could help reduce the computational cost without sacrificing performance.

4.4.4 Discussion on Privacy

Conventional parameter-based clustering methods, usually require training a global model in certain communication rounds to determine the difference between client models. Based on the similarity of parameters, the server groups the client and FL training is then performed independently for each client cluster to produce multiple federated models. Our proposed

approach clusters the clients before the FL training, which saves the communication rounds for the clustering procedure. In terms of privacy, it is worth noting that the summary of the data distribution sent by the client to the server does not reveal any specific information regarding the video of the user. Since the client sends a single value to the server, our approach is more private than clustering based on label distribution or feature distribution.

4.5 Conclusion

In this work, we established the first FLVS benchmark with three key technical directions. 1) For the federated learning foundation, we investigated the Frame-Based Aggregation tailored for the VS problem with different video lengths. 2) A Community-Aware Clustering Federated Learning (CFed-VS) framework was proposed. By completing the proxy registration before FL training, community clients with similar distances of data distribution effectively mitigated the data heterogeneity. 3) The proposed Mixture Attention Transformer alleviated the bottleneck problem and significantly improved the model generalization in the non-IID setting. Our thorough evaluation on both datasets suggested favorable outcomes of our method compared to existing established FL and VS frameworks. Future development of the CFed-VS framework can investigate how it could be adapted or improved to handle even larger and more diverse datasets.

Epilogue

In this chapter, we aimed to address the research questions 2.1 and 2.2, focusing on developing advanced methodologies for video summarization within federated learning frameworks that ensure both data privacy and computational efficiency.

For RQ 2.1, which investigates how video summarization techniques can preserve data privacy while extracting valuable content from video data, we introduced a novel Cluster-based Federated Learning (CFed-VS) approach. This method divides clients into distinct clusters based on their data distribution, allowing localized training that minimizes privacy risks while effectively summarizing the inherent complexities of video data. By leveraging Frame-based Aggregation techniques, our model is capable of handling diverse video lengths and structures, ensuring that video summarization is performed in a secure and effective manner without compromising the temporal and structural characteristics of the video content.

Regarding RQ 2.2, which explores the design of federated learning approaches that balance privacy protection and computational efficiency, we proposed an innovative clustering method that performs data distribution estimation before the federated learning training

begins. This pre-clustering step reduces communication costs and enhances computational efficiency by eliminating the need for repeated clustering operations during the training process. Furthermore, we introduced the Mixture Transformer, a model tailored to address the complexities of video summarization in non-IID federated environments. Through the use of Associated Queries and a mixture attention map, the Mixture Transformer provides a more refined attention mechanism, effectively addressing the "Softmax bottleneck" and improving generalization across heterogeneous video datasets.

In summary, the CFed-VS framework successfully meets the dual goals of privacy preservation and computational efficiency in video summarization. By addressing the unique challenges posed by video data—such as its volume, complexity, and variability in length—while ensuring robust privacy protections, CFed-VS represents a significant advancement in the field of federated video analytics. Future work may focus on further optimizing these techniques to handle even larger datasets and more complex video structures, enhancing both privacy and performance.

Chapter 5

Privacy-Enhanced Zero-Shot Learning via Data-Free Knowledge Transfer

Prologue

Building on the discussion of Community-Aware Federated Video Summarization (CFed-VS) and its contributions to privacy-preserving video processing, this chapter shifts focus to the critical intersection of data utilization and privacy in machine learning. While video summarization emphasizes privacy, the need for privacy-preserving methodologies extends across all data-driven tasks, particularly in the context of annotated datasets used for training complex models.

This chapter introduces **Privacy-Enhanced Zero-Shot Learning (PE-ZSL)**, a framework designed to tackle the challenges of utilizing annotated data while maintaining strict privacy standards. PE-ZSL enables models to classify objects in unseen categories without accessing sensitive data, addressing the limitations of traditional data-sharing practices and the scarcity of labeled data.

At the core of PE-ZSL is a novel data-free knowledge transfer mechanism, which combines zero-shot learning principles with advanced techniques for secure model training. This approach allows models to learn from abstract, non-sensitive information, striking a balance between maximizing data utility and preserving privacy. The development of PE-ZSL represents a significant step toward building robust machine-learning models that respect ethical data use while maintaining strong performance.

Declaration: This chapter is a modified version of **"Privacy-Enhanced Zero-Shot Learning via Data-Free Knowledge Transfer"**, published in IEEE International Conference on Multimedia & Expo(ICME), 2023. Code Link

5.1 Introduction

The blossom of deep learning technologies embraces the development of high-performance computing and large-scale multi-modal data. However, sharing data across different institutes and even between different countries has become increasingly difficult and sensitive. The increasing awareness of data copyright, expensive data annotation, and restricted access to data in expert domains have hindered the development of interdisciplinary and intercultural deep models. However, sharing data across different institutes and even between different countries has become increasingly difficult and sensitive. The increasing awareness of data copyright, expensive data annotation, and restricted access to data in expert domains have hindered the development of interdisciplinary and intercultural deep models.

As shown in Fig.5.1, datasets may contain sensitive data, *i.e.*, healthcare and face information, which cost data owners billions and tens of years to collect. Strict regulations, such as the GDPR [162] in Europe have been enforced to control the risk of data leaking. **In our work, we focus on protecting data copyright and eliminating data sensitivity from data owners when data contain confidential user information.** Concretely, AI service providers (*i.e.*, AI companies) maintain close cooperation with the data owners (*i.e.*, scientific institutions and hospitals) and they need to obtain data to provide related services for customers. However, the healthcare dataset is expensive to collect so the hospital cannot share the data directly with AI service providers due to copyright protection. When AI companies need access to such sensitive data to provide AI services, the shared data is exposed to leaking risks even though tedious confidentiality agreements have been signed. Take another example of the situation related to surveillance data, AI service providers take on the task, *i.e.*, pedestrian re-identification, while it is necessary to eliminate the sensitive user face information. **Motivated by these challenges, this work explores a theoretical case when an AI developer needs to train a new model, the data owner (from different institutes) can provide a data-free teacher service as an API so that knowledge can be transferred without any data sharing.**

As a promising machine learning paradigm, zero-shot learning (ZSL) shows good potential to tackle data-free problems which investigates an extreme case when such deep transfer can go beyond seen classes in the teacher dataset. Existing ZSL models are established based on real data from either seen or unseen classes. When adapting a pre-trained model to a new

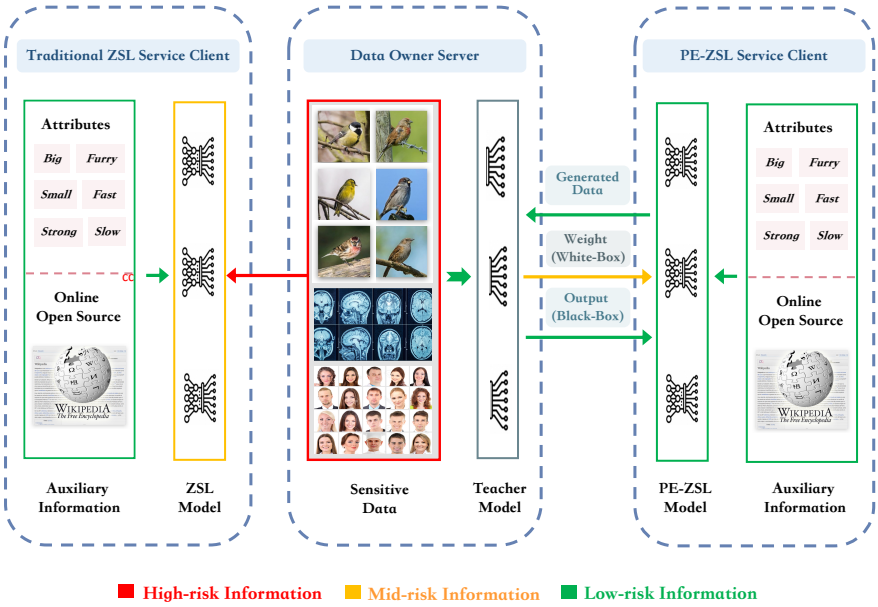


Fig. 5.1 Traditional ZSL models require access to real images from the data owner to learn the visual-semantic associations. PE-ZSL suggests an extra data safeguard using a teacher model so that a PE-ZSL model can achieve GZSL without access to any real images. The training of PE-ZSL only involves generated data and prior auxiliary information and guidance from the teacher model.

task domain, existing ZSL models assume a large amount of labeled seen class or unlabeled unseen class data are available to establish the visual-semantic relationship. However, sharing data across different institutes and even different countries is often infeasible. Different from existing ZSL settings, we focus on establishing the ZSL model without data sharing during training. In this work, we propose a new paradigm dubbed Privacy-Enhanced Zero-Shot Learning (PE-ZSL) to avoid sensitive data leaking while still enabling AI model can be trained. Figure 5.1 briefly illustrates the difference between ZSL and PE-ZSL tasks. Our PE-ZSL task suggests replacing data with a teacher model (pre-trained on real data) to guide the ZSL model training. The teacher model can be regarded as the implicit representation of data so the PE-ZSL model can be established through the supervision of the teacher model, which can prevent real data from being shared.

To comprehensively explore our proposed PE-ZSL framework, we also present extensive discussion from the perspective of privacy issues and knowledge space of the teacher model. First, we propose two PE-ZSL scenarios in terms of framework privacy. In the 'black-box' scenario, the teacher only provides output classification scores but does not share weights. In the 'white-box' scenario, the teacher will also share the model weights during training, which is more informative. These two scenarios indicate different levels of communication between

data owners and AI service providers, which will lead to different ZSL recognition performances. In terms of teacher model privacy, we adopt differential privacy [163] in teacher training, which protects against the adversary who has access to model information, *i.e.*, parameters. Furthermore, we propose omniscient and quasi-omniscient teachers according to the knowledge space, *i.e.*, whether unseen classes are involved in the pre-training teacher model. In summary, our contributions are three-fold:

- Privacy-Enhanced Zero-Shot Learning aims to achieve zero-shot classification without access to real data. The paradigm can be applied to real-world applications for data copyright protection and sensitivity elimination.
- We develop a novel data-free knowledge transfer framework for the PE-ZSL task. In addition to zero data sharing setting, we propose 'black-' and 'white-box' scenarios and discuss the pros and cons of model sharing problems. We also present an analysis of the teacher model in both omniscient and quasi-omniscient settings according to the knowledge space.
- We show experimental results for conventional and generalized ZSL tasks in two scenarios. Though the PE-ZSL model is established without data sharing during training, it achieves promising performance.

5.2 Related Work

The most widely used framework for data privacy enhancement is Federated Learning [164]. A global model is shared with clients to avoid data leaking. Knowledge distillation [165] utilizes the domain-expert teacher model to train a compact student model and it can prevent the teacher model from being attacked. Yet, none of these methods have explored the potential in zero-shot learning situations. This work presents the first work exploring a privacy-enhanced zero-shot learning paradigm via data-free knowledge transfer.

Zero-Shot learning [4] enables deep learning model[137, 125, 126, 122] to recognise unknown/unseen classes by establishing the relationship between seen and unseen classes via class semantic information. Some work [66] aims to build the mapping between visual and semantic space. Other works [166] focus on unseen class data generation to alleviate the data-missing problem. According to whether unseen data is adopted during training, existing ZSL methods can be categorised into inductive [126] and transductive [167] settings. As for the test phase, conventional ZSL (CZSL) methods [67] assume test data only come from unseen classes, while generalized ZSL (GZSL) [68] is then proposed to assign both

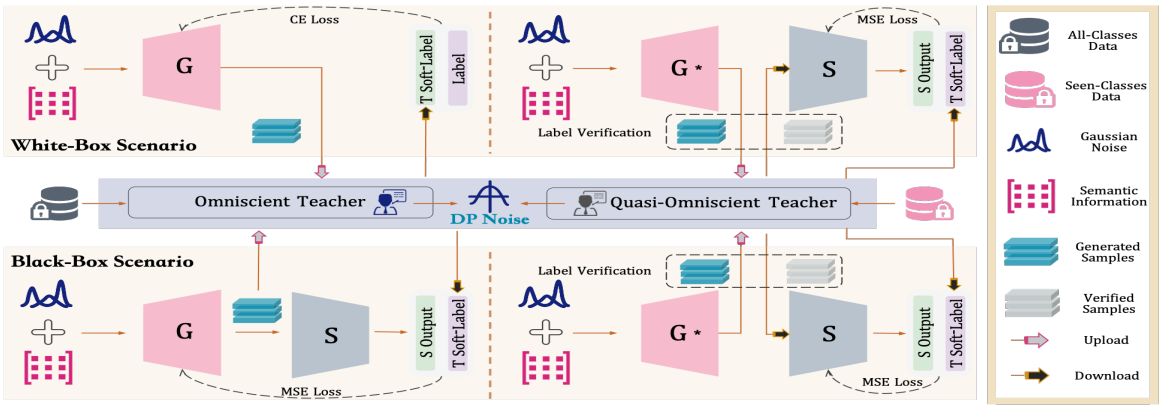


Fig. 5.2 Overall framework in the black-box and white-box scenarios. In the white-box scenario, the generator has access to teacher weights during training while the teacher only provides output guidance in the black-box scenario.

seen and unseen data into corresponding classes. There has been little research on zero-shot learning whilst enhancing data privacy, so we propose a privacy-enhanced zero-shot learning paradigm, which aims to accomplish zero-shot recognition without access to real data during training.

5.3 Privacy-Enhanced Zero-Shot Learning

As shown in Fig.5.2, PE-ZSL addresses the problem when sensitive data is secured on the *Data Owner* domain. The key idea is to introduce a teacher model as the data safeguard and guide the model deployed on the *AI Service Provider* domain to train a classifier with zero real data. In addition to data privacy-enhancing, we introduce white- and black-box scenarios to discuss the teacher model sharing problem regarding the balance between performance and security.

5.3.1 Problem Formulation

The basic PE-ZSL setting involves secured images and their extracted visual features $x \in \mathscr{X}$. The data safeguard is provided by a pre-trained teacher model on the data owner domain. For simplicity, we consider a supervised learning model $f_T : \mathscr{X} \rightarrow \mathscr{Y}$, where $y \in \mathscr{Y}$ is the label space. The ultimate goal is to train a student model in the AI service provider's domain using an objective function ℓ that learns from the guidance of the teacher model:

$$\ell\left(f_{PE-ZSL}\left(\tilde{x}\right), f_T\left(\tilde{x}\right)\right), \tag{5.1}$$

where $\tilde{x} \in \tilde{\mathscr{X}}$ is the generated data, ensuring that no real data is accessed. Here, $f_{PE-ZSL}(\tilde{x})$ represents the full model in the PE-ZSL framework, which is composed of two parts: the generator and the student model. The generator is responsible for synthesizing the data $\tilde{x}$, while the student model learns from this generated data. The loss function $\ell(\cdot,\cdot)$ measures the discrepancy between the output of the student model within f_{PE-ZSL} and the output of the teacher model f_T. The goal is to minimize this loss, transferring knowledge from the teacher model to the student model without ever accessing the real data.

PE-ZSL with Omniscient & Quasi-Omniscient Teacher On the data owner domain, we further break down the PE-ZSL into omniscient and quasi-omniscient teachers according to the label space. Seen classes are defined as $\mathscr{S} = \{(x_s, a_s, y_s) \mid x_s \in \mathscr{X}_s, a_s \in \mathscr{A}, y_s \in \mathscr{Y}_s\}$, where $x_s \in \mathbb{R}^{d_x}$ denotes the d_x-dimensional visual feature in the set of seen class features, $a_s \in \mathbb{R}^{d_a}$ denotes the d_a-dimensional auxiliary class-level semantic embedding, and $\mathscr{Y}_s$ stands for the set of labels for seen classes. Unseen classes are defined as $\mathscr{U} = \{(x_u, a_u, y_u) \mid x_u \in \mathscr{X}_u, a_u \in \mathscr{A}, y_u \in \mathscr{Y}_u\}$, where x_u represents the unseen class features, a_u denotes the semantic embedding of unseen classes and y_u denotes the unseen class labels. Seen and unseen classes are disjoint, *i.e.*, $\mathscr{Y}_s \cap \mathscr{Y}_u = \emptyset$. For PE-ZSL, both seen and unseen features, x_s and x_u, are unavailable for the service provider. Available information for the service provider can be represented as $\mathscr{T}_r = \{(a, y) \mid a \in \mathscr{A}, y \in \mathscr{Y}\}$, which means only semantic embedding and class labels can be accessed. The teacher model pre-trained by real data is provided for model training guidance. In this way, the basic PE-ZSL with omniscient teacher considers $f_T : \mathscr{X} \to \mathscr{Y}$ because the source domain contains both seen and unseen classes. A more challenging PE-ZSL with quasi-omniscient teacher considers $f_T : \mathscr{X}_s \to \mathscr{Y}_s$.

ZSL vs GZSL On AI service provider domain, PE-ZSL aims to classify test images $f_{ZSL} : \mathscr{X}_u \to \mathscr{Y}_u$ for CZSL, and $f_{GZSL} : \mathscr{X} \to \mathscr{Y}$ for GZSL. Training of the above classifiers using absolutely generated data will be introduced next.

5.3.2 White-Box & Black-Box Scenarios

The objective function of PE-ZSL in Eq.(6.1) defines a data-free knowledge transfer framework for PE-ZSL task, *i.e.*, through the guidance of the teacher model, our proposed PE-ZSL formula consists of two components: a *Generator G* and a *Student* network *S*. Generator *G* is to synthesize features and student *S* aims to match the performance of the teacher model. Figure 5.2 depicts the detailed PE-ZSL framework in both white-box and black-box scenarios. The system consists of 1) the secured data and teacher model on the data owner; 2) the PE-ZSL model on the AI service provider; 3) and the information exchange channels. Considering that model inversion can attack shared models, we also investigate the security levels regarding model sharing in addition to data privacy enhancement. For the white-box

scenario, the teacher weights are involved in computing the gradient for the generator and student network training. In the black-box scenario, the teacher only provides the output as pseudo labels, *i.e.*, the teacher model is not involved in backpropagation for optimization of the PE-ZSL framework.

White-Box Scenario. In the white-box scenario, the teacher model provides both gradient and softmax output as the PE-ZSL training guidance as follows. **1) Uploading generated data:** the generator synthesizes features of the same classes with the supervision of a teacher based on noise **z** and class-level semantic embedding **a** (attributes or BERT model of class names [109] as the condition). Specifically, we aim to synthesize the features that can be classified into corresponding classes with the constraint of the teacher network. $\tilde{x} = G(z|a;\theta_G)$ represents the generated features that are uploaded to the data owner. **2) Gradient and softmax guidance:** The teacher model receives $\tilde{x}$ and process the data using the loss function:

$$\min_{\theta_G} \mathscr{L}(\tilde{x}, y; \theta_G) + \alpha \mathscr{R}(\tilde{x}), \tag{5.2}$$

where $\mathscr{L}(\cdot)$ represents cross-entropy loss by teacher model for classification, $\mathscr{R}(\cdot)$ refers to the regularization term during feature generation with hyperparameter α. The regularization term aims to minimize the distribution distance of real and generated features. Note that regularization is also completed on the data owner side and real data will not be accessed by the AI service provider. **3) Feedback downloading**: a request is sent to the service provider so that the gradient, the regularization of distribution divergence and softmax output can be downloaded. **4) Label verification:** Using softmax to compute pseudo labels and filter out misclassified generated samples:

$$\begin{aligned}(\tilde{x}^*, y^*) \in \{(\tilde{x}, y) | y = argmax\, T(\tilde{x}; \theta_T^*),\\ \tilde{x} = G(z|a;\theta_G^*)\},\end{aligned} \tag{5.3}$$

where T represents teacher model, θ_T^* and θ_G^* are the optimised parameters of teacher and generator, $\tilde{x}^*$ is the high-quality generated features, y^* is the corresponding class labels. **5): Training the student model:**

$$\min_{\theta_S} \|T^*(\tilde{x}^*;\theta_T^*) - S(\tilde{x}^*;\theta_S)\|_2^2, \tag{5.4}$$

where S and θ_S denotes the student model and its parameters. In the white-box scenario, the gradient is imposed directly onto generated features and can massively improve the performance of the generator. As a trade-off, the gradient feedback is mid-risk information

Algorithm 2 Training Procedure in Both Protocols

Require: Pre-trained Teacher network θ_T^*, class labels $\mathscr{Y}_{tr}$ and their auxiliary semantic embedding $\mathscr{A}$; the maximal number of training epochs T_g and T_s for generator and student network, respectively.
Ensure: The learned parameters θ_G, θ_S for generator G and student network S, respectively.

```
Initialize θ_G, θ_S. Set iteration epochs t_g = 1, t_s = 1.
while t_g < T_g do
    if White-Box Protocol then
        Train generator with gradient guidance from teacher network using Eq.(5.2).
    else if Black-Box Protocol then
        Train generator with output guidance from teacher network using Eq.(5.5).
    end if
    t_g ← t_g + 1
end while
Conduct label verification using Eq.(5.3).
while t_s < T_s do
    Train student network with output guidance from the teacher using Eq.(5.4).
    t_s ← t_s + 1
end while
```

(may lead to teacher model leaking) whereas the softmax and regularization feedback are low-risk.

Black-Box Scenario. The black-box scenario only differs from the white-box scenario in the guidance provided by the teacher model in the second step. Only low-risk regularization and softmax output can be requested from the teacher model so as to avoid the model leaking risk. Specifically, generated features $\tilde{x} = G(z|a;\theta_G)$ are uploaded to the data owner to compute the softmax and divergence regularization. The data owner then creates a request so that the feedback can be downloaded. Generated data can validate whether its conditional class input can match the teacher softmax output and misclassified samples are filtered out. Generator G and student network S are then trained as an end-to-end model as follows:

$$\min_{\theta_G,\theta_S} \|T^*(\tilde{x};\theta_T^*) - S(\tilde{x};\theta_S)\|_2^2 + \alpha\mathscr{R}(\tilde{x}), \tag{5.5}$$

where θ_G, θ_S are parameters of generator and student model. The comprehensive training procedures for both protocols are delineated in Algorithm 2.

This work mainly focuses on investigating the following research questions (**RQ**): **1)** What are the impacts of different teacher feedback information on the quality and diversity of generated data? **2)** different semantic information as generation condition and their impacts; **3)** trade-off between performance and security in white-box and black-box; **4)** can student generate new knowledge beyond the limitation of a quasi-omniscient teacher? **5)** previous work uses real seen data and generated unseen data, which causes bias towards seen classes.

Table 5.1 Detailed dataset statistics and data split in PE-ZSL. Notation: 'att' - attribute; 'S' - seen class; 'U' - unseen class; 'Om' - omniscient teacher; 'Q-Om' - quasi-omniscient teacher.

Dataset	Semantics	Class Number	Image	Teacher (Om/Q-Om)		PE-ZSL Training		PE-ZSL Evaluation (Om/Q-Om)	
		S/U		S	U	S	U	S	U
AWA1 [4]	BERT/att	40/10	30475	19832	4542/0	**0**	**0**	4958	1143/5685
AWA2 [62]	BERT/att	40/10	37322	23527	6328/0	**0**	**0**	5882	1585/7913
aPY [168]	BERT/att	20/12	15539	5932	6333/0	**0**	**0**	1483	1591/7924

In PE-ZSL, both seen and unseen classes are trained using generated data, which improves consistency between seen and unseen classifiers in the GZSL problem.

5.3.3 Privacy-Enhanced Zero-Shot Classification

After the training process, the generator can synthesize features of good quality and the student network can predict class labels of test features. With the omniscient teacher, where seen and unseen classes are available, the generator can synthesize features of all classes. Given the test features, we can obtain the predicted class labels as follows:

$$y^* = \underset{y \in \mathscr{Y}}{argmax}\, p(y|x, \theta_S^*), \tag{5.6}$$

where θ_S^* denotes optimised parameters of student.

With the quasi-omniscient teacher, where only seen class data is available, the problem is more challenging. The generator is utilized to synthesize data of unseen classes. Given the noise z and unseen class semantic embedding, the generated features can be obtained as $\tilde{x} = G(z|a; \theta_G^*)$. Then it is converted into a supervised learning task. The generated features are adopted to train a classifier C and class labels of test features can be predicted through an optimized classifier.

5.4 Experiments

Datasets and Implementation Details. We evaluate our PE-ZSL model on three benchmark datasets: AWA1[4], AWA2 [62]) and aPY [168]. AWA1 and AWA2 consist of 30,475 and 37,322 images of 50 classes. aPY contains 15,539 images of 32 classes. As a semantic representation, we use 768-dimensional word embedding generated by BERT [109]. Following [62], we adopt the 2048-dimensional ResNet101 features as image representation. As for data split, we follow the proposed data split in [62] for quasi-omniscient teacher.

Table 5.2 Comparison results in CZSL and GZSL tasks. 'WB' & 'BB' represent white- & black-box scenario, '*' represents TZSL method. 'PE-ZSL+WB/BB*' and 'PE-ZSL+WB/BB' represent our model with the omniscient and quasi-omniscient teacher.

Method	Zero-Shot Learning			Generalized Zero-Shot Learning								
	AWA1	AWA2	aPY	AWA1			AWA2			aPY		
	T1	T1	T1	u	s	H	u	s	H	u	s	H
DAP [4]	44.1	46.1	33.8	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
ALE [66]	59.9	62.5	39.7	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEM [60]	68.4	67.1	35.0	32.8	84.7	47.3	30.5	86.4	45.1	11.1	75.1	19.4
f-CLSWGAN [108]	68.2	-	-	57.9	61.4	59.6	-	-	-	-	-	-
CE-GZSL [169]	71.0	70.4	-	65.3	73.4	69.1	63.1	78.6	70.0	-	-	-
SDGZSL [170]	-	74.3	47.0	-	-	-	69.6	78.2	73.7	39.1	60.7	47.5
ICCE [171]	74.2	72.7	49.5	67.4	81.2	73.6	65.3	82.3	72.8	45.2	46.3	45.7
DTN* [172]	69.0	-	41.5	54.8	88.5	67.7	-	-	-	37.4	**87.9**	52.5
GMSADE* [173]	81.3	80.7	49.9	71.2	87.7	78.6	71.3	86.1	78.0	76.1	39.3	51.8
EDE* [174]	**85.3**	77.5	31.3	71.4	**90.1**	79.7	68.4	**93.2**	78.9	29.8	79.4	43.3
BGT* [175]	-	**82.4**	49.8	-	-	-	56.2	82.2	66.7	39.3	72.9	51.0
PE-ZSL+BB	14.1	19.9	12.3	4.1	3.7	3.9	3.5	3.7	3.6	6.8	4.0	5.1
PE-ZSL+WB	34.5	36.5	18.7	23.4	34.3	27.8	27.3	44.3	33.7	17.9	52.5	26.7
PE-ZSL+BB*	33.5	29.0	30.2	33.5	28.6	30.9	29.0	25.3	27.0	30.2	42.2	35.2
PE-ZSL+WB*	77.9	79.0	**83.9**	**77.9**	81.8	**79.8**	**79.0**	86.7	**82.7**	**83.9**	85.7	**84.8**

Table 5.3 Experimental results in the black-box scenario with the omniscient teacher in both CZSL and GZSL tasks.

Method	Zero-Shot Learning			Generalized Zero-Shot Learning								
	AWA1	AWA2	aPY	AWA1			AWA2			aPY		
	T1	T1	T1	u	s	H	u	s	H	u	s	H
Label-Conditioned	15.5	10.0	7.0	15.5	24.3	18.9	10.0	17.8	12.8	7.0	3.8	4.9
Attribute-Conditioned	10.1	23.0	8.2	10.1	11.3	10.7	23.0	17.6	20.0	8.2	5.0	6.3
w/o Label Verification	25.6	24.7	11.8	25.6	15.6	19.4	24.7	18.1	20.9	11.8	20.9	15.0
w/o Regularization	26.8	23.7	23.2	26.8	26.7	26.8	23.7	23.2	23.4	23.2	25.6	24.3
PE-ZSL+BB	**33.5**	**29.0**	**30.2**	**33.5**	**28.6**	**30.9**	**29.0**	**25.3**	**27.0**	**30.2**	**42.2**	**35.2**

The omniscient teacher is trained with all classes, so we split unseen classes randomly into training and test sets following [172]. Student and teacher models have the same architecture, which has two hidden layers with 1024 and 512 units. Generator contains a single hidden layer with 4096 hidden units. The dimension of the **noise vector** z is set to 20 for all datasets. The regularization term weight is set to 0.5 for AWA1 and AWA2, and 1 for aPY. The number of generated features is 400 in average per class for all datasets. For the training epochs T_g and T_s, we selected values that balance convergence and prevent overfitting or underfitting for both the generator and student network. Experimentally, we found performance plateaus in both networks beyond certain iterations, indicating an optimal stopping point for training. Consequently, $T_g = 50$ and $T_s = 80$ were set to optimize both computational efficiency and model effectiveness.

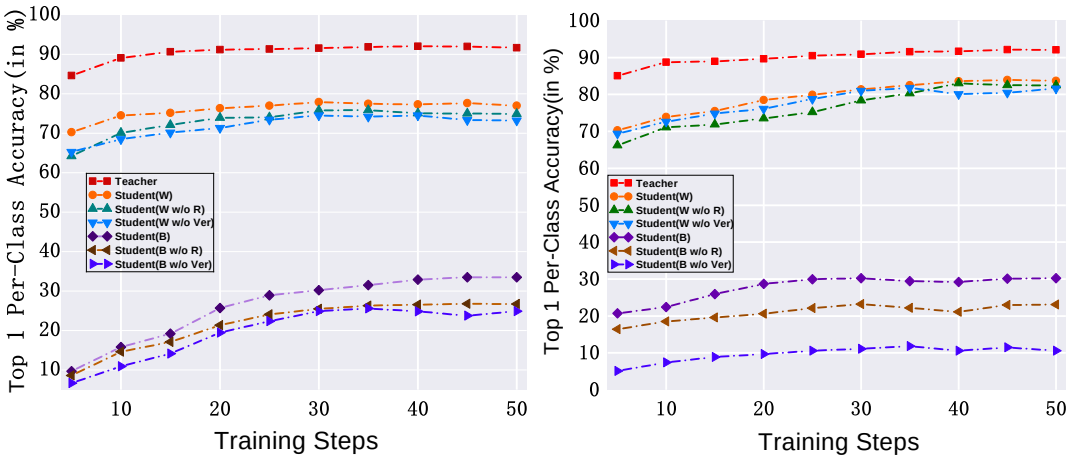
Evaluation Protocol. Following [62], we adopt the per-class average top-1 accuracy (T1) for CZSL task. We use harmonic mean $H = (2 \times u \times s)/(u+s)$ for evaluation in GZSL, where u and s denote average per-class top-1 accuracy on unseen and seen classes, respectively.

5.4.1 Main Results

Comparisons with State-of-Arts. We present experimental results in both CZSL and GZSL tasks in Table 5.2. Considering this is the first PE-ZSL work, we provide a comparison with traditional state-of-the-arts as a reference. To investigate **RQ1**, we show results under two kinds of feedback from omniscient and quasi-omniscient teachers. PE-ZSL model with omniscient teacher achieves promising performance in both CZSL and GZSL in white-box scenario. We achieve the best performance in GZSL, especially on aPY, with an increase in harmonic mean of 32.3% than DTN* method, which indicates an improved balance of seen and unseen classes. As for the black-box scenario, the accuracy on unseen classes is 4.9% higher than seen classes on AWA1. It indicates that the PE-ZSL model is promising to mitigate the class-level overfitting issue in the GZSL task proposed in **RQ5**. Compared with inductive ZSL methods, results show that our model with the quasi-omniscient teacher in the white-box scenario gains satisfactory performance in GZSL, especially on aPY, with 7.3% higher performance on the harmonic mean. It is very impressive that the student model can generate new knowledge beyond the source data of the teacher model as discussed in **RQ4**. For the black-box scenario, results show our PE-ZSL model outperforms random guessing, which is around 10% on AWA1, AWA2 and 8% on aPY. The white-box achieves better performance than the black-box, indicating that gradient guidance provides more information.

Comparisons in Black-Box Scenario. As it is the first time to propose this setting, we provide several baselines for comparison in Table 5.3. We provide label and attribute for conditional feature generation to investigate **RQ2**. Our proposed framework with BERT embedding achieves the best performance, *i.e.*, with 18.0% and 23.4% increases in unseen accuracies on AWA1. Results show that our framework gains obvious improvement in accuracy with label verification, *i.e.*, with 20.2% higher performance on harmonic mean on aPY dataset. And results indicate the effectiveness to adopt regularization, *i.e.*, it achieves 3.6% and 10.9% increases in Harmonic mean on AWA2 and aPY. The comparison with baselines demonstrates the effectiveness of our PE-ZSL model in black-box scenario with omniscient teacher.

Performance vs Framework Privacy. Compared to traditional ZSL methods, the performance under the white-box scenario is very promising, since data privacy is already preserved and our model can still achieve adequate performance. Compared with white-box scenario,



(a) AWA1 (b) aPY

Fig. 5.3 Epoch analysis for unseen accuracy. 'Ver': label verification. 'R': regularization term.

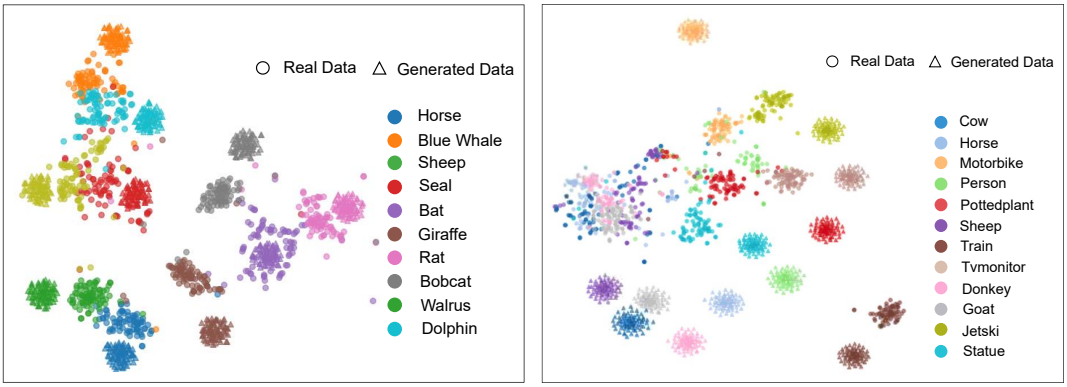
black-box is more secure but sacrifices classification performance. Thus, the performance of black-box scenario is reasonable because both data privacy and model safety are guaranteed as proposed in **RQ3**.

Semantic Attribute Representation and Privacy Considerations. In our framework, we use **BERT-generated word embeddings** based on class names, rather than conventional expert-annotated attributes commonly used in zero-shot learning (ZSL). Unlike manual annotations, which could expose sensitive information, BERT embeddings are derived from publicly available data and capture general semantic relationships. This ensures the attributes used are **non-sensitive** and do not contain personal or proprietary information. By leveraging these public embeddings, our framework preserves privacy in both white-box and black-box settings. The embeddings, which are broad and general, ensure that no real user data is exposed, aligning with the privacy goals of the proposed model.

5.4.2 Analysis and Discussion

Student Performance Analysis. We present performance of teacher and student model with increasing training steps in both scenarios on AWA1 and aPY in Figure 5.3. Student in white-box scenario obtains results close to teacher, indicating the effectiveness of gradient guidance. Besides, results show that model achieves better performance with regularization term, indicating the effectiveness of feature distribution during training. And statistics also show that framework performs better with label verification in both scenarios, which indicates

its necessity because it can mitigate the negative influence caused by generated features of inferior quality.



(a) AWA1 (b) aPY

Fig. 5.4 t-SNE visualization on AWA1 and aPY

Quality of Generated Features. Figure 5.4 shows the t-SNE visualization of real and generated unseen features in both scenarios with omniscient teacher on AWA1 and aPY. We randomly choose a part of features for clear visualization. Generated features have distribution close to real ones and they are more class-level clustered, indicating effectiveness of feature generation under the supervision of teacher guidance, even though real data is unavailable. Therefore, generated features can be viewed as a suitable replacement for real features.

5.5 Conclusion

This work has presented a privacy-enhanced ZSL paradigm via data-free knowledge transfer. A pre-trained teacher model was deployed on the data owner as data safeguard to provide guidance for model training. We extensively studied the 'black-box' and 'white-box' scenarios and their trade-off in performance and framework privacy. Our model maintains promising performance in CZSL and GZSL tasks despite the absence of real data during training. Future development of PE-ZSL can focus on model design, reducing communication costs and improving performance in the inductive setting.

Epilogue

This chapter has focused on addressing RQ 3.1, which examines how innovative mechanisms can enhance both the utility and confidentiality of data in machine learning, specifically in the context of recognizing both seen and unseen classes while maintaining data privacy.

To respond to RQ 3.1, we introduced the Privacy-Preserving Zero-Shot Learning (PP-ZSL) framework. This framework presents a novel approach by integrating data-free knowledge distillation techniques to protect the privacy of data while maintaining its value. In this approach, data owners can utilize their private datasets to train a teacher model, which can then be used by others through an API. By doing so, privacy is preserved, as no raw data is exposed, and data owners can also create new revenue opportunities by offering access to the trained models.

The PP-ZSL framework further addresses the challenge of recognizing unseen classes without sharing real data. By leveraging semantic information and guiding the training of the generator and student model under different levels of security, the system ensures that seen and unseen classes are effectively recognized. This process not only enhances the utility of the data but also ensures that privacy remains intact, even in scenarios where sensitive information could otherwise be exposed.

Additionally, the Sentinel-Guided Zero-Shot Learning (SG-ZSL) framework introduces both black-box and white-box training scenarios, allowing for varying levels of privacy protection. By incorporating differential privacy into the training of the teacher model, data owners are given the flexibility to set privacy levels according to their specific needs, balancing the trade-off between privacy and model performance. This fine-grained control over privacy offers a practical solution to address different levels of sensitivity in data while still maintaining robust performance in machine learning tasks.

In summary, the innovative mechanisms presented in this chapter successfully tackle the core challenge of RQ 3.1 by providing robust privacy-preserving solutions that maximize the utility of the data for machine learning tasks, particularly for the recognition of both seen and unseen classes. Future research could focus on refining these approaches to further improve model performance while maintaining strict privacy standards across an even wider range of applications.

Chapter 6

Sentinel-Guided Zero-Shot Learning

Prologue

In the preceding chapter, we explored Privacy-Enhanced Zero-Shot Learning (PE-ZSL), a framework designed to address the challenges of data sensitivity and copyright protection by enabling zero-shot learning without direct exposure to real data. PE-ZSL set the stage for privacy-preserving machine learning, focusing on data-free knowledge transfer and ensuring that sensitive information remains protected throughout the training process.

Building on the foundation of PE-ZSL, this chapter introduces **Sentinel-Guided Zero-Shot Learning (SG-ZSL)**, an evolution that addresses some of the critical limitations of PE-ZSL. SG-ZSL not only enhances the data privacy mechanisms introduced in the previous framework but also integrates a more comprehensive comparison between SG-ZSL and Inductive Zero-Shot Learning and Transductive Zero-Shot Learning. Through extensive experiments conducted on the various datasets, SG-ZSL demonstrates superior performance in both ZSL and GZSL tasks, surpassing existing methods by a significant margin. Differential privacy further strengthens the privacy guarantees during the model's training and deployment, ensuring minimal leakage of sensitive information while maintaining high classification accuracy.

Moreover, the chapter presents in-depth experimental results in both white-box and black-box protocols, highlighting the flexibility and robustness of SG-ZSL across different privacy settings.

The objective of this chapter is to demonstrate how SG-ZSL not only builds upon but significantly expands the privacy-preserving capabilities of PE-ZSL. By leveraging enhanced experimental analysis and in-depth comparisons with ZSL variants, SG-ZSL stands as a more versatile and adaptable solution in the zero-shot learning landscape.

Declaration: This chapter is a modified version of **"Sentinel-Guided Zero-Shot Learning: A Collaborative Paradigm without Real Data Exposure"**, published in IEEE Transactions on Circuits and Systems for Video Technology(TCSVT), 2024. Code Link

6.1 Introduction

Zero-Shot Learning (ZSL) is a promising machine learning paradigm that addresses the challenge of classifying unseen classes by leveraging semantic information. Traditional ZSL approaches often rely on real data to establish visual-semantic associations, which raises significant concerns regarding data privacy, security, and ownership. As machine learning expands into sensitive domains like healthcare and finance, safeguarding data privacy during model training has become increasingly critical.

To address these concerns, Privacy-Enhanced Zero-Shot Learning (PE-ZSL) was introduced. PE-ZSL eliminates the reliance on real data during training by employing omniscient and quasi-omniscient teacher models that guide student models using synthetic data. Additionally, it offers two security protocols—white-box and black-box—allowing clients to balance performance and security according to their specific needs. This flexibility not only enhances the robustness of the PE-ZSL framework but also strengthens its privacy guarantees (see Figure 6.1). However, as real-world applications demand even greater data security and intellectual property protection, further advancements were necessary. Many critical issues in PE-ZSL remain unexplored, such as the impact of different types of semantic information on model performance, the specific roles of various loss functions employed, and whether PE-ZSL provides sufficient protection for model intellectual property.

Building on this foundation, Sentinel-Guided Zero-Shot Learning (SG-ZSL) introduces key improvements to address both data privacy and model security more comprehensively. SG-ZSL retains PE-ZSL's dual-teacher model architecture (omniscient and quasi-omniscient) and dual training protocols, but takes privacy protection a step further by incorporating differential privacy during the training of the teacher models. This ensures that sensitive data remains secure throughout the training process, reducing the risk of data leaks. The addition of differential privacy enhances the framework's robustness in environments requiring high confidentiality. In addition to strengthening privacy safeguards, SG-ZSL provides a more thorough approach to protecting model ownership and intellectual property. In distributed machine learning environments, preventing unauthorized access or misuse of the model's knowledge is critical. SG-ZSL addresses this through detailed ablation studies, evaluating each module's contribution to privacy protection and security. These studies offer insights into

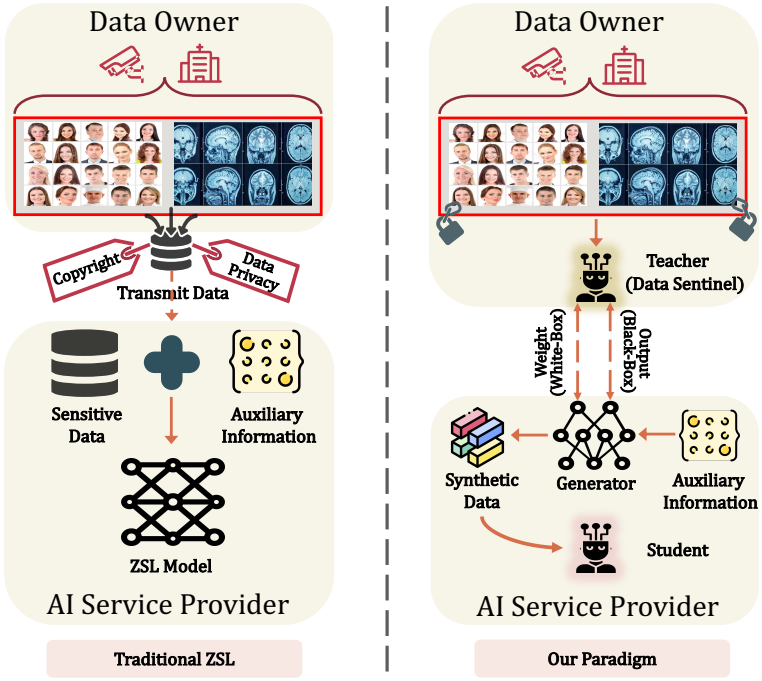


Fig. 6.1 In traditional ZSL approaches, real data is necessitated to establish the visual-semantic association. Conversely, SG-ZSL introduces a teacher model, which acts as a data sentinel, enabling the execution of ZSL tasks without the need for direct access to real data.

how differential privacy and sentinel models collaborate to protect both data and intellectual property, ensuring SG-ZSL's effectiveness in safeguarding against potential breaches.

SG-ZSL also distinguishes itself from traditional ZSL methods such as Inductive ZSL and Transductive ZSL. As shown in Figure 6.1, SG-ZSL utilizes a sentinel model to generate synthetic data and perform knowledge transfer using semantic embeddings, effectively eliminating the need for direct access to real data. This ensures high performance across both seen and unseen classes, all while maintaining strong privacy protection. The sentinel-guided mechanism strikes a balance between data security and model utility, making it highly suitable for privacy-sensitive environments.

Moreover, SG-ZSL includes a comprehensive ablation study exploring the impact of its various components. These evaluations provide a clear understanding of how differential privacy, model generation, and sentinel models work together to optimize both performance and privacy. By advancing data security and model protection, SG-ZSL marks a significant step forward in privacy-preserving machine learning, paving the way for its responsible application in sensitive real-world scenarios.

With these innovations, SG-ZSL not only enhances privacy protection but also introduces robust mechanisms for safeguarding model intellectual property, ensuring secure, efficient, and ethical machine learning for Zero-Shot Learning tasks.

6.2 Related Work

The realm of machine learning has recently experienced a significant shift towards prioritizing data privacy, particularly when handling sensitive information across diverse domains. Federated Learning [176] has been recognized as a formidable framework, designed to mitigate potential data leakage by decentralizing the training process. Recent advancements in this domain have been characterized by the exploration of various architectures and optimization strategies, all aimed at enhancing model performance without sacrificing data privacy. For example, studies [177, 178, 24, 179] have been dedicated to optimizing communication efficiency in federated learning setups, while research such as [33, 180] has delved into the application of federated learning in edge computing, ensuring data privacy at its source.

Differential Privacy [163] has been seamlessly integrated into numerous machine learning paradigms to bolster data privacy. Recent contributions, including [181–183], have investigated the fusion of DP with deep learning, ensuring that while models remain proficient, the privacy of their training data remains uncompromised. For example, Guo *et al.*[181] developed 'TOP-DP', a topology-aware differential privacy approach for decentralized image classification systems, which innovatively utilizes decentralized communication topologies to enhance privacy protection while achieving an improved balance between model usability and data privacy.

Knowledge Distillation [165], on the other hand, has emerged as a pivotal strategy for protecting intricate teacher models by training a streamlined student model, thereby thwarting potential adversarial attacks. Recent endeavors, such as [184–187], have showcased the versatility of knowledge distillation across domains of computer vision. For example, Zhang *et al.* [187] introduced an evolutionary knowledge distillation approach, where an adaptive, online-evolving teacher model continuously transfers intermediate knowledge to a student network, significantly enhancing learning effectiveness, especially in low-resolution and few-sample scenarios.

It is imperative to note, however, that both Federated Learning and Knowledge Distillation are predominantly confined to supervised learning. This confines their utility in scenarios necessitating the recognition and categorization of previously unseen data categories, a domain where Zero-Shot Learning protocols excel. ZSL, with its prowess in recognizing

unseen classes by establishing semantic relationships, transcends the limitations inherent to the supervised nature of both Federated Learning and Knowledge Distillation.

In this work, an innovative SG-ZSL paradigm is introduced. This paradigm, distinct in its data-free knowledge transfer, is adept at addressing unseen data categories, especially in contexts where data sensitivity and privacy are paramount. The incorporation of DP within the teacher model further enhances data privacy, ensuring that the traditional ZSL generalization properties to unseen classes are preserved without additional training, all while safeguarding data and model privacy.

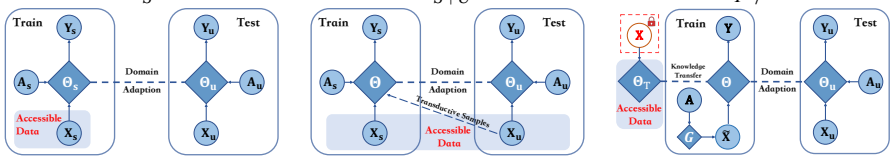
Zero-Shot Learning [188–190, 29, 191] is predicated on recognizing unseen classes by establishing connections between seen and unseen classes through semantic information, such as attributes [192–195], word embeddings [196] and predefined similes [197, 198]. Numerous studies [199–201] have been dedicated to mapping from visual to semantic space, while others [202, 166, 108, 203] focus on generating unseen class data to mitigate data scarcity issues. Effective spaces for visual and semantic embedding have been investigated in [66, 204–207]. Depending on the utilization of unseen data during training, ZSL methods can be categorized into inductive [208, 209] and transductive settings [210, 167]. As for the test phase, conventional ZSL methods [66, 67] operate under the assumption that test data originates exclusively from unseen classes, while Generalized ZSL (GZSL) [68–70] aims to classify both seen and unseen data into their respective classes.

The distinctions between SG-ZSL and traditional ZSL settings are elucidated in Table 6.1. In terms of data access during training, IZSL and Transductive ZSL (TZSL) access labeled seen data and data from both seen and unseen classes, respectively. In contrast, the SG-ZSL setting operates without direct data access, relying solely on a teacher model, trained on sensitive real data, for guidance (as indicated by the red 'X' in Table 6.1). Concerning model security, weight accessibility refers to the accessibility of weights trained on real data. While ZSL models in both inductive and transductive settings possess accessible weights, the SG-ZSL paradigm introduces a teacher model pre-trained on real data. In assessing teacher weight privacy, we introduce the black-box and white-box protocols. In the white-box protocol, teacher weights are accessible for guidance during SG-ZSL model training, whereas the black-box protocol restricts weight sharing, thereby preserving the privacy of both data and model weights.

6.3 Methodology

As depicted in Fig. 6.1, in scenarios where the Data Owner's sensitive data is inaccessible yet a collaboration with the AI Service Provider is sought to leverage the data's value, the

Table 6.1 The distinctions between SG-ZSL and traditional ZSL settings are delineated in the table. Herein, 'S' and 'U' denote the seen and unseen classes, respectively. '$\mathscr{X}$' signifies visual features, while '$\tilde{\mathscr{X}}$' pertains to generated features. The semantics of the seen and unseen classes are represented by 'A_s' and 'A_u', respectively. The red 'X' symbolizes sensitive real data. The ZSL model is denoted by 'θ', whereas 'θ_T' corresponds to the pre-trained teacher model specific to the SG-ZSL task. 'θ_U' can be associated with either the conventional ZSL model or the SG-ZSL model. It should be noted that the SG-ZSL model is constructed under the guidance of the teacher model, effectively eliminating the need for sharing actual data.

	IZSL	**TZSL**	**SG-ZSL**
Accessible Data	$\mathscr{X}_s$	$\mathscr{X}_s \cup \mathscr{X}_u$	**0**
Accessible Weights	θ_S	θ_{S+U}	θ_T / **0**
Paradigm			

proposed SG-ZSL paradigm emerges as a solution. The Data Owner employs a teacher model, serving as a data sentinel, which guides the AI Service Provider's models in training classifiers without real data access. Recognizing the balance between privacy preservation and performance optimization, two distinct training protocols with varying security levels, namely the white-box and black-box protocols, are introduced to enhance the paradigm's adaptability.

6.3.1 Problem Definition

The SG-ZSL paradigm fosters collaboration between the Data Owner, housing a teacher model, and the AI Service Provider, hosting a student model and a generator. The teacher model, represented as $\mathscr{F}_{\theta_T} : \mathscr{X} \to \mathscr{Y}$, serves as a data sentinel. Central to the SG-ZSL paradigm is the utilization of the teacher model at the Data Owner's end to direct the training of the student model at the AI Service Provider's end. This objective is achieved through synthetic data generated by the generator $\mathscr{F}_{\theta_G}$, with the aim of enabling the student model to match the teacher's performance or explore domain not covered by the teacher without the transmission of real data. The objective function is given by:

$$\ell\left(\mathscr{F}_{\{\theta_S,\theta_G\}}\left(\tilde{x}\right), \mathscr{F}_{\theta_T}\left(\tilde{x}\right)\right), \tag{6.1}$$

where ℓ denotes the objective function guided by the teacher, and $\tilde{x} \in \tilde{\mathscr{X}}$ signifies the data generated by the generator, ensuring no real data access.

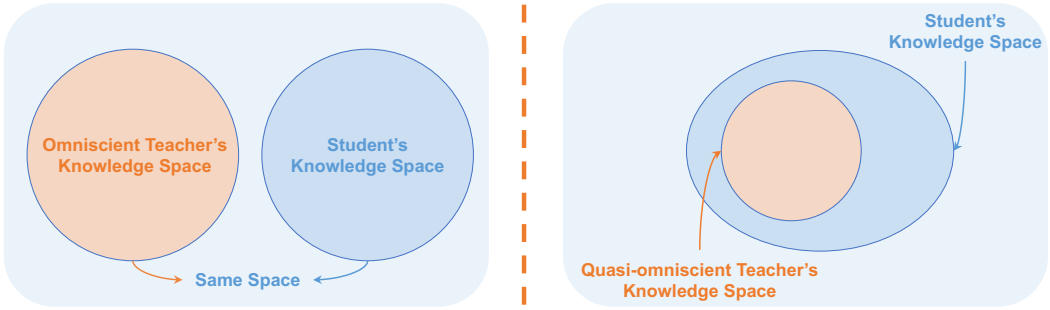


Fig. 6.2 Differences between the Omniscient and the Quasi-omniscient teacher.

6.3.2 Data Sentinel at the Data Owner's End

Omniscient and Quasi-omniscient Teachers

Given the potential inconsistency between the teacher's data categories and the student model's objective categories, there may be unseen class data absent in the teacher's domain but essential for the student model. Thus, teacher models are further categorized into omniscient and quasi-omniscient types as shown in Fig.6.2. The omniscient model encompasses all categories, covering both seen and unseen class data, while the quasi-omniscient model is limited to seen class data.

Here, we define the seen class as $\mathscr{S} = \{(x_s, a_s, y_s) \mid x_s \in \mathscr{X}_s, a_s \in \mathscr{A}, y_s \in \mathscr{Y}_s\}$, where $x_s \in \mathbb{R}^{d_x}$ denotes the d_x-dimensional visual feature in the set of seen class features, $a_s \in \mathbb{R}^{d_a}$ denotes the d_a-dimensional auxiliary class-level semantic embedding, and $\mathscr{Y}_s$ stands for the set of labels for seen classes. Unseen classes are defined as $\mathscr{U} = \{(x_u, a_u, y_u) \mid x_u \in \mathscr{X}_u, a_u \in \mathscr{A}, y_u \in \mathscr{Y}_u\}$, where x_u represents the unseen class features, a_u denotes the semantic embedding of unseen classes and y_u denotes the unseen class labels. The seen and unseen classes are disjoint, *i.e.*, $\mathscr{Y}_s \cap \mathscr{Y}_u = \emptyset$.

In the SG-ZSL paradigm, a key constraint is the inaccessibility of both seen and unseen real features at the Data Owner's end during the student model and generator training at the AI Service Provider's end. The available information for the AI service provider is represented as $\mathscr{T}_r = \{(a, y) \mid a \in \mathscr{A}, y \in \mathscr{Y}\}$, indicating only semantic embeddings a and class labels y are available during training. Additionally, a teacher model, pre-trained on real data, is provided to guide the training of the student model and generator. Depending on the teacher model type, different teacher objectives are considered.

Teacher Objectives

The teacher models guide the student model in mastering various ZSL tasks. For the CZSL task, the student model's objective is to classify test images, represented by $f_{ZSL} : \mathscr{X}_u \rightarrow \mathscr{Y}_u$. For the GZSL task, the student model aims to recognize test images, denoted by $f_{GZSL} : \mathscr{X} \rightarrow \mathscr{Y}$.

Incorporating DP in Teacher Model Training

To bolster the protection of sensitive data at the Data Owner's end, differential privacy techniques are seamlessly integrated into the teacher model's training process. Differential privacy stands as a preeminent mechanism for ensuring data and model security. Denote an algorithm with the differential privacy property by $M(.)$. The algorithm is randomized to make it difficult to have access to the privacy information of the input data. The formal definition of DP is provided below:
Definition 1 [163]. Given a pair of neighboring datasets D and D', for every set of outcomes S, a mechanism M satisfies DP if the following holds:

$$\mathbb{P}(M(D) \in S) \leq e^{\varepsilon} \cdot \mathbb{P}(M(D') \in S) + \delta. \tag{6.2}$$

Here, M(D) and M(D') represent the algorithm's outputs for input datasets D and D', respectively, and $\mathbb{P}$ captures the algorithm's inherent noise randomness. Both ε (privacy budget) and δ (failure probability) influence the privacy strength: smaller values of ε and δ ensure enhanced privacy. In the realm of deep learning, DP is typically realized by introducing the subsampled Gaussian mechanism to safeguard the minibatch gradients during the training process [211–213]. The distinction between deep learning with DP and conventional deep learning hinges on the private release of the gradient. The Gaussian mechanism is defined as:
Definition 2 (Gaussian Mechanism) [212]. Let Δf be the sensitivity of function f, defined as $\Delta f = \max_{D,D'} \|f(D) - f(D')\|_2$. The Gaussian Mechanism, $\hat{f}(D) = f(D) + \sigma \Delta f \cdot \mathscr{N}(0, \mathscr{I})$, is deemed ($\varepsilon$, δ)-differentially private for specific values of ε and δ contingent on σ.

During our teacher models' training, random noise is introduced to perturb the original data distribution, thereby enhancing data privacy. Leveraging the post-processing property of differential privacy, as elucidated in [212], ensures that any subsequent operation on a differentially private output remains privacy-preserving. Thus, data generation under the guidance of the pre-trained teacher model is deemed secure. Specifically, random Gaussian noise is incorporated during the teacher model's training as follows:

$$g_T \leftarrow g_T + N(0, \sigma_n^2 c_g^2 I). \tag{6.3}$$

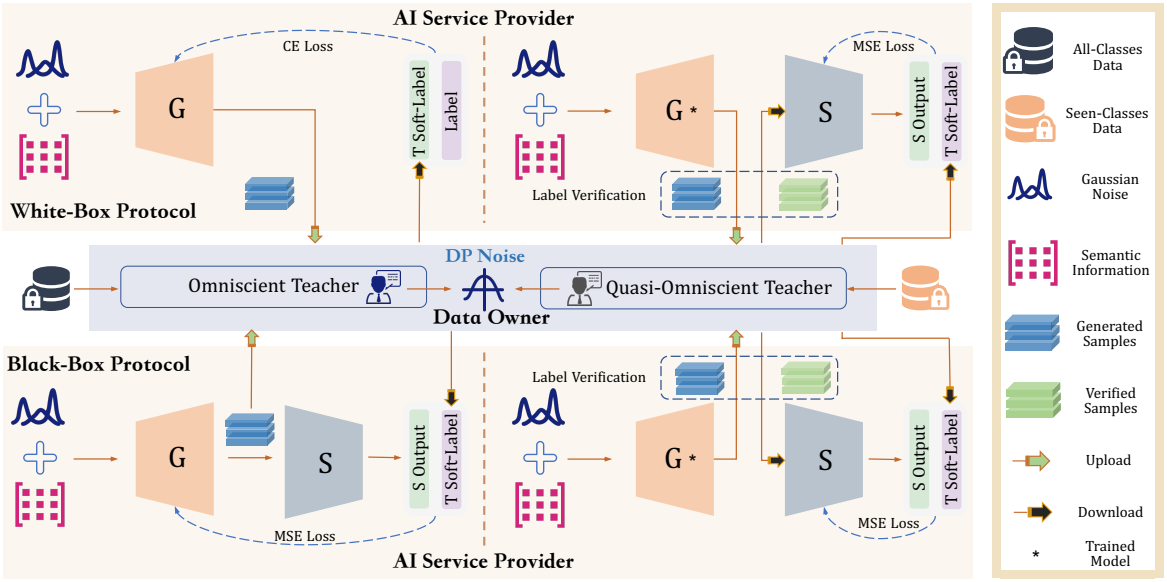


Fig. 6.3 The overarching paradigm for both black-box and white-box protocols. In the white-box protocol, the generator accesses teacher weights during training, whereas in the black-box protocol, only output guidance from the teacher is utilized.

Here, g_T represents the teacher's gradients, σ_n is the noise scale, and c_g signifies the gradient function's sensitivity. Subsequently, the teacher model's weight parameters are updated and truncated within the range (-c, c) to optimize the model:

$$w \leftarrow clip(w + \alpha \cdot Adam(w, g_T), -c, c). \tag{6.4}$$

For practical implementation, we use Opacus [214], Facebook's specialized library for training PyTorch models with differential privacy.

6.3.3 Dual Training Protocols

To address varying privacy and performance needs, SG-ZSL offers two distinct training protocols: the white-box and black-box protocols, each tailored for different levels of security and model interaction. In the white-box protocol, the teacher model provides both gradient and softmax outputs, allowing for direct backpropagation and a more effective generator. Meanwhile, in the black-box protocol, only the softmax output is used, preventing the teacher model's parameters from being exposed or used in backpropagation, which reduces the risk of model leakage. Both protocols employ synthetic data generation via the generator and semantic embeddings to train a student model. The white-box protocol offers more detailed guidance but carries a higher privacy risk, while the black-box protocol ensures stronger

privacy protection by limiting model exposure. Together, these dual training strategies allow SG-ZSL to cater to diverse privacy requirements while ensuring efficient learning without direct access to real data.

6.3.4 Absolute Zero-Shot Classification

In the testing phase, the omniscient teacher, having been trained on both seen and unseen features at the Data Owner's end, facilitates the generator in synthesizing features for all classes. Consequently, the student network is equipped to predict class labels for test features. Given these test features, the predicted class labels are determined as:

$$y^* = \underset{y \in \mathscr{Y}}{argmax}\, p(y|x, \theta_S^*), \tag{6.5}$$

where θ_S^* represents the optimized parameters of the student model.

For the quasi-omniscient teacher model, the challenge confronting the student model intensifies. This heightened challenge arises because, during the training phase, neither the data owner nor the AI service provider possesses information regarding the unseen classes. In the testing phase, an initial step involves synthesizing a data batch for these unseen classes via the generator, denoted as $\tilde{x} = G(z|a; \theta_G^*)$, with z indicating noise and a representing the semantic embedding of the unseen class. Utilizing this synthesized data, the classifier C undergoes training in a supervised learning task with the generated features, as formalized in the following equation:

$$\min_{\theta_C} -\mathbb{E}\left[\log(y|\tilde{x}; \theta_C))\right], \tag{6.6}$$

the function calculates the softmax loss by comparing the predicted label probabilities from synthesized features $\tilde{x}$ against actual labels y to minimize the negative log-likelihood of correct class predictions, optimizing classifier C for accurate unseen class label prediction.

Subsequently, the prediction of class labels for test features is executed as follows:

$$y^* = \underset{y \in \tilde{\mathscr{Y}}}{argmax}\, p(y|x, \theta_C^*), \tag{6.7}$$

where $\tilde{\mathscr{Y}} = \mathscr{Y}_u$ is designated for the conventional ZSL task, and $\tilde{\mathscr{Y}} = \mathscr{Y}_s \cup \mathscr{Y}_u$ for the GZSL task.

In the context of the first SG-ZSL work, this work primarily seeks to address the ensuing research questions:

- **RQ1:** How does the variation in teacher feedback influence the quality and diversity of the synthesized data?

- **RQ2:** How does the alteration in semantic information, when employed as generative conditions, affect the student model's performance?
- **RQ3:** Compare with the traditional ZSL methods, how do the SG-ZSL perform under the black-box and white-box protocols in terms of data privacy, model security, and classification accuracy?
- **RQ4:** Is the student model capable of transcending the constraints of the quasi-omniscient teacher model to generate novel knowledge (on unseen class)?
- **RQ5:** Does the SG-ZSL paradigm, which trains on both seen and unseen classes using synthesized data, enhance the congruence between seen and unseen classifiers in the GZSL challenge? Specifically, is there an improvement over prior ZSL approaches that employed real seen data and synthesized unseen data, potentially introducing a bias towards seen classes?

6.4 Experiments

6.4.1 Datasets

Our SG-ZSL model is evaluated on three benchmark datasets: AWA1 [4], AWA2 [62], and aPY [168]. Both AWA1 and AWA2 encompass 30,475 and 37,322 images, respectively, distributed across 50 classes. The aPY dataset contains 15,539 images spanning 32 classes. For semantic representation, embeddings generated by the BERT language model [109] are adopted, with a consistent dimensionality of 768 across all datasets. The data splits differ based on the type of teacher model. For quasi-omniscient teachers, we adopt the data split proposed in [62], wherein only seen class data is accessible to the teacher. Conversely, the omniscient teacher is trained across all classes. In alignment with prior ZSL studies [172], unseen classes are randomly divided into training and test sets.

6.4.2 Implementation Details

For image representation, 2048-dimensional ResNet101 features [215] are utilized, consistent with [62]. Within our proposed paradigm, all networks are constructed using Multi-Layer Perceptrons equipped with LeakyReLU activations [216]. Both the teacher and student models share the same architecture comprising two hidden layers with 1024 and 512 units, respectively. The generator contains a single hidden layer with 4096 hidden units and its output layer is ReLU. During the training process, we adopt the Adam optimizer and the

learning rate of each network is set to 10^{-5}. The dimension of the noise vector z is a hyper-parameter, which is empirically set to 20 for all datasets. The weight of the regularization term is empirically set to 0.5 for AWA1 and AWA2, and 1 for aPY. A trade-off between accuracy and computational efficiency is taken into consideration when determining the number of generated features. In practice, we generate 400 synthetic features on average per class for all datasets.

6.4.3 Evaluation Protocol

We follow the evaluation metrics proposed in [62]. For conventional ZSL tasks, we use the per-class average top-1 accuracy to evaluate classification performance to alleviate the data imbalance of classes. For the GZSL task, we use harmonic mean $H = (2 \times u \times s)/(u+s)$ for evaluation, where u and s denote average per-class top-1 accuracy on unseen and seen classes, respectively. It is noteworthy that existing methods aim to classify unseen data into corresponding unseen classes in conventional ZSL tasks, while the class space at test time involves both unseen and seen classes in SG-ZSL with the omniscient teacher. This makes SG-ZSL with an omniscient teacher more difficult compared with existing ZSL methods.

6.4.4 Main Results

Comparisons with State-of-Arts

Table 6.2 presents results for both CZSL and GZSL tasks. Given that this is the inaugural SG-ZSL study, a comparison with traditional state-of-the-art methods serves as a reference. The selected methods can be categorized into inductive and transductive ZSL methods. Methods in the upper part of Table 6.2, *i.e.*, IAP, are inductive ZSL methods, which access only labeled seen class data during the training process. The rest of the four methods, *i.e.*, DTN, are transductive methods, which utilize both labeled seen class data and unlabeled unseen class data for model training.

To investigate **RQ1**, we show results under two kinds of feedback from omniscient and quasi-omniscient teachers. SG-ZSL student model with omniscient teacher achieves promising performance in both CZSL and GZSL in the white-box protocol. Our method achieve the best performance in GZSL, especially on aPY, with an increase in harmonic mean of 32.3%, which indicates an improved balance of seen and unseen classes. As for the black-box protocol, the accuracy on unseen classes is 4.9% higher than on seen classes on AWA1. It indicates that the SG-ZSL student model is promising to mitigate the class-level overfitting issue in the GZSL task proposed in **RQ5**. Compared with inductive ZSL methods,

Table 6.2 Comparison results with the state-of-the-art methods in CZSL and GZSL tasks. CZSL measures per-class average top-1 accuracy (T1) on unseen classes. GZSL measures u = T1 on unseen classes, s = T1 on seen classes, H = harmonic mean. 'WB' & 'BB': white- & black-box protocol; 'Om' - omniscient teacher, 'Q-Om' - quasi-omniscient teacher. 'SG-ZSL+WB/BB*' and 'SG-ZSL+WB/BB' represent our model with omniscient and quasi-omniscient teachers, respectively. The best results are in bold.

Method	**Zero-Shot Learning**			**Generalized Zero-Shot Learning**								
	AWA1	**AWA2**	**aPY**	**AWA1**			**AWA2**			**aPY**		
	T1	T1	T1	u	s	H	u	s	H	u	s	H
IAP [4]	35.9	35.9	36.6	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
DAP [4]	44.1	46.1	33.8	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
ALE [66]	59.9	62.5	39.7	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [217]	54.2	59.7	39.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
CONSE [67]	45.6	44.5	26.9	0.4	88.6	0.8	0.5	90.6	1.0	0.0	**91.2**	0.0
ESZSL [209]	58.2	58.6	38.3	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [218]	54.0	46.6	23.9	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
DEM [60]	68.4	67.1	35.0	32.8	84.7	47.3	30.5	86.4	45.1	11.1	75.1	19.4
f-CLSWGAN [108]	68.2	-	-	57.9	61.4	59.6	-	-	-	-	-	-
CE-GZSL [169]	71.0	70.4	-	65.3	73.4	69.1	63.1	78.6	70.0	-	-	-
SDGZSL [170]	-	74.3	47.0	-	-	-	69.6	78.2	73.7	39.1	60.7	47.5
ICCE [171]	74.2	72.7	49.5	67.4	81.2	73.6	65.3	82.3	72.8	45.2	46.3	45.7
DTN [172]	69.0	-	41.5	54.8	88.5	67.7	-	-	-	37.4	87.9	52.5
GMSADE [173]	81.3	80.7	49.9	71.2	87.7	78.6	71.3	86.1	78.0	76.1	39.3	51.8
EDE [174]	**85.3**	77.5	31.3	71.4	**90.1**	79.7	68.4	**93.2**	78.9	29.8	79.4	43.3
BGT [175]	-	**82.4**	49.8	-	-	-	56.2	82.2	66.7	39.3	72.9	51.0
Q-Om Teacher	0.0	0.0	0.0	0.0	92.9	0.0	0.0	93.1	0.0	0.0	91.6	0.0
Om Teacher	92.1	91.7	90.8	92.1	92.5	92.3	91.7	92.2	91.9	90.8	91.4	91.1
SG-ZSL+BB	14.1	19.9	12.3	4.1	3.7	3.9	3.5	3.7	3.6	6.8	4.0	5.1
SG-ZSL+WB	34.5	36.5	18.7	23.4	34.3	27.8	27.3	44.3	33.7	17.9	52.5	26.7
SG-ZSL+BB*	33.5	29.0	30.2	33.5	28.6	30.9	29.0	25.3	27.0	30.2	42.2	35.2
SG-ZSL+WB*	77.9	79.0	**83.9**	**77.9**	81.8	**79.8**	**79.0**	86.7	**82.7**	**83.9**	85.7	**84.8**

results show that our model with the quasi-omniscient teacher in a white-box protocol gains satisfactory performance in GZSL, especially on aPY, with 7.3% higher performance on the harmonic mean compared with non-generative inductive ZSL methods. Despite the quasi-omniscient teacher model's inability to recognize unseen classes and the student model's lack of access to real seen and unseen data, our student model still secures robust accuracy across various ZSL scenarios. For example, it achieves 34.5% accuracy in inductive ZSL settings on AWA1 and a harmonic mean of 26.7% in GZSL on the aPY dataset. This underscores the student model's capacity to extrapolate and generalize from the teacher's knowledge without data exposure, as explored in **RQ4**. Additionally, when contrasted with traditional TZSL methods, our model exhibits significant accuracy enhancements in GZSL, especially for unseen classes (*i.e.*demonstrate a 6.5% and 7.8% improvement on AWA1 and aPY datasets, respectively), and presents a reduced discrepancy between seen and unseen class accuracies, showcasing an advanced ability to mitigate seen class bias as mentioned in RQ5.

For the black-box protocol, results show our SG-ZSL student model outperforms random guessing, which is around 10% on AWA1, AWA2, and 8% on aPY. The white-box protocol demonstrates better performance than the black-box protocol for the student, indicating that gradient guidance provides more information.

Performance vs Paradigm Privacy

Compared to traditional ZSL methods, the performance under the white-box protocol is very promising, since data privacy is already preserved and our model can still achieve adequate performance. Compare with the white-box protocol, the black-box protocol indeed operates under a more constrained information flow, where only softmax outputs from the teacher model are used as pseudo-labels for the student model, without direct gradient exchange. This design choice inherently poses challenges to optimization efficiency compared to direct gradient-based methods. However, this constraint is a deliberate design choice to enhance privacy. Thus, the performance of the black-box protocol is reasonable because both data privacy and model safety are guaranteed as proposed in **RQ3**.

As for model copyright reservation, traditional ZSL methods often involve sharing model details across entities, raising potential issues related to intellectual property and copyright infringement. Our SG-ZSL paradigm circumvents these issues by utilizing a sentinel mechanism that facilitates the learning process without exposing the internal architecture of the models involved. This is achieved by guiding the generation of synthetic data as a medium for communication between the AI Service provider and the Data Owner, enabling both parties without directly sharing the models themselves. This approach ensures that copyright and intellectual property rights are respected and protected, offering a sustainable model for collaborative AI development and usage.

6.4.5 Analysis and Discussion

Feature Generation Regularization Analysis

The key issue in our data-free knowledge transfer framework is to generate high-quality features, which are expected to have a similar distribution to real data. To show the influence of different constraints during the feature generation process, we provide analysis with different regularization terms for generator training in Table 6.3. KL and MMD loss [219] aim to minimize the distribution difference between real and generated features. Results show that adding distribution constraints of synthesized data is beneficial for feature generation. For example, the harmonic mean increases 2.7% and 2.9% with MMD and KL loss respectively compared with the baseline that only contains cross-entropy loss. Besides, results indicate

that KL and MMD loss are both effective and KL loss performs better to a small extent, which shows the effectiveness of KL regularization.

Table 6.3 Experimental results with different constraints for feature generation in GZSL task in the **white-box** protocol. 'CE' represents cross-entropy loss, 'MMD' represents MMD distance loss, and 'KL' represents KL divergence loss.

Method	**AWA2**			**aPY**		
	u	s	H	u	s	H
CE	76.1	83.8	79.8	83.0	84.5	83.7
CE+MMD	**79.9**	85.1	82.5	81.5	85.5	83.5
CE+KL	79.0	**86.7**	**82.7**	**83.9**	**85.7**	**84.8**

Table 6.4 Experimental results with different constraints for feature generation in GZSL task in the **black-box** protocol. 'CE' represents cross-entropy loss, 'MMD' represents MMD distance loss, and 'KL' represents KL divergence loss.

Method	**AWA1**			**AWA2**			**aPY**		
	u	s	H	u	s	H	u	s	H
CE	26.8	26.7	26.8	23.7	23.2	23.4	23.2	25.6	24.3
CE+MMD	31.8	25.3	28.2	**33.8**	20.5	25.5	26.0	36.3	30.3
CE+KL	**33.5**	**28.6**	**30.9**	29.0	**25.3**	**27.0**	**30.2**	**42.2**	**35.2**

We also provide an extensive analysis of the impact of different feature generation regularizations in the black-box scenario in Table 6.4. Similarly, we provide MMD and KL loss as regularization for feature synthesis in the GZSL task as the regularization term is essential for the generalization ability of the SG-ZSL model. The experimental results show that the SG-ZSL model with regularization term outperforms the one with only cross-entropy loss, *i.e.*, with 6% and 10.9% improvement on harmonic mean with MMD and KL loss on aPY, indicating the effectiveness of the constraint for feature generation. Besides, the SG-ZSL model with KL constraint achieves the best performance in harmonic mean, with 4.9% and 2.7% increases on aPY and AWA1 datasets respectively, which indicates that the SG-ZSL model with KL loss can make a better balance between seen and unseen classes.

Teacher Model Privacy Evaluation

Table 6.5 displays the performance corresponding to various privacy budgets ε when DP is incorporated into teacher training. Here, $\varepsilon = \infty$ signifies the baseline non-private performance, i.e., absent DP in teacher training. The results demonstrate that larger ε values correspond to enhanced performances for both the teacher and student models, indicating that a smaller ε yields heightened data security protection. A trade-off between performance and privacy level is observed, allowing for an adjustment of the privacy budget to achieve a balance.

Quality of Generated Features

Fig. 6.4 displays t-SNE visualizations of synthetic unseen features under the white-box protocol guided by the quasi-omniscient teacher model across the AWA1 and aPY datasets.

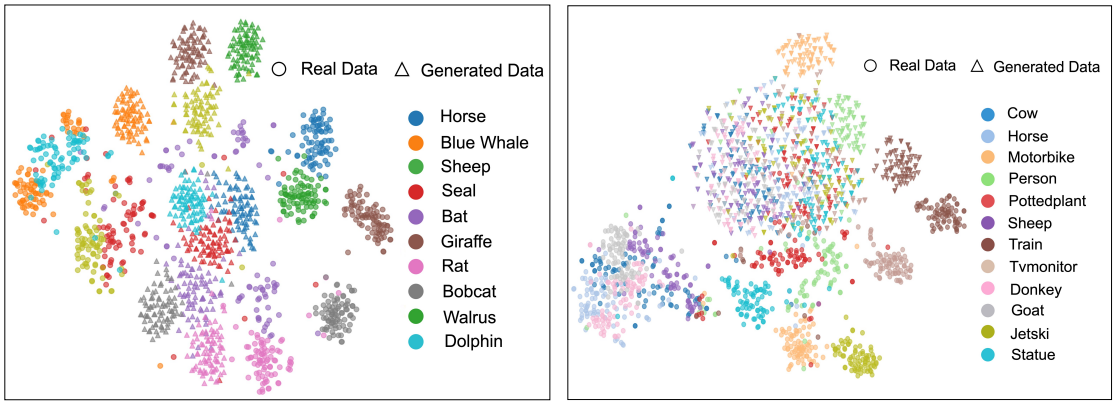
Table 6.5 Results in the white-box protocol with an omniscient teacher under different privacy budgets ε.

Dataset	Accuracy	$\varepsilon = 30$	$\varepsilon = 50$	$\varepsilon = \infty$
AWA1	Teacher Model	56.7	68.4	92.1
	Harmonic Mean	41.7	56.4	79.8
AWA2	Teacher Model	59.1	70.5	91.7
	Harmonic Mean	46.8	60.3	82.7
aPY	Teacher Model	60.6	72.4	90.8
	Harmonic Mean	43.6	62.2	84.8

For clarity, a subset of features is randomly selected for visualization. The unseen class features synthesized under the guidance of the quasi-omniscient teacher, as shown in (c) and (d), demonstrate a slight decline in quality compared to those guided by the omniscient teacher. Specifically, the generated feature distributions are farther from the real data distributions, indicating certain limitations of the generator in capturing unseen class characteristics precisely. However, it's notable that the quasi-omniscient teacher synthesizes these features without access to unseen class information, demonstrating the model's ability to create plausible novel knowledge. This capability is promising for generating meaningful features in Zero-Shot Learning (ZSL) scenarios, even under stringent privacy constraints.

Hyper-Parameter Analysis

We assess the impact of two pivotal hyper-parameters, namely, noise dimension and regularization weight, on our student model. Two ablation studies are conducted on the AWA1 and aPY datasets within a white-box protocol framework, engaging an omniscient teacher, as illustrated in Fig. 6.5. We select four disparate noise dimensions 20, 100, 400, and 768 to elucidate their relationship with the harmonic mean. The findings reveal a performance decrement correlating with the expansion of the noise dimension across both datasets, suggesting that higher-dimensional noise may engender significant interference. Concerning the regularization weight, we designate the values of α as 0.1, 0.5, 1, and 10 for the experimental analysis. As shown in Fig. 6.5, the harmonic mean on both datasets exhibits marginal fluctuation with varying α values. Optimal performance is attained at α values of 0.5 and 1 for AWA1 and aPY datasets, respectively, demonstrating a nuanced interaction between regularization weight and model performance.



(a) AWA1 (b) aPY

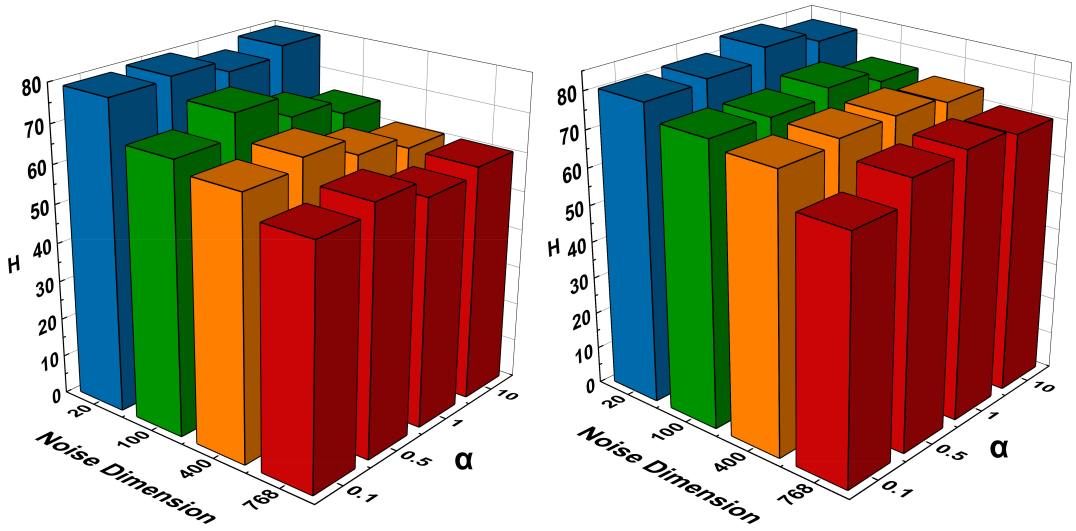
Fig. 6.4 The t-SNE visualizations on AWA1 and aPY datasets under the white-box protocol. The synthetic features in (c) and (d) are generated by the quasi-omniscient teacher-guided generator, illustrating the model's ability to synthesize unseen class data without direct access to unseen information.

Impact of Semantic Information.

We further investigate the influence of various semantic embeddings on the GZSL task. The experimental analysis encompasses three distinct semantic typologies, namely, attributes, Word2Vec, and BERT, serving as the evaluation benchmarks. As delineated in Table 6.6, the comparative outcomes across all three semantic modalities in the GZSL task are relatively aligned, manifesting the robustness of our model with respect to semantic embedding. Notably, the BERT embedding outperforms, signifying the superior efficacy of BERT representation in capturing semantic nuances.

Table 6.6 Experimental results in white-box protocol with omniscient teacher using different semantic information in GZSL task.

Semantics	AWA1			AWA2		
	u	s	H	u	s	H
Attribute	64.7	81.1	72.0	76.8	82.7	79.6
Word2vec	61.6	80.0	69.5	71.4	81.9	76.3
BERT	**77.9**	**81.8**	**79.8**	**79.0**	**86.7**	**82.7**



(a) AWA1 (b) aPY

Fig. 6.5 Noise dimension and parameter α analysis with omniscient teacher in white-box protocol.

Robustness of Student Network

We elucidate the robustness inherent to the student network in this section. Given that the teacher network remains undisclosed by the Data Owner within the black-box protocol, it becomes imperative to showcase the results across diverse student models in this black-box scenario. As illustrated in Table 6.7, the performances across various student models are closely aligned, denoting the stability and consistency afforded by our method.

Table 6.7 Results with different student models in black-box protocol with omniscient teacher in GZSL task.

Student Model	**AWA1**			**aPY**		
	u	s	H	u	s	H
1 hidden layer	31.9	25.9	28.6	**30.4**	34.4	32.3
3 hidden layer	27.9	26.3	27.1	28.5	36.4	32.0
Ours	**33.5**	**28.6**	**30.9**	30.2	**42.2**	**35.2**

6.4.6 Potential Applications

As for potential applications, our SG-ZSL paradigm could carry profound implications for industries where data privacy is paramount. In healthcare, SG-ZSL can facilitate the sharing

of medical insights without exposing patient data, thus advancing research while complying with stringent confidentiality regulations. Similarly, in finance, SG-ZSL enables the collaborative development of predictive models without risking sensitive financial information. Consequently, SG-ZSL fosters a collaborative environment where both data owners and AI service providers can thrive, leveraging the strengths of each party without compromising on security or copyright.

6.4.7 Limitations

Although our research raises awareness of data and model privacy in the ZSL field, balancing privacy with performance remains challenging. The white-box protocol offers high performance through the guidance of teacher model weights and outputs but demands a careful balance between privacy and performance using differential privacy techniques. Meanwhile, the inherently secure black-box protocol may lag in optimization and performance due to its exclusive reliance on output-based supervision. Future efforts aim to bridge these gaps by enhancing the generator's capabilities, notably by incorporating common-sense knowledge from large-scale models to establish a more robust knowledge space, thus improving knowledge transfer from seen to unseen classes.

6.5 Conclusion

In this work, we introduced an SG-ZSL paradigm facilitating through data-free knowledge transfer. A pre-trained teacher model was instantiated at the data owner's end, acting as a data sentinel to render guidance for model training. A thorough evaluation was conducted for both 'black-box' and 'white-box' protocols, elucidating the trade-off between model performance and data privacy. Based on the proposed paradigm, the real data does not participate in the training at the AI service provider end, our model exhibits comparable performance against CZSL and GZSL while the data privacy is also secured. Future advancements in SG-ZSL can explore advanced optimization strategies based on more representative common knowledge (*i.e.*from Large Language Models), and investigate more robust privacy protections, ensuring data owner interests are preserved without compromising model performance.

Epilogue

This chapter addressed RQ 3.2, focusing on the effective use of semantic information for knowledge transfer in zero-shot learning and the optimization of the trade-off between privacy and performance.

To respond to RQ 3.2, we first investigated various types of semantic information, evaluating their role in enhancing knowledge transfer. We found that BERT-generated semantic embeddings significantly improved the recognition accuracy of unseen classes by providing rich contextual representations, thereby facilitating more effective knowledge transfer. These embeddings outperform traditional attribute-based semantics, offering a superior mechanism for bridging the gap between seen and unseen classes in ZSL tasks.

In addressing the issue of bias towards seen classes in traditional ZSL methods, we developed a data-free knowledge transfer approach that reduces the reliance on seen class data. By leveraging Sentinel-Guided Zero-Shot Learning (SG-ZSL), particularly under the black-box protocol, our method improves the accuracy of unseen classes while preventing overfitting on seen class data. This mitigates the inherent bias found in conventional ZSL models and ensures a fairer balance between seen and unseen category recognition, which is crucial in Generalized Zero-Shot Learning (GZSL) tasks.

To tackle privacy concerns and the protection of intellectual property, we incorporated differential privacy (DP) techniques into the training process of the ZSL framework. The integration of DP ensures that the model's outputs are resistant to privacy attacks, allowing sensitive data to remain secure even when model parameters are accessed. By adding controlled noise to the training process, differential privacy guarantees that individual data points cannot be inferred from the model, providing a strong layer of protection for both the data owners and the intellectual property contained within the model.

Additionally, the sentinel mechanism within SG-ZSL allows for secure, collaborative learning by hiding internal model architectures and parameters from external access. This combination of differential privacy and sentinel-guided security ensures that privacy is preserved without compromising performance. It also offers data owners flexibility in balancing the trade-off between privacy protection and model accuracy, allowing for tailored privacy settings in different deployment scenarios.

In summary, the approaches introduced in this chapter successfully optimize the trade-off between privacy and performance in ZSL. By identifying optimal semantic representations, reducing bias, and integrating privacy-preserving techniques such as differential privacy, our work establishes a foundation for secure and efficient knowledge transfer in zero-shot learning. Future research could further refine these methods, expanding their application to more complex and sensitive domains while maintaining robust privacy protections.

Chapter 7

Conclusion and Future Work

In this thesis, we have proposed several innovative frameworks aimed at addressing critical challenges in machine learning, with a specific focus on enhancing data privacy, mitigating data heterogeneity, and optimizing knowledge transfer in decentralized and privacy-sensitive environments.

Key Contributions and Impact

The research presented in this thesis contributes significant advancements in federated learning, zero-shot learning, and video summarization, particularly through the development of privacy-preserving methodologies. The Asynchronous Personalized Federated Learning (AP-FL) framework and Community-Aware Federated Video Summarization (CFed-VS) system are pivotal examples of how federated learning can be extended to complex tasks such as video summarization, which involve large, heterogeneous datasets. By integrating personalized clustering techniques and frame-based aggregation, these models demonstrate improved robustness, efficiency, and adaptability in decentralized settings.

In zero-shot learning, I introduced the Privacy-Enhanced Zero-Shot Learning (PE-ZSL) and Sentinel-Guided Zero-Shot Learning (SG-ZSL) frameworks. These models break new ground by ensuring privacy during knowledge transfer processes, which traditionally expose sensitive data to risks. These innovations contribute to the broader discourse on how machine learning models can effectively generalize to unseen categories while simultaneously preserving the privacy and integrity of the data used for training.

Reflection and Practical Implications

The frameworks developed in this thesis are particularly relevant to domains such as healthcare, surveillance, and personalized content delivery, where both privacy and efficiency are of utmost importance. The research shows that it is possible to achieve a balance between data utility and privacy by leveraging advanced machine learning techniques, such as federated learning and zero-shot learning, while safeguarding sensitive information. Furthermore, the methodologies proposed here lay the groundwork for future applications of privacy-preserving machine learning, potentially transforming how sensitive data is processed across decentralized networks.

Limitations and Future Directions

Despite these advancements, the research in this thesis is not without limitations. One of the key challenges faced was the computational cost associated with implementing certain privacy-preserving algorithms, such as differential privacy in a federated setting. Additionally, while this thesis provides theoretical contributions and demonstrates practical applications, real-world deployment of the proposed frameworks, especially in highly regulated industries, will require further exploration of scalability and security risks.

Future work can extend these frameworks by exploring their adaptability in environments with even more stringent privacy requirements, such as those involving medical data or legal documents. Moreover, expanding the application of these models to other machine learning paradigms, such as reinforcement learning or unsupervised learning, could offer new insights into privacy-preserving techniques for broader classes of problems.

Final Thoughts

In conclusion, this thesis not only addresses current challenges in machine learning but also sets a foundation for future research in creating secure, efficient, and privacy-conscious systems. The contributions made here demonstrate that innovation in machine learning need not come at the cost of data privacy, and that with the right methodologies, both objectives can be pursued in tandem.

References

[1] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[2] L. Chen, P. Bolton, B. R. Holmström, E. Maskin, C. A. Pissarides, A. M. Spence, T. Sun, T. Sun, W. Xiong, L. Yang *et al.*, "Understanding big data: Data calculus in the digital era," 2021.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE TPAMI*, vol. 36, no. 3, pp. 453–465, 2013.

[5] M. A. Smith and T. Kanade, *Video skimming for quick browsing based on audio and image characterization*, 1995.

[6] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European conference on computer vision*. Springer, 2014, pp. 505–520.

[7] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.

[8] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[9] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: an extension of mnist to handwritten letters," *arXiv preprint arXiv:1702.05373*, 2017.

[10] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[11] "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 44, no. 1.2, pp. 206–226, 2000.

[12] "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[13] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[14] "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] A. Gupta and J. Wadhwa, "An analysis of the impact of artificial intelligence on business," *Asian Journal of Management and Commerce*, vol. 4, no. 2, pp. 151–158, 2023. [Online]. Available: https://www.allcommercejournal.com/article/206/4-2-30-400.pdf

[20] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] Y. Zou, A. H. Mhaidli, A. McCall, and F. Schaub, ""' i've got nothing to lose": Consumers' risk perceptions and protective actions after the equifax data breach," in *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, 2018, pp. 197–216.

[23] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[24] F. Wan, J. Wang, H. Duan, Y. Song, M. Pagnucco, and Y. Long, "Community-aware federated video summarization," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[25] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Computing Surveys (Csur)*, vol. 51, no. 4, pp. 1–35, 2018.

[26] A. C. Yao, "Protocols for secure computations," in *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE, 1982, pp. 160–164.

[27] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model." *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[28] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[29] R. Gao, F. Wan, D. Organisciak, J. Pu, H. Duan, P. Zhang, X. Hou, and Y. Long, "Privacy-enhanced zero-shot learning via data-free knowledge transfer," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 432–437.

[30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[31] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[32] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[33] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE journal on selected areas in communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[34] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[35] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "On the convergence of federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, vol. 3, p. 3, 2018.

[36] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

[37] H. Yu, Z. Liu, Y. Liu, T. Chen, M. Cong, X. Weng, D. Niyato, and Q. Yang, "A fairness-aware incentive scheme for federated learning," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 393–399.

[38] K. Pillutla, K. Malik, A.-R. Mohamed, M. Rabbat, M. Sanjabi, and L. Xiao, "Federated learning with partial model personalization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 716–17 758.

[39] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[40] M. Duan, D. Liu, X. Ji, R. Liu, L. Liang, X. Chen, and Y. Tan, "Fedgroup: Efficient federated learning via decomposed similarity-based clustering," in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 228–237.

[41] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[42] C. Feng, H. H. Yang, D. Hu, Z. Zhao, T. Q. Quek, and G. Min, "Mobility-aware cluster federated learning in hierarchical wireless networks," *IEEE Transactions on Wireless Communications*, 2022.

[43] Y. Kim, E. Al Hakim, J. Haraldson, H. Eriksson, J. M. B. da Silva, and C. Fischione, "Dynamic clustering in federated learning," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[44] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[45] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.

[46] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[47] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[48] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[49] X. Wu, F. Huang, Z. Hu, and H. Huang, "Faster adaptive federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 10 379–10 387.

[50] Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive personalized federated learning," *arXiv preprint arXiv:2003.13461*, 2020.

[51] S. M. Shah and V. K. Lau, "Model compression for communication efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 5937–5951, 2021.

[52] Y. Xu, Y. Liao, H. Xu, Z. Ma, L. Wang, and J. Liu, "Adaptive control of local updating and model compression for efficient federated learning," *IEEE Transactions on Mobile Computing*, 2022.

[53] T. Chen, X. Jin, Y. Sun, and W. Yin, "Vafl: a method of vertical asynchronous federated learning," *arXiv preprint arXiv:2007.06081*, 2020.

[54] C. Xu, Y. Qu, Y. Xiang, and L. Gao, "Asynchronous federated learning on heterogeneous devices: A survey," *Computer Science Review*, vol. 50, p. 100595, 2023.

[55] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[56] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Online incremental attribute-based zero-shot learning," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3657–3664.

[57] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," *Advances in neural information processing systems*, vol. 22, 2009.

[58] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.

[59] R. Qiao, L. Liu, C. Shen, and A. Van Den Hengel, "Less is more: zero-shot learning from online textual documents with noise suppression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2249–2257.

[60] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *CVPR*, 2017.

[61] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 487–11 496.

[62] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly," in *CVPR*, 2017.

[63] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7402–7411.

[64] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *CVPR Workshop*, 2018.

[65] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8247–8255.

[66] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *CVPR*, 2013.

[67] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *ICLR*, 2014.

[68] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *ECCV*, 2016.

[69] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *CVPR*, 2020.

[70] Y. Hu, L. Feng, H. Jiang, M. Liu, and B. Yin, "Domain-aware prototype network for generalized zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[71] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[72] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2794–2802.

[73] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," 2017.

[74] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[75] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2604–2613.

[76] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[77] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 3048–3068, 2021.

[78] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3903–3911.

[79] H. Xu, Y. Su, Z. Zhao, Y. Zhou, M. R. Lyu, and I. King, "Deepobfuscation: Securing the structure of convolutional neural networks via knowledge distillation," *arXiv preprint arXiv:1806.10313*, 2018.

[80] K. Yoshida and T. Fujino, "Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks," in *Proceedings of the 13th ACM workshop on artificial intelligence and security*, 2020, pp. 117–127.

[81] J. C. Jiang, B. Kantarci, S. Oktug, and T. Soyata, "Federated learning in smart city sensing: Challenges and opportunities," *Sensors*, vol. 20, no. 21, p. 6230, 2020.

[82] B. Qolomany, K. Ahmad, A. Al-Fuqaha, and J. Qadir, "Particle swarm optimized federated learning for industrial iot and smart city services," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[83] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021.

[84] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, and J. Qadir, "Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge," *IEEE Open Journal of the Computer Society*, vol. 3, pp. 172–184, 2022.

[85] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[86] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Communication-efficient federated learning for digital twin edge networks in industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5709–5718, 2020.

[87] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[88] X. Li, Z. Qu, B. Tang, and Z. Lu, "Stragglers are not disaster: A hybrid federated learning algorithm with delayed gradients," *arXiv preprint arXiv:2102.06329*, 2021.

[89] A. Reisizadeh, I. Tziotis, H. Hassani, A. Mokhtari, and R. Pedarsani, "Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity," *arXiv preprint arXiv:2012.14453*, 2020.

[90] J. Park, D.-J. Han, M. Choi, and J. Moon, "Sageflow: Robust federated learning against both stragglers and adversaries," *Advances in Neural Information Processing Systems*, vol. 34, pp. 840–851, 2021.

[91] Z. Chai, Y. Chen, L. Zhao, Y. Cheng, and H. Rangwala, "Fedat: A communication-efficient federated learning method with asynchronous tiers under non-iid data," *ArXivorg*, 2020.

[92] Z. Chai, A. Ali, S. Zawad, S. Truex, A. Anwar, N. Baracaldo, Y. Zhou, H. Ludwig, F. Yan, and Y. Cheng, "Tifl: A tier-based federated learning system," in *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, 2020, pp. 125–136.

[93] J. Hao, Y. Zhao, and J. Zhang, "Time efficient federated learning with semi-asynchronous communication," in *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2020, pp. 156–163.

[94] A. K. Sahu, S. Ravi, W.-C. Lo, J. Konečnỳ, and H. B. McMahan, "Optimizing federated learning on non-iid data with reinforcement learning," in *International Conference on Learning Representations*, 2020.

[95] P. Fallah, J. Tao, Q. Yang, Y. Liu, and J. Xie, "Adaptive federated optimization," in *NeurIPS*, 2020.

[96] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *IEEE INFOCOM 2022-IEEE conference on computer communications*. IEEE, 2022, pp. 1739–1748.

[97] H. Wang and J. Xu, "Friends to help: Saving federated learning from client dropout," *arXiv preprint arXiv:2205.13222*, 2022.

[98] Y. Wang, J. Wolfrath, N. Sreekumar, D. Kumar, and A. Chandra, "Accelerated training via device similarity in federated learning," in *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*, 2021, pp. 31–36.

[99] M. Rasouli, T. Sun, and R. Rajagopal, "Fedgan: Federated generative adversarial networks for distributed data," *arXiv preprint arXiv:2006.07228*, 2020.

[100] P. K. Chauhan, V. P. K. Kuppili, S. Munaga, G. Saini, and P. Malhotra, "Feddpgan: Federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[101] Y. Liu, B.-G. Kang, J.-G. Song, and D. Lee, "Federated learning for covid-19 detection with generative adversarial networks in edge cloud computing," *Sensors*, vol. 20, no. 21, p. 6164, 2020.

[102] T. Zhao, Q. Li, F. Liu, and T.-Y. Wang, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," in *Advances in Neural Information Processing Systems*, 2018, pp. 2371–2381.

[103] Q. Wang, P. Xu, Y. Zhang, Y. Gao, and G. Sun, "Federated learning with gan-based data synthesis for non-iid," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2020, pp. 1913–1916.

[104] S. Yun, M. Jeong, and J. Yoon, "Ensemble distillation for robust model fusion in federated learning," *arXiv preprint arXiv:2002.06715*, 2020.

[105] X. Wang, P. Peng, K. Yu, Y. Liu, Z. Liu, J. Yang, and T. Abdelzaher, "FedMD: Heterogenous federated learning via model distillation," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 2019, pp. 6388–6398.

[106] M. Mohri, A. T. Suresh, A. Liu, and M. Zinkevich, "Fedaux: Leveraging unlabeled auxiliary data in federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6887–6896.

[107] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.

[108] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *CVPR*, 2018.

[109] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[110] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[111] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2351–2363, 2020.

[112] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[113] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," *arXiv preprint arXiv:2102.02079*, 2021.

[114] X. Gong, A. Sharma, S. Karanam, Z. Wu, T. Chen, D. Doermann, and A. Innanje, "Ensemble attention distillation for privacy-preserving federated learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 076–15 086.

[115] G. Fang, J. Song, C. Shen, X. Wang, D. Chen, and M. Song, "Data-free adversarial distillation," *arXiv preprint arXiv:1912.11006*, 2019.

[116] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7596–7604.

[117] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[118] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[119] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[120] J. Wang, Y. Long, M. Pagnucco, and Y. Song, "Dynamic graph warping transformer for video alignment," 2020.

[121] J. Wang, L. Qin, P. Zhang, Y. Long, B. Hu, M. Pagnucco, S. Wang, and Y. Song, "Towards unified multi-excitation for unsupervised video prediction," 2022.

[122] J. Wang, Z. Sun, Y. Qian, D. Gong, X. Sun, M. Lin, M. Pagnucco, and Y. Song, "Maximizing spatio-temporal entropy of deep 3d cnns for efficient video recognition," *arXiv preprint arXiv:2303.02693*, 2023.

[123] J. Wang, B. Hu, Y. Long, and Y. Guan, "Order matters: Shuffling sequence generation for video prediction," *arXiv preprint arXiv:1907.08845*, 2019.

[124] Y. Bai, J. Wang, Y. Long, B. Hu, Y. Song, M. Pagnucco, and Y. Guan, "Discriminative latent semantic graph for video captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3556–3564.

[125] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE TPAMI*, 2021.

[126] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li, "Zero-shot learning using synthesised unseen visual data with diffusion regularisation," *IEEE TPAMI*, vol. 40, no. 10, pp. 2498–2512, 2017.

[127] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[128] O. Morere, H. Goh, A. Veillard, V. Chandrasekhar, and J. Lin, "Co-regularized deep representations for video summarization," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3165–3169.

[129] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 347–363.

[130] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[131] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[132] J. Wang, Y. Bai, Y. Long, B. Hu, Z. Chai, Y. Guan, and X. Wei, "Query twice: Dual mixture attention meta learning for video summarization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4023–4031.

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

[134] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," *arXiv preprint arXiv:2002.05516*, 2020.

[135] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[136] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[137] H. Duan, Y. Long, S. Wang, H. Zhang, C. G. Willcocks, and L. Shao, "Dynamic unary convolution in transformers," *IEEE TPAMI*, 2023.

[138] H. Duan, S. Wang, and Y. Guan, "Sofa-net: Second-order and first-order attention network for crowd counting," *arXiv preprint arXiv:2008.03723*, 2020.

[139] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[140] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[141] Z. Wang, X. Li, S. Yu, H. Duan, X. Zhang, J. Zhang, and S. Chen, "Vsp-fuse: Multifocus image fusion model using the knowledge transferred from visual salience priors," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[142] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[143] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[144] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.

[145] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "Fedvision: An online visual object detection platform powered by federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13 172–13 179.

[146] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.

[147] M. Xie, G. Long, T. Shen, T. Zhou, X. Wang, J. Jiang, and C. Zhang, "Multi-center federated learning," *arXiv preprint arXiv:2005.01026*, 2020.

[148] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[149] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*, 1967, pp. 281–297.

[150] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the softmax bottleneck: A high-rank rnn language model," *arXiv preprint arXiv:1711.03953*, 2017.

[151] G. Conti, "Analytic combinatorics."

[152] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[153] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[154] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[155] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597, 2020.

[156] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[157] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[158] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[159] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization," *arXiv preprint arXiv:1904.08265*, 2019.

[160] J. Park, J. Lee, I.-J. Kim, and K. Sohn, "Sumgraph: Video summarization via recursive graph modeling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 647–663.

[161] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2793–2801, 2021.

[162] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[163] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[164] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[165] G. Xu, Z. Liu, and C. C. Loy, "Computation-efficient knowledge distillation via uncertainty-aware mixup," *arXiv preprint arXiv:2012.09413*, 2020.

[166] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, and L. Shao, "Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning," *IEEE Transactions on Image Processing*, 2020.

[167] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE TPAMI*, vol. 37, no. 11, pp. 2332–2345, 2015.

[168] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[169] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *CVPR*, 2021.

[170] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *ICCV*, 2021.

[171] X. Kong, Z. Gao, X. Li, M. Hong, J. Liu, C. Wang, Y. Xie, and Y. Qu, "En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning," in *CVPR*, 2022.

[172] H. Zhang, L. Liu, Y. Long, Z. Zhang, and L. Shao, "Deep transductive network for generalized zero shot learning," *Pattern Recognition*, vol. 105, p. 107370, 2020.

[173] O. Gune, M. Pal, P. Mukherjee, B. Banerjee, and S. Chaudhuri, "Generative model-driven structure aligning discriminative embeddings for transductive zero-shot learning," *arXiv preprint arXiv:2005.04492*, 2020.

[174] L. Zhang, P. Wang, L. Liu, C. Shen, W. Wei, Y. Zhang, and A. Van Den Hengel, "Towards effective deep embedding for zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[175] X. Li, D. Zhang, M. Ye, X. Li, Q. Dou, and Q. Lv, "Bidirectional generative transductive zero-shot learning," *Neural computing and applications*, pp. 5313–5326, 2021.

[176] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[177] L. Zhang, G. Gao, and H. Zhang, "Spatial-temporal federated learning for lifelong person re-identification on distributed edges," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[178] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.

[179] H. Ren, J. Deng, X. Xie, X. Ma, and Y. Wang, "Fedboosting: Federated learning with gradient protected boosting for text recognition," *Neurocomputing*, vol. 569, p. 127126, 2024.

[180] R. Yu and P. Li, "Toward resource-efficient federated learning in mobile edge computing," *IEEE Network*, vol. 35, no. 1, pp. 148–155, 2021.

[181] S. Guo, T. Zhang, G. Xu, H. Yu, T. Xiang, and Y. Liu, "Topology-aware differential privacy for decentralized image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 4016–4027, 2021.

[182] N. Fu, W. Ni, S. Zhang, L. Hou, and D. Zhang, "Gc-nldp: A graph clustering algorithm with local differential privacy," *Computers & Security*, vol. 124, p. 102967, 2023.

[183] G. Xu, G. Li, S. Guo, T. Zhang, and H. Li, "Secure decentralized image classification with multiparty homomorphic encryption," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3185–3198, 2023.

[184] H. Liu, X. Zhu, Z. Lei, D. Cao, and S. Z. Li, "Fast adapting without forgetting for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3093–3104, 2021.

[185] K. Xu, L. Wang, J. Xin, S. Li, and B. Yin, "Learning from teacher's failure: A reflective learning paradigm for knowledge distillation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[186] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 925–10 934.

[187] K. Zhang, C. Zhang, S. Li, D. Zeng, and S. Ge, "Student network learning via evolutionary knowledge distillation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2251–2263, 2022.

[188] H. Zhang, Y. Long, Y. Guan, and L. Shao, "Triple verification network for generalized zero-shot learning," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 506–517, 2018.

[189] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks." in *AAAI*, 2008.

[190] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging seen and unseen semantic relationships for generative zero-shot learning," in *ECCV*, 2020.

[191] H. Zhang, H. Mao, Y. Long, W. Yang, and L. Shao, "A probabilistic zero-shot learning method via latent nonnegative prototype synthesis of unseen classes," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2361–2375, 2019.

[192] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *NeurIPS*, 2014.

[193] S. Li, L. Wang, S. Wang, D. Kong, and B. Yin, "Hierarchical coupled discriminative dictionary learning for zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[194] Y. Long, L. Liu, and L. Shao, "Towards fine-grained open zero-shot learning: Inferring unseen visual features from attributes," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 944–952.

[195] Y. Guo, G. Ding, J. Han, and S. Tang, "Zero-shot learning with attribute selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[196] Z. Zhang and V. Saligrama, "Zero-shot recognition via structured prediction," in *ECCV*, 2016.

[197] Y. Long and L. Shao, "Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble," in *WACV*. IEEE, 2017, pp. 907–915.

[198] H. Zhang, Y. Long, and L. Shao, "Zero-shot leaning and hashing with binary visual similes," *Multimedia Tools and Applications*, vol. 78, pp. 24 147–24 165, 2019.

[199] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao, "Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition." *IEEE SPL*, vol. 23, no. 11, pp. 1667–1671, 2016.

[200] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *CVPR*, 2017.

[201] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *ECCV*, 2018.

[202] J. Wang, Y. Jiang, Y. Long, X. Sun, M. Pagnucco, and Y. Song, "Deconfounding causal inference for zero-shot action recognition," *IEEE Transactions on Multimedia*, 2023.

[203] D. Cheng, G. Wang, B. Wang, Q. Zhang, J. Han, and D. Zhang, "Hybrid routing transformer for zero-shot learning," *Pattern Recognition*, vol. 137, p. 109270, 2023.

[204] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *CVPR*, 2015.

[205] Y. Tian, Y. Kong, Q. Ruan, G. An, and Y. Fu, "Aligned dynamic-preserving embedding for zero-shot action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1597–1612, 2019.

[206] Y. Liu, Q. Gao, J. Li, J. Han, and L. Shao, "Zero shot learning via low-rank embedded semantic autoencoder." in *IJCAI*, vol. 8, no. 9, 2018, p. 10.

[207] Y. Liu, X. Gao, J. Han, L. Liu, and L. Shao, "Zero-shot learning via a specific rank-controlled semantic autoencoder," *Pattern Recognition*, vol. 122, p. 108237, 2022.

[208] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in *CVPR*, 2017.

[209] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *ICML*, 2015.

[210] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, "Transductive unbiased embedding for zero-shot learning," in *CVPR*, 2018, pp. 1024–1033.

[211] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Automatic clipping: Differentially private deep learning made easier and stronger," *arXiv preprint arXiv:2206.07136*, 2022.

[212] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[213] L. Zhang, B. Shen, A. Barnawi, S. Xi, N. Kumar, and Y. Wu, "Feddpgan: federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1403–1415, 2021.

[214] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao *et al.*, "Opacus: User-friendly differential privacy library in pytorch," *arXiv preprint arXiv:2109.12298*, 2021.

[215] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[216] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," in *Proc. ICML Deep Learning Workshop*, 2015.

[217] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013.

[218] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *CVPR*, 2016.

[219] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.