Robust Knowledge Adaptation for Federated Unsupervised Person ReID

Jianfeng Weng

This thesis is presented as part of the requirements for the conferral of the degree:

Master of Philosophy (Engineering)

Supervisor:
A/Prof. Zhiyong Wang

Auxiliary Supervisor:
Prof. Dacheng Tao

The University of Sydney
School of Computer Science

2023

Declaration

I, *Jianfeng Weng*, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree *Master of Philosophy (Engineering)*, from the University of Sydney, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Jianfeng Weng

September 6, 2023

Abstract

Person Re-identification (ReID) has been extensively studied in recent years due to the increasing demand in the public security sector. However, collecting and modelling with sensitive personal data raises privacy concerns. Therefore, federated learning has been studied for Person ReID, which aims to share minimal sensitive data between different parties (clients). However, the statistical heterogeneity between client domains under a federated setting remains as a challenge, which limits the process of knowledge aggregation across clients and often leads to inferior identification accuracy. Additionally, existing federated learning-based person ReID methods generally rely on laborious and time-consuming data annotations, which has the scalability issues to deploy them for real-world Person ReID. Therefore, this thesis aims to address the unsupervised person ReID problem under a federated learning scheme. Specifically, two methods are devised:

- In Chapter 3, we introduce a federated unsupervised cluster-contrastive (FedUCC) learning method for Person ReID. FedUCC introduces a three-stage modelling strategy following a coarse-to-fine manner. In detail, generic knowledge, specialized knowledge and patch knowledge are discovered using a deep neural network. This enables mutual knowledge sharing among clients while retaining local domain-specific knowledge based on the categories of the network components and their parameter settings.

- In Chapter 4, we propose a novel deep transformer-based architecture - *context* and *camera* invariant transformer, namely C^2IT, in pursuit of homoge-

neous image representations by ignoring the irrelevant *context* and *camera* bias. First, regarding the contents within images, a contrastive learning-based context-invariant learning strategy is proposed to identify clustered transformer tokens that are not closely associated with the ReID task as a context token and to disentangle them from final image representations. Second, a camera-invariant learning strategy is further proposed to mitigate the difference incurred by camera-specific characteristics, where camera tokens are introduced and a uniformity loss is devised to reduce variations caused by cameras.

The proposed methods are capable of establishing a universal model to formulate generalizable Person ReID representations and deliver consistent and robust performance across diverse client domains. Comprehensive experiments demonstrate the effectiveness of these methods, which show the state-of-the-art performance on eight public datasets.

Acknowledgements

First and foremost, I express my heartfelt gratitude to my supervisor, Associate Professor Zhiyong Wang, for his unwavering encouragement and guidance throughout my Master's journey. His perpetual support has been instrumental in helping me overcome various difficulties and pursue challenging problems. Then, I would like to express my gratitude to my co-supervisor Professor Dacheng Tao, who has also served as my supervisor during my honours degree studies at the University of Sydney. His valuable support and guidance were instrumental in helping me pursue my Master's degree. I am also deeply grateful to Assistant Professor Jingya Wang, for her invaluable support and critical feedback. Her mentorship has helped me navigate the many nuances of my research and has played a significant role in shaping my academic growth. I want to extend my sincere thanks to Dr.Kun Hu for his invaluable suggestions and consistent support throughout my research work. His expertise and insights have been a source of inspiration and guidance, and I am truly grateful for his suggestions for my Master's thesis. I would like to thank all my colleagues and friends at the University of Sydney for their unwavering friendship and support. Finally, I would like to thank my parents Yazhen Lin and Qizhi Weng for their love and support over the years.

JianFeng Weng

Authorship Attribution Statement

Title of the paper (Chapter 3)	Robust Knowledge Adaptation for Federated Unsupervised Person ReID
Publication status	Published
Publication details	Weng, Jianfeng and Hu, Kun and Yao, Tingting and Wang, Jingya and Wang Zhiyong. "Robust Knowledge Adaptation for Federated Unsupervised Person ReID". In: Digital Image Computing: Techniques and Applications (DICTA). 2022 **Awards:** DICTA 2022 APRS/IAPR Best Paper Award
Contribution	85% Conceptualisation, literature review, data collection, experiment (code) setup and modification, result analysis, writing the manuscript, reviewing and editing the manuscript.

Title of the paper (Chapter 4)	CCIT: Context and Camera Invariant Transformer for Federated Unsupervised Person ReID
Publication status	Under review for publication
Publication details	Weng, Jianfeng and Hu, Kun and Yao, Tingting and Wang, Jingya and Wang Zhiyong. "CCIT: Context and Camera Invariant Transformer for Federated Unsupervised Person ReID". 2023
Contribution	85% Conceptualisation, literature review, data collection, experiment (code) setup and modification, result analysis, writing the manuscript, reviewing and editing the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Student Name: JianFeng Weng

Student Signature:

Date: 07/03/2023

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Zhiyong Wang

Supervisor Signature:

Date: 07/03/2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background & Research Motivation

Person Re-identification (ReID) focuses on identifying a person across multiple cameras. This has become a highly important research field in the past decades due to its potential for many real-world security applications [75]. Person ReID provides the capability of tracking, and timely identifying and preventing crimes, which is critical for public safety. Recently, with the advanced deep learning techniques and the availability of large-scale datasets, person ReID has a rapid development [46, 67, 37].

Despite the continual breakthrough, deploying accurate and stable Person ReID solutions in real-world still remains a challenging problem [72]. Firstly, the demands of annotated training data is unrealistic, which is time-consuming and labour-intensive to obtain [31]. This encourages the development of unsupervised learning approach for person ReID, which does not involve labelled images for training [40, 31, 41]. However, purely unsupervised learning naturally suffers from the noise pseudo-labelling, which hampers the learning of discriminative features. Many efforts have been made to exploit the knowledge learned from labelled source datasets and assisted in the feature learning on unlabelled target datasets [5, 18, 74]. However, due

to the increasing awareness of the privacy for these sensitive data, accessing out-of-domain datasets can be infeasible. Therefore, federated learning scheme becomes attractive for unsupervised Person ReID [81]. However, current unsupervised Person ReID research utilizing federated learning fails to address the domain discrepancy issue between clients for achieving optimal pseudo-label prediction accuracy [81].

Therefore, this thesis aims to address the issue of domain discrepancy among person ReID datasets for modelling in the context of unsupervised person ReID. Specifically, this thesis explores both personalized model learning and global model learning approaches to address the data discrepancy challenge. For personalized model learning, it focuses on disentangling domain-specific knowledge from generic knowledge for cross-domain knowledge aggregation. Generally, domain-specific knowledge learned from a dataset can negatively affect a model's capability to distinguish samples in a different data domain. However, this domain-specific knowledge can be extremely beneficial for the task within the associated dataset. Following this intuition, in Chapter 3, we propose a simple yet effective strategy for learning specialized models which are in pursuit of personalized models for different clients.

On the other hand, global model learning focuses on learning a single global model that is robust towards all client domains. However, enforcing the conflictual domain-specific knowledge fusion is difficult especially with a conventional scheme such as a simple parameter aggregation. To enable more efficient knowledge aggregation, we take into account the influence of context (background) information and camera bias for image feature generation. Specifically, a novel deep transformer-based architecture - *context* and *camera* invariant transformer is proposed in Chapter 4, namely C^2IT. First, regarding the contents within images, a contrastive learning-based context-invariant learning strategy is proposed to identify clustered transformer tokens that are not closely associated with the Person ReID task as a context token and to disentangle them from image representations. Second, a camera-invariant learning strategy is further proposed to mitigate the difference incurred by camera

properties, where camera tokens are introduced and a uniformity loss is devised to reduce variations caused by cameras.

The proposed methods are capable of establishing a universal model to formulate generalizable Person ReID representations and deliver consistent and robust performance across diverse client domains. Comprehensive experiments demonstrate the effectiveness of these methods, showing the state-of-the-art performance on eight public datasets.

1.2 Summary of Contributions

The contribution of this thesis are as follows,

1. A federated unsupervised cluster-contrastive (FedUCC) learning method is proposed for Person ReID. FedUCC introduces a three-stage modelling strategy following a coarse-to-fine manner. In detail, generic knowledge, specialized knowledge and patch knowledge are discovered using a deep neural network. This enables mutual knowledge sharing among clients while retaining local domain-specific knowledge based on the kinds of network layers and their parameters.

2. To alleviate the varied data distribution issue for federated person ReID, a novel deep transformer-based architecture - *context* and *camera* invariant transformer is proposed, namely C^2IT, in pursuit of homogeneous image representations by ignoring the irrelevant *context* and *camera* bias. First, regarding the contents within images, a contrastive learning-based context-invariant learning strategy is proposed to identify clustered transformer tokens that are not closely associated with the ReID task as a context token and to disentangle them from image representations. Second, a camera-invariant learning strategy is further proposed to mitigate the difference incurred by camera properties, where camera tokens are introduced and a uniformity loss is devised to reduce variations caused by cameras.

1.3 Thesis Structure

The rest of the thesis is organized as follows:

Chapter 2 provides the background and the related studies for Person ReID and federated learning.

Chapter 3 proposes a CNN-based method to iteratively learn coarse-to-fine representations by exploiting a three-stage modelling strategy.

Chapter 4 devises a transformer-based method that is based on token feature learning to reduce domain-specific bias.

Chapter 5 concludes this thesis with a discussion of future work.

1.4 List of abbreviations

Table 1.1: Enumerated below are the prevalent symbols employed within Chapter 3.

Abbreviation	Meaning
K	number of clients in the federated learning
n^k	number of images in client k
X^k	dataset within client k
x	training image
u	2-d feature representation of training image before classifier
v	3-d feature representation of training image before GAP layer
$\theta(x, y)$	deep learning model parameterized by x and y
W_g	shared generic parameter across clients
W_g	client-specific parameters
$\mathcal{H}^k$	set of clustering results derived from k-th client dataset
c	cluster centroid
q	patch-level feature representation
p	number of horizontally stripped features derived from one image feature
D	set of patch-level cluster centroids within one cluster

Table 1.2: Enumerated below are the prevalent symbols employed within Chapter 4.

Abbreviation	Meaning
N	number of clients in the federated learning
C_i	i-th client in the federated learning
D_i	dataset within client i
x	training image
y	training image feature representation (output embedding token)
R	total number of federated learning optimization rounds
$f_i^r(x\|\mathbf{\Theta}_i^r)$	r-th round local model optimized by the i-th client, parameterized by $\mathbf{\Theta}_i^r$
c	cluster centroid
p	number of transformer input/output tokens
T	set of input patch tokens for a transformer
Y	set of output patch tokens from a transformer
ρ	local density value of a training image
δ	distance score of a training image
y^{emb}	embedding token of a training image
y^{context}	context token of a training image
Q_i	number of cameras on the i-th client
$y^{\text{camera,q}}$	q-th camera token of a training image

Chapter 2

Literature Review

2.1 Supervised Person Re-identification

Person ReID aims to identify the same individual across a network of cameras. Various studies in this field have been proposed with a metric learning scheme to formulate Person ReID representations, where the representations are with high similarity between the same person and the representations are dissimilar between different people. Specifically, different losses for metric learning are explored to learn appropriate feature representations between images. For example, verification losses [77, 4] were studied to learn from pairs of images by identifying their correspondence. Contrastive loss [63, 53] was adopted for pair-wise learning of images by the constraints on the feature space with the guidance of positive and negative image pairs. Triplet loss [25, 49] extended the learning to address the distance of an anchor image to both a positive and a negative images simultaneously. Various methods further combined these metric learning losses to address the challenges in Person ReID considering the variations in pose, illumination, occlusion, and image background across different cameras. Specifically, human body pose information was adopted for learning robust and discriminative features with an enhanced attention on human body regions [54, 17, 68]. Predefined strip segmentation on the

input image or feature maps were utilised to extract fine-grained local features [56, 42, 57] and helped alleviate the widely happened occlusion issue [73]. Compared with the aforementioned studies, which mostly addressed the Person ReID problem in a supervised learning manner, this thesis addresses an increasing demand for Person ReID: unsupervised multi-domain learning under a decentralized setting. In addition to the challenge of image variation across cameras, additional issues are explored including domain discrepancies arising from differences between datasets, as well as the accurate assignment of pseudo-labels for training images.

2.2 Unsupervised Person ReID

Unsupervised person ReID has received increased attention in recent years due to its less reliance on human annotations [66, 27, 64, 18]. Generally, these methods can be categorised into two streams: fully unsupervised methods [27, 66] and unsupervised domain adaptation (UDA) [64, 5, 18].

2.2.1 Fully Unsupervised Person ReID

Fully Unsupervised Person ReID can be more challenging as there is no prior knowledge (i.e., label annotations). Clustering-based methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-mean were used to estimate image identity relationships [28, 41], which introduced pseudo-labels for the images based on the clustering results. However, noises can be found because of the misclustered images, which degrades the accuracy. A number of methods were proposed to address this by aligning image feature representations towards cluster centroids using contrastive learning strategies [23, 6, 61]. Thus, the mislabeled images of the same identity are not optimized to drift apart in the latent space. Specifically, contrastive learning involves the construction of a memory bank with a mean image representation for every cluster. Then, the loss function is adopted to

minimise the image distance to its associated cluster representation, while the distances to the rest of cluster representations are maximized. A discrepancy among camera views was considered in [61], which further split a single cluster into multiple camera-specific centroids for inter-camera and intra-camera representation learning. BUC (Bottom Up Clustering) [41] adopted a hierarchical clustering approach to iteratively associate similar images, which started with each image to build a single cluster, and terminated until a predefined number of clusters had remained. Besides clustering, multi-labels were used to consider the close-identity relationship for image optimization [60], thus compensating the potentially wrongly assigned labelling. For unsupervised learning in the federated setting, FedUReID [81] followed the BUC [41] strategy to iteratively associate neighbouring clusters, and designed the cluster stopping criteria according to the batch precision score. However, the positive correlation between clustering accuracy and batch precision score cannot be guaranteed as batch precision scores are highly sensitive to the distribution of a dataset. Thus, both under-clustering and over-clustering issues can be found. Recently, transformer-based methods were studied for unsupervised person ReID [44] under the self-supervised learning framework. However, the lack of reliable supervision still limits the discriminability of the learned representations.

2.2.2 Unsupervised Domain Adaptation for Person ReID

Unsupervised domain adaptation for Person ReID involves knowledge transferring between the datasets of two domains: a labeled source domain and an unlabeled target domain. Given a source domain with labeled data, the knowledge from the source domain is exploited to supervise the learning on an unlabeled target domain. The major challenge is caused by the data distribution (e.g. image style), which may differ significantly between the two different domains. Existing studies have explored various transfer learning techniques. In [64, 5, 9], domain gaps were minimised by transferring the source image style to the target image style by using a GAN, while

preserving the contents. Specifically, [5] employed an object segmentation model to accurately separate the foreground and background of the target image, and by combines it with the source image to synthesis a background-transferred image. This approach enables the produced images to have a consistent content style. Additionally, knowledge distillation has been utilised to offer guidance during the training on unlabelled datasets [18]. Typically, this process involves soft targets generated by a pre-trained teacher model to optimize a student model being trained on an unlabeled dataset. This technique allows the student model to learn from the knowledge distilled by the teacher model, which can lead to better generalization and improved performance on the unlabeled data. Meta-learning is also a strategy used for obtaining generic feature that is transferable to an unlabelled domain [74]. However, the need for computing the second-order derivatives entails a significant computational cost. While unsupervised domain adaptation methods typically are more effective than pure unsupervised learning, the knowledge transferring procedure heavily relies on the similarity between the source and target domains. Simultaneously accessing multiple datasets increases the risk of sensitive information leakage.

2.3 Federated Learning and Person ReID

2.3.1 General Federated Learning Methods

Federated learning addresses decentralized training and reduces the risk of privacy leakage. Its intuition is to conduct knowledge exchanges between the clients with abstract information rather than raw inputs, which contain the sensitive information. For example, modelling weights and gradients can be transferred between clients for this purpose [45]. The prevailing majority of existing federated learning methods are formulated within a supervised learning framework. [43, 82, 71, 52], to address the issue of statistical heterogeneity. A widely used strategy is to introduce regularization terms to ensure the similarity between the local models and the aver-

aged global model [33, 32], which used distance metrics such as ℓ_2-normalization [33] and image representation similarity [32]. While this can limit the impact of local updates and alleviate the client drift problem, it can restrict the model's ability to achieve maximum performance on all clients. Meta-learning techniques such as MAML (Model-Agnostic Meta-Learning) [15] were used to obtain an initial global model that can adapt quickly to a new heterogeneous task with few local gradient descents [13]. Feddg [43] attempted to share local image distributions across clients by first decoupling style and content patterns using Fourier Transform [47]. Then, re-synthesizing images with these patterns for image style transfer allowed local models to train on augmented images with the distributions across clients. However, the Fourier transform is only capable of decoding generic image colour and texture, which is difficult to formulate the statistical heterogeneity between Person ReID datasets. Compared with the abovementioned approaches which attempt to learn a strong global model that adapts to all clients, the model personalization approach trains individual models for each client to focus on client-specific patterns [36, 1]. Instead of conducting model aggregation to exchange cross-client knowledge, knowledge consensus is obtained by gathering image predictions on a public dataset and using knowledge distillation on averaged image predictions to conduct a knowledge exchange [30]. Alternatively, parameter personalization approaches are proposed to store generic and specialized knowledge on separate parameters of the model, and the client-specific model is reformulated by integrating generic parameters and client-specific parameters. [36, 1, 38, 51]. The selections of personal parameters were investigated in various works, including top layers [1] and bottom layers [38] of the model network, as well as batch normalization [36] and convolutional channel parameters [51], which are evenly distributed across the model network depth. These methods, however, fail to address potential domain over-fitting, in which local specialized knowledge overtakes generic knowledge, causing performance degrading for the more biased clients. Another direction explores a client-clustering approach to

associate clients of similar data distribution for collaborative learning [19, 20]. As a result, multiple central models will be established and each client will update and synchronize the model parameter toward one of the assigned central models. For instance, upon receiving multiple central models from the server, [19] determined the most appropriate central model by selecting the one with the minimum empirical loss value for a local image batch. This approach can estimate the similarity between clients based on the proximity of their central models, thereby facilitating collaborative learning among similar clients. However, determining the appropriate number of central models can be challenging, and the approach for estimating client-to-client closeness may not always be accurate.

2.3.2 Federated Person ReID

As Person ReID involves sensitive personal information, various studies proposed federated learning methods for Person ReID using data from multiple sources to reduce the risks of privacy leakage [55, 81, 65]. However, due to the domain discrepancy issue, local trained Person ReID models have their own weights, which makes it challenging to obtain a global model for optimal performance across all the clients. A number of existing methods have been proposed to address this issue and can be organised into two approaches: unified learning and personalized learning methods. Unified learning formulates a single shared global model with reduced domain discrepancy for all clients. [82] used a public dataset for an additional post-modelling stage to aggregate knowledge from clients. [32] exploited a contrastive learning scheme to enforce the similarity of local-global model distance. Personalized learning formulates both local and global knowledge for better domain adaptation. [55] used separate weights to encode general and personalized knowledge, where personalized weights are kept locally and combined with the general global weights for local modelling. The aforementioned methods explored federated settings for supervised person ReID, whilst the unsupervised method was studied recently [81],

which proposed a personal update mechanism that interpolates the global and local model weights to update the models and mitigate the heterogeneity. However, the proposed personalized model learning is yet to resolve the domain over-fitting issue. In Chapter 3, we introduce a patch-level feature alignment with fine-feature consistency across clients. Moreover, note that these methods are all based on conventional CNN architectures, which tend to over-fit on local high-frequency patterns [48]. This could result in heterogeneous local models and negatively impact the aggregation of the global knowledge for person ReID. In Chapter 4, we investigate the transformer-based architecture to alleviate such an issue.

As vision transformers have achieved promising performance in various computer vision tasks, it has been explored for person ReID as well [24, 39, 70]. Specifically, [24] first adopted transformer architecture for person ReID. It introduced a set of learnable side information embeddings (SIE), which encodes camera side patterns and appends them to the image tokens to mitigate feature divergence caused by scene bias. [39] merged tokens that encode similar background features and used the rest tokens to characterise more informative regions. However, the background information still contributes to the formulation of holistic image representations. Similarly, [70] filtered out the less informative tokens which are less relevant to image-level person ReID knowledge. However, the ratio of such tokens is predefined, which limits its capability for generalisation. Note that these methods are for supervised person ReID, whilst few studies involve transformers for federated unsupervised person ReID and proper methods are required for accurate federated modelling.

Chapter 3

Robust Knowledge Adaptation for Federated Unsupervised Person ReID

3.1 Introduction

Decentralised learning paradigm, specifically federated learning, has gained increasing attention recently. It allows models to be trained collaboratively across multiple sources or clients without data centralisation [45]. Figure 3.1 illustrates the comparison between centralised learning and federated learning. FedAVG [45], for example, aggregated the model parameters at a central server after a number of local epochs, and the new global model is re-distributed to clients as new parameters for local optimization. Using this simple approach, multiple clients can collaborate without explicit data sharing. Hence, federated learning has been widely utilised in various applications where data privacy is seriously concerned, such as facial expression recognition [52] and medical image segmentation [43].

For Person ReID, a number of recent studies have explored federated learning strategy [81, 82, 71, 55, 65]. For example, FedReID [82] adopted FedAvg to im-

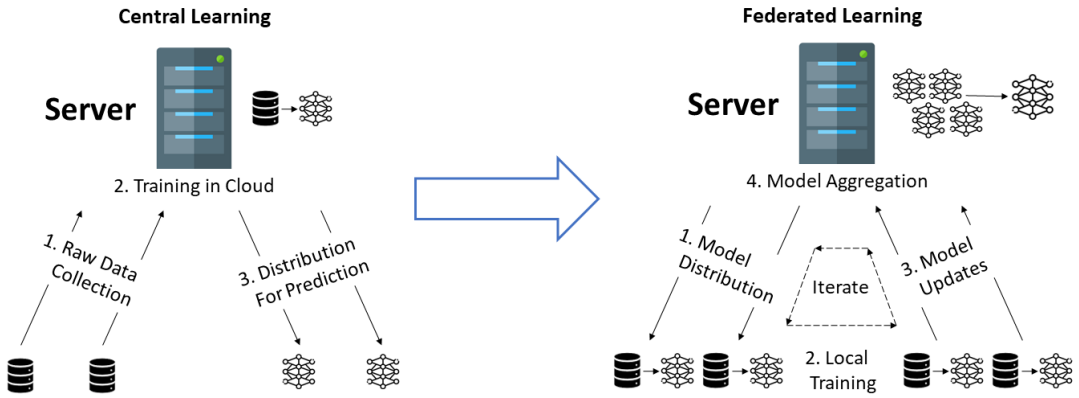


Figure 3.1: Illustrative comparison between centralised learning pipeline and decentralised pipeline of federated learning.

prove the generalizability of learned models via 9 multi-domain datasets. However, most existing methods typically assume that the dataset of each client is annotated, which can be time-consuming and laborious to obtain. This limits the scalability for large-scale real-world deployment. FedUReID [81] attempted to resolve this issue by introducing an unsupervised learning method for federated Person ReID. Specifically, hierarchical clustering was used to progressively merge neighbouring clusters during training. Nonetheless, determining the stopping criteria for clustering on each client is challenging, as the number of images and identities often varies widely. In addition, the variation between clients could be impacted by statistical heterogeneity, such as label skew, feature skew and concept shift [29]. Existing studies often fine-tune the global model using public datasets [65], and utilise adversarial learning [71] to alleviate such a issue. However, a unified model is still not sufficient to perform well in this scenario, especially when considering the nature of the extreme statistical heterogeneity in the person ReID task. Model personalization methods attempted to incorporate generic and specialized knowledge using different model parameters [1, 38, 36, 51]. However, these methods have not addressed the possible domain over-fitting issues yet, with the potential generic knowledge overwhelmed by potential local knowledge bias.

To address the aforementioned challenges, a three-stage learning strategy in a

coarse-to-fine manner is proposed for federated unsupervised learning-based Person ReID. First, the task is modelled with generic knowledge by a conventional federated learning scheme based on DBSCAN [12] clustering. This stage aggregates and distributes all parameters from local clients. Second, specialized knowledge is explored in pursuit of client personalization. This is achieved by decoupling the client-specific knowledge from generic knowledge via parameter localization. Third, we further enhance fine-grained patterns by incorporating patch-level feature alignment across clients. Unlike holistic person images, which present potentially large variations from changing poses and view angles, patch-level information provides fine and consistent person ReID representations. Finally, comprehensive experiments and analyses have been conducted to demonstrate the state-of-the-art performance of our proposed FedUCC on 8 public Person ReID datasets.

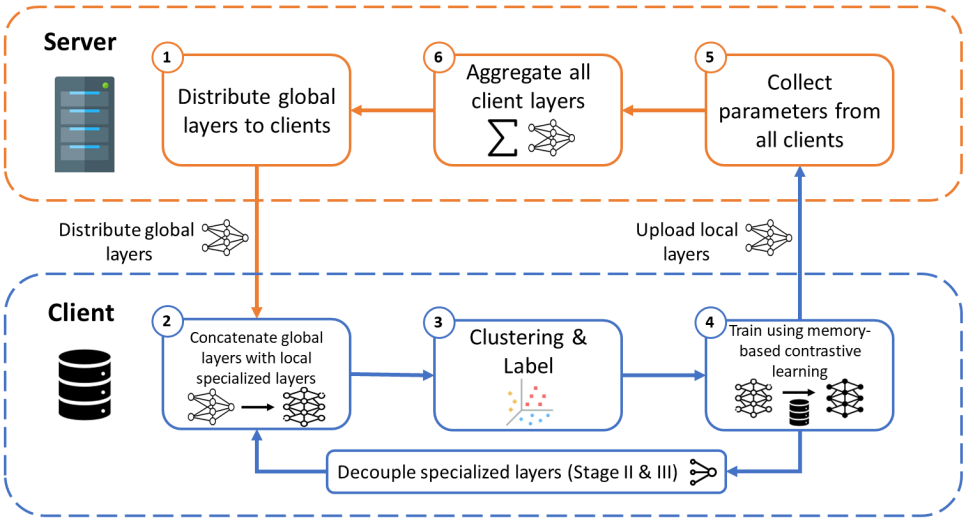
In summary, the major contributions of this chapter are three-fold as follows:

- A novel unsupervised federated learning method is proposed for Person ReID with a coarse-to-fine strategy including three learning stages.
- A novel strategy to explore client-specific knowledge by incorporating patch-level representations.
- Comprehensive experiments demonstrate the state-of-the-art performance of the proposed method on 8 benchmark datasets.

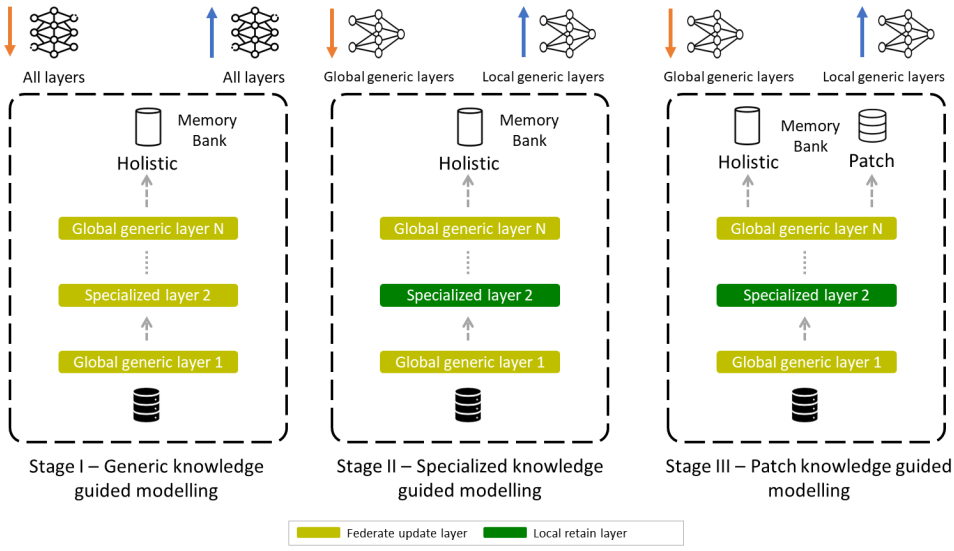
3.2 Proposed Method

3.2.1 Overview

Fig. 3.2a illustrates the overview of the proposed federated learning method - FedUCC. The overall architecture is composed of a central server and a number of clients, each of which has a multi-camera ReID dataset captured at different geographical locations. The task of the central server is to aggregate the trained



(a) A global round contains local training, global aggregation and distribution.



(b) Three-stage local training.

Figure 3.2: Illustration of the proposed FedUCC method: (a) a global round, and (b) a three-stage strategy for local training.

models from these local clients, and redistribute an updated global model. The clients, on the other hand, are responsible for inheriting the global model's generic knowledge and conducting local optimization to properly integrate their specialised local knowledge.

The learning of FedUCC consists of three stages $s = \text{I, II, III}$ following a coarse-to-fine manner. Each training stage contains T_s global round, including E_s parallel client-wise (local) epochs and a global aggregation/redistribution. Particularly, the three stages are: I) **Generic knowledge guided modelling** formulates a model by adopting all available knowledge from the clients. II) **Specialized knowledge guided modelling** decouples the client-specific knowledge from the generic knowledge via parameter localization, and achieves personalized client models; and III) **Patch knowledge guided modelling** strengths the fine-grained patterns via local feature alignment. The details of the proposed method are discussed in the rest of this section.

3.2.2 Stage I: Generic Knowledge Guided Modelling

At this stage, the goal is to obtain a global model that can roughly discriminate images across all the clients, of which statistical heterogeneity is not explicitly explored. It helps obtain a robust initial weight for the model to conduct further training in the next two stages.

Given K clients, where the k-th client has a Person ReID dataset $X^k = \left\{x_1^k, x_2^k, \ldots, x_{n^k}^k\right\}$ with n^k images, our objective is to obtain model $\theta(W_g, W_l)$ on $\mathcal{X} = \{X^1, ..., X^k\}$. Note that the model $\theta(W_g, W_l)$ is parameterized by (W_g, W_l) using shared generic parameters W_g and client-specific parameters W_l. Particularly, the specialized parameters W_l are the neuron weights of batch normalization layers and the generic parameters are the remaining model parameters. We define (W_g^k, W_l^k) to be the trained generic parameters and specialized parameters on the k-th client, respectively. At the current stage we focus on learning robust generic knowledge shared

among clients, so both (W_g, W_l) are updated to the central server for redistribution. In the following part, we introduce the training steps taken by clients and the server respectively.

Client Design

Person ReID is a metric learning problem, thus at inference, the representations of image pairs capturing the same person are supposed to be as close as possible. Therefore, a clustering algorithm is adopted, namely DBSCAN, where images with distance below a predefined threshold are grouped into the same cluster.

Denote $\mathcal{H}^k = \{H^{k,1}, H^{k,2}, \ldots, H^{k,m^k}\}$ as the clusters obtained from the k-th client with dataset X^k, where m^k denotes the number of clusters in X^k. Note that each cluster $H^{k,i}$ is expected to contain all image features of a particular person. We first obtain a set of cluster centroids $\mathcal{C}^k = \{c_1^k, \ldots, c_{m^k}^k\}$ from X^k. Each cluster centroid is obtained by the mean feature vector of all images with the same pseudo-label:

$$c_j^k = \frac{1}{|H^{k,j}|} \sum_{u_i^k \in H^{k,j}} u_i^k, \tag{3.1}$$

where u_i^k denotes the feature representation of image x_i^k, obtained from the model $\theta(W_g, W_l)$, and $\left|H^{k,j}\right|$ indicates the number of images in the cluster.

Cluster centroid contrastive learning focuses on cluster-level feature alignment. For each client k, the objective is to minimise the distance of each image feature to their cluster centroid feature via an InfoNCE loss [58]:

$$\mathcal{L}_{\mathrm{u}} = \mathbb{E}\left[-\log \frac{\exp\left(u_i^k \cdot c_j^k / \tau\right)}{\sum_{j'=1}^{m^k} \exp\left(u_i^k \cdot c_{j'}^k / \tau\right)}\right], \tag{3.2}$$

where the distances between the query image feature u_i^k and its cluster centroid c_j^k are minimised and out-cluster centroid distances are maximised, optimizing with a similar principle as a softmax entropy. And τ is a temperature hyper-parameter.

During optimization, the centroid feature vectors c_j^k of query instance u_i^k are momentously updated in the memory dictionary as follows:

$$c_j^k \leftarrow \lambda c_j^k + (1-\lambda)u_i^k, \tag{3.3}$$

where $\lambda \in [0,1]$ is a hyper-parameter controlling the extent to update the cluster centroid. At the end of local client training, all parameters (W^g, W^l) are forwarded to the server for multi-model aggregation.

Server Design

The responsibility of the central server is to coordinate client knowledge sharing via an aggregation step. Basically, a strategy similar to FedAvg [45] is adopted in this study, which conducts the aggregation at the $(t+1)$-th global round as follows:

$$(W_g, W_l)^{t+1} = \sum_{k=1}^{K} \frac{n^k}{n} (W_g^k, W_l^k)^t, \tag{3.4}$$

where $n = \sum_{k=1}^{K} n^k$ are the total number of images of all clients. Instead of using equivalent average, the weighted sum is determined by the size of each client dataset. As a result, it allows the larger dataset clients to contribute richer knowledge for model aggregation, which is expected to contain more robust Person ReID patterns.

However, the issue of statistical heterogeneity is yet to be addressed. Existing methods attempted to alleviate this by combining the latest generic parameters with old specialized parameters to obtain a client-specific model [65]. However, this doesn't fully resolve the distribution difference between clients as the model performance degrades after global aggregation/redistribution is still severe. Therefore, we further introduce the next modelling stage to solve this issue.

3.2.3 Stage II: Specialized Knowledge Guided Modelling

The first stage has produced a generic model by taking into account the knowledge of all clients. This process aggregates all client-specific domain knowledge into the global model. Due to the variety of different client data distributions, there are inevitably parameter collisions between the local models. The knowledge aggregation effect could be degraded, hence causing decreased performance on particular clients. We propose a knowledge guided modelling stage to further improve model accuracy by considering statistical heterogeneity. It focuses on retrieving client-specific patterns to conduct knowledge personalization, which aims to maximize the model performance on all clients.

This stage is based on the local batch normalization (BN). Specifically, batch normalization layers are added in the backbone networks to normalize the activation. They adopt statistics in a mini-batch, which inherently captures the information related to the client data distribution [36]. To achieve knowledge personalization, we preserve the learned local BN parameters by decoupling them from the global aggregation. This not only avoids parameter collision of the BN weights between clients but also allows the updated global model to be used in conjunction with the local BN parameters to achieve personalization. At this stage, the model aggregation and updates are as follows:

$$(W_g)^{t+1} = \sum_{k=1}^{K} \frac{n^k}{n} (W_g^k)^t. \tag{3.5}$$

And model localization for the k-th client is as follows:

$$\theta^{k,t+1} = \theta((W_g)^{t+1}, (W_l^k)^t). \tag{3.6}$$

Note that BN layers for personalization could lead to sub-optimal performance due to the over-fitting issue on a client. We further enhance image feature learning by

introducing patch-level representation to enable fine-grained aspects for facilitating parameter alignment across clients.

3.2.4 Stage III: Patch Knowledge Guided Modelling

Patch-based representations show their advantages for the robustness of unseen identities that are not involved in the training process [69]. Hence, additional model parameters which are used to characterize patch-level features are introduced and optimized by following a similar strategy as in Stage II.

Specifically, given the dataset X^k on the k-th client, we obtain the set of 3D tensor features of images before the global average pooling (GAP) layer in the backbone model as $V^k = \{v_1^k, \ldots, v_{n^k}^k\}$. We then extract the set of patch features for all images as $\mathcal{Q}^k = \{Q_1^k, \ldots, Q_{n^k}^k\}$ by equally splitting v_i^k into p horizontal strips and further adopt GAP to obtain patch-level image features. In detail, we represent them as $Q_i^k = \{q_{i,1}^k, \ldots, q_{i,p}^k\}$.

Finally, for each cluster $H^{k,j} \in \mathcal{H}^k$, we further extend it to p clusters for features within particular patches, of which the centroids are: $D_j^k = \{d_{j,1}^k, \ldots, d_{j,p}^k\}$ associated with the corresponding patch-level features.

After that, given a patch feature representation $q_{i,r}^k$ and a centroid $d_{j,s}^k$, a loss function can be defined as follows:

$$\mathcal{L}_{\mathrm{p}} = \mathbb{E} \sum_{r=1}^{p} \left[-\log \frac{\exp\left(q_{i,r}^k \cdot d_{j,s}^k / \tau\right)}{\sum_{j'=1}^{m^K} \exp\left(q_{i,r}^k \cdot d_{j',s}^k / \tau\right)} \right], \tag{3.7}$$

which is similar to Eq.(3.2).

To this end, a loss function jointly optimized L_u and L_p can be adopted to achieve a robust federated Person ReID:

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_p. \tag{3.8}$$

Table 3.1: Statistics of the eight person ReID datasets.

		Train		Test	
Datasets	# Cam	# ID	# Img	# Query	# Gallery
Duke	8	702	16,522	2,228	17,611
Market1501	6	751	12,936	3,368	19,732
CUHK03	2	767	7,365	1,400	5,332
PRID	2	285	3,744	100	649
CUHK01	2	485	1,940	972	972
VIPeR	2	316	632	316	316
3DPeS	2	93	450	246	316
iLIDS	2	59	248	98	130

Table 3.2: Overall performance with ablation studies on eight Person ReID datasets.

Methods	Duke		Market		CUHK03		PRID	
Metric (%)	R1	mAP	R1	mAP	R1	mAP	R1	mAP
FedUReID [81]	51.0	-	65.2	-	8.9	-	38.0	-
Backbone	14.4	6.6	42.2	21.3	1.4	2.0	19.9	24.4
Stage I	65.7	44.4	74.6	45.6	**11.3**	9.4	24.0	27.2
Stage I + II	75.2	55.6	85.7	63.5	8.9	9.0	52.9	57.2
Stage I + III	76.3	56.3	82.6	58.2	9.6	8.8	46.9	52.1
FedUCC	**78.8**	**60.5**	**86.5**	**65.5**	9.6	**9.7**	**58.9**	**63.1**

Methods	CUHK01		VIPeR		3Dpes		iLIDS		Avg	
Metric (%)	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
FedUReID [81]	43.6	-	26.6	-	65.3	-	73.5	-	46.5	-
Backbone	26.7	26.1	22.7	28.2	61.2	43.4	63.9	48.5	31.5	25.1
Stage I	60.6	57.3	**35.4**	**39.7**	60.3	41.7	**81.0**	**66.2**	51.6	41.4
Stage I + II	69.5	65.7	31.0	36.1	65.5	46.4	80.1	60.7	58.6	49.3
Stage I + III	74.5	71.4	28.1	33.1	65.5	46.6	67.5	53.6	56.4	47.5
FedUCC	**78.3**	**75.3**	31.3	36.7	**68.9**	**50.9**	74.7	59.7	**60.9**	**52.7**

3.3 Experiments

3.3.1 Dataset

Eight Person Re-ID datasets were adopted to comprehensively evaluate our proposed method,as detailed in 3.1, including DukeMTMC-ReID [50], Market1501 [76], CUHK03-NP [35], PRID [26], CUHK01 [34], VIPeR [21], 3DPeS [2] and iLIDS [62]. Based on the federated learning setting, each dataset is considered to be held by a single client. Therefore, each client's data is collected at a unique location. This is consistent to real-world scenarios when delivering a Person Re-ID model with the

consideration of the privacy requirements.

Standard Person ReID evaluation metrics were adopted to measure the performance, including Cumulative Match Characteristic (CMC) curve and mean Average Precision (mAP) [75]. Given a query image and the gallery containing a group of watch list images, CMC first ranks the gallery images in line with their similarity to the query. Then, each image ranked within the rank-K (i.e., top-K) results is verified whether a matching identity is within the results or not. In addition, rank-1 accuracy is introduced to indicate the average probability of the most similar image matching the query image, and mAP reflects the mean average precision of all queries.

Table 3.3: Comparisons to the state-of-the-art methods.

Methods	Types	Duke				Market			
Metric (%)		R1	R5	R10	mAP	R1	R5	R10	mAP
PUL [14]	Domain Adaptation	30.4	46.4	50.7	16.4	44.7	59.1	65.6	20.1
HHL [78]	Domain Adaptation	46.9	61.0	66.7	27.2	62.2	78.8	84.0	31.4
SPGAN [9]	Domain Adaptation	46.9	62.6	68.5	26.4	58.1	76.0	82.7	26.7
MMT [18]	Domain Adaptation	**83.8**	**91.2**	**93.4**	**70.4**	**90.5**	**96.6**	**97.8**	**78.2**
BUC [41]	Purely Unsupervised	40.4	52.5	58.2	22.1	61.9	73.5	78.2	29.6
MMCL [60]	Purely Unsupervised	65.2	75.9	80.0	40.2	80.3	89.4	92.3	45.5
MMT [18]	Purely Unsupervised	82.4	90.8	93.0	67.1	88.1	96.0	97.5	74.3
FedUReID [81]	Federated Unsupervised	51.0	62.4	67.6	29.5	65.2	77.8	82.2	34.2
FedUCC (Ours)	Federated Unsupervised	78.8	87.2	89.8	60.5	86.5	94.5	96.7	65.5

3.3.2 Experimental Settings

Backbone Architectures. ResNet-50 [22] pretrained on ImagNet [8] was adopted as the backbone network for feature extraction.

Implementation Details. For each client, an Adam optimizer with Nesterov momentum 0.9 was adopted. The learning rate was set to 3.5×10^{-4} for local model optimization, and the batch size was set to 64 empirically. The number of patch-feature segmentation p was set to 2, indicating the upper-half and lower-half of the

human body. We employed DBSCAN [12] for image feature clustering based on Jaccard distance with k-reciprocal encoding [79] and set the number of minimum samples as 2 to discard un-clustered images.We empirically set the temperature hyper-parameter τ as 0.05. The implementation was with an NVIDIA V100 for the training and inference stages.

Training Settings. The model parameters were optimized through 3 training stages as discussed in Sec. 3.2. In particular, during the initial phase, Stage I is carried out independently for $T_I = 100$ global rounds to establish the initial model, wherein the local epoch E_I is set to 1 in order to promptly aggregate and update the model parameters from the local clients. Following this, Stage II commences for a total of $T_{II} = 100$ global rounds, with a local epoch of $E_{II} = 5$ to acquire sufficient local knowledge. Note that Stage II and III/IV are compatible. The model optimized through Stage II training continues to be refined by combined Stage II and Stage III training with global round of $T_{III} = 100$ and local epoch of $E_{III} = 5$. In this joint optimization, loss functions from Stage II and III are merged with equal significance. The ultimate FedUCC retrieval outcomes are based on the fusion of Stage II holistic features and Stage III patch features. More precisely, the similarity of image features between query and gallery, rooted in both the holistic view of Stage II and the part-based view of Stage III, is concatenated through addition. This combined similarity is then used for accuracy assessment.

3.3.3 Comparisons with the State-of-the-Art Unsupervised ReID

We compared the proposed method with a number of existing methods following their evaluation protocols to demonstrate the effectiveness of our method. In Table 3.2, a recently proposed unsupervised federated Person Re-ID method FedUReID [65] was compared to ours on 8 datasets. Our proposed method performed consistently better than FedUReID on all datasets. In particular, our result outper-

Figure 3.3: Ranking result on three queries

forms FedUReID with a large margin on the Duke, Market, and CUHK01 datasets. Note that Duke and Market are the two largest datasets. Such result indicates significant incremental in terms of the top-1 accuracy from 51.0% to 78.8% on Duke and 65.2% to 86.8% on Market. Other datasets also demonstrate the effectiveness of FedUCC. CUHK01 shows the largest improvement regarding the top-1 accuracy 43.6% to 78.3%, PRID shows 20.9% incremental, and the remaining CUHK03, VIPeR, 3DPeS, iLIDS also show performance improvement of 0.7%, 4.7%, 3.6% and 1.2%, respectively.

In Table 3.3, FedUCC is further compared with two additional categories of methods including the unsupervised domain adaption (PUL [14], HHL [78], SPGAN [9] and MMT [18]) and the fully unsupervised approaches (BUC [41] , MMCL [60] and MMT [18]). The evaluation was conducted on the two largest datasets: Duke and Market. Note that the unsupervised domain adaptation requires the source dataset to have annotated labels. Without the additional supervision using annotations, our method still performs better than most domain adaptation methods. For example, our method and SPGAN [9] achieve 78.8% and 46.9% rank-1 accuracy on Duke, respectively; our method and HHL [78] achieve 86.8% and 62.2% rank-1 accuracy on Market. On the contrary, MMT has exhibited superior performance compared to our method, achieving rank-1 accuracy of 83.8% and 78.8% on Duke, 90.5% and 86.5% on Market datasets. This showcases that, while multi-domain feature learning through federated learning offers advantages, it might not necessarily surpass the Domain Adaptation strategy by a substantial margin. Additionally, we compare our results with fully unsupervised methods, where each ReID model is trained independently on individual datasets without collaboration. Superior performance was also demonstrated over some of previous works, given the comparison of our result to MMCL [60] with 78.8% and 65.2% rank-1 accuracy on Duke, 86.6% and 80.3% on Market. This also highlights the advantage of collaboration training using federated learning. Nevertheless, even within a purely unsupervised context, MMT

has exhibited a slight performance edge over our approach, attaining rank-1 accuracies of 82.4% and 78.8% on Duke, and 88.1% and 86.5% on Market datasets. While one might intuitively anticipate that federated unsupervised learning benefits from a more diverse data distribution across clients, this assumption doesn't always hold true.This is evidenced by the performance dip observed after the weight aggregation stage, which implies that federated learning still constrains clients from fully adapting the model to their specific domains. Nevertheless, this insight hints at a potential avenue for future exploration—further domain adaptation on the larger-sized clients to attain maximal performance. However, this extended setting would inherently belong to the domain adaptation approach, exceeding the scope of the current study.

3.3.4 Ablation Studies

Ablation studies were conducted to demonstrate the effectiveness of the individual strategies proposed, which helped explore the role of each stage in the entire pipeline of FedUCC. Particularly, Table 3.2 lists the comparisons of several architectures in terms of Rank-1 accuracy and mAP. These architectures include a backbone only method, a Stage I only method, Stage I + II method and Stage I + III method. Overall, our proposed FedUCC achieves the best overall performance with all these strategies. We discuss the details of these individual architectures in the following sections.

Backbone Method: Generic Knowledge Guided Modelling. The backbone setting adopted conventional DBSCAN-based label assignment and conducted optimization using a cross-entropy loss for classification. For federated learning, we chose a local epoch of 5 that balances the model convergence and the cross-client knowledge exchange. Each client model was updated with a fixed number of training batches in line with the dataset size. FedAvg considered this by limiting smaller client models with fewer averaging ratios for aggregation. However, skewness can

usually be identified among the clients and the local model parameters were likely to conflict. As a result, the local training led to significant fluctuation of the global model performance between the global rounds and eventually limited the person ReID accuracy.

Stage I: Generic Knowledge Guided Modelling with Memory-based Contrastive Learning. Compared to the backbone method, Stage I further introduces the memory-based contrastive learning on each client as indicated in Eq. (3.1) - (3.3). As shown in TABLE 3.2, the performance was superior on all clients. Particularly, the Rank-1 accuracy increased on the largest dataset Duke and Market 51.3% and 32.4% and achieved an average increase of 20.1% on all datasets. Such performance improvement shows the reduced influence of statistical heterogeneity. Due to the more effective model training strategy and the unresolved issue of client knowledge conflict, the model performance fluctuation on all clients is still severe as illustrated in Fig. 3.4.

Stage I + II: Specialized Knowledge Guided Modelling with Batch Normalization. Stage II helps explore personalized knowledge, which retains the batch normalization patterns within each client. Specifically, Stage II has shown an average of 7.9% increase in terms of Rank-1 accuracy. Additionally, the model performance is much more stable compared with Stage I, as shown in Fig. 3.4. This indicates the problem of statistical heterogeneity is greatly reduced with effective personal knowledge decoupling. Note that the improvements were not consistent across all the clients. For smaller datasets such as CUHK03, VIPeR, and iLIDS, the performance even dropped slightly. It suggested a potential domain over-fitting issue, such that the generic knowledge from the shared parameters was overwhelmed by the local knowledge. Therefore, this motivated us to further explore patch-based feature learning to enhance the robustness of the feature spaces.

Stage I+III: Patch Knowledge Guided Modelling. The patch-based features were explored for fine-grained person representation. As shown in TABLE 3.2,

the average Rank-1 accuracy has decreased by 1.8% on average compared with using the holistic person feature. However by ensemble of both holistic and patch features (FedUCC), better discriminability was achieved compared with using standalone representation, with an average increase of rank-1 accuracy by 3.4% Specifically, a consistent performance gain can be observed on the majority of those datasets (i.e., CUHK03, PRID, CUHK01, VIPeR, 3DPeS, iLIDS) compared with the strategy of Stage I + II. In our observations, iLIDS shows the highest performance in Stage I training, and no performance gain has been shown for later stages of training. Given that it is the smallest dataset, the number of images is likely insufficient for the Stage I model to converge better.

3.3.5 Qualitative Study

Fig. 3.3 visualizes the top-10 results of three queries to demonstrate the effectiveness of our three-stage modelling method. It can be seen that the backbone-only method retrieves the most incorrect images. Even though the general appearances are similar, attention to more detailed patterns is lacking. Visual clues such as the backpack are rarely matched in the corresponding query results. Stage I shows a much better retrieval accuracy in Fig. 3.3, where more discriminative patterns are identified to mine the matching query images. Stage II training further improves retrieval accuracy with an enhanced ability to mine discriminative patterns despite viewpoint variations. Better retrieval accuracy is shown with more front and side views of the person despite the larger visual variation introduced by the backpack. Stage III model shows a comparable result to Stage II, with more focus on regional pattern matching. This is indicated by the 10-th negative retrieval image where the camo patterns in the query's backpack and short are closely matched. Lastly, the FedUCC (Stage I + II + III) model ensembles holistic and patch-level features to obtain a refined feature representation, improving query results with both holistic and fine-grained patterns, as illustrated in Fig. 3.3.

3.4 Summary

In this chapter, we present a federated unsupervised cluster-contrastive learning method, namely FedUCC, to address the Person ReID problem. FedUCC introduces a three-stage modelling strategy following a coarse-to-fine manner. In detail, generic knowledge, specialized knowledge and patch knowledge are explored with a deep neural network. It enables the sharing of mutual knowledge while retaining local domain-specified knowledge through network layers and their parameters. Comprehensive experiments on 8 public benchmark datasets consistently demonstrated the performance improvement of our proposed FedUCC, compared with existing methods.

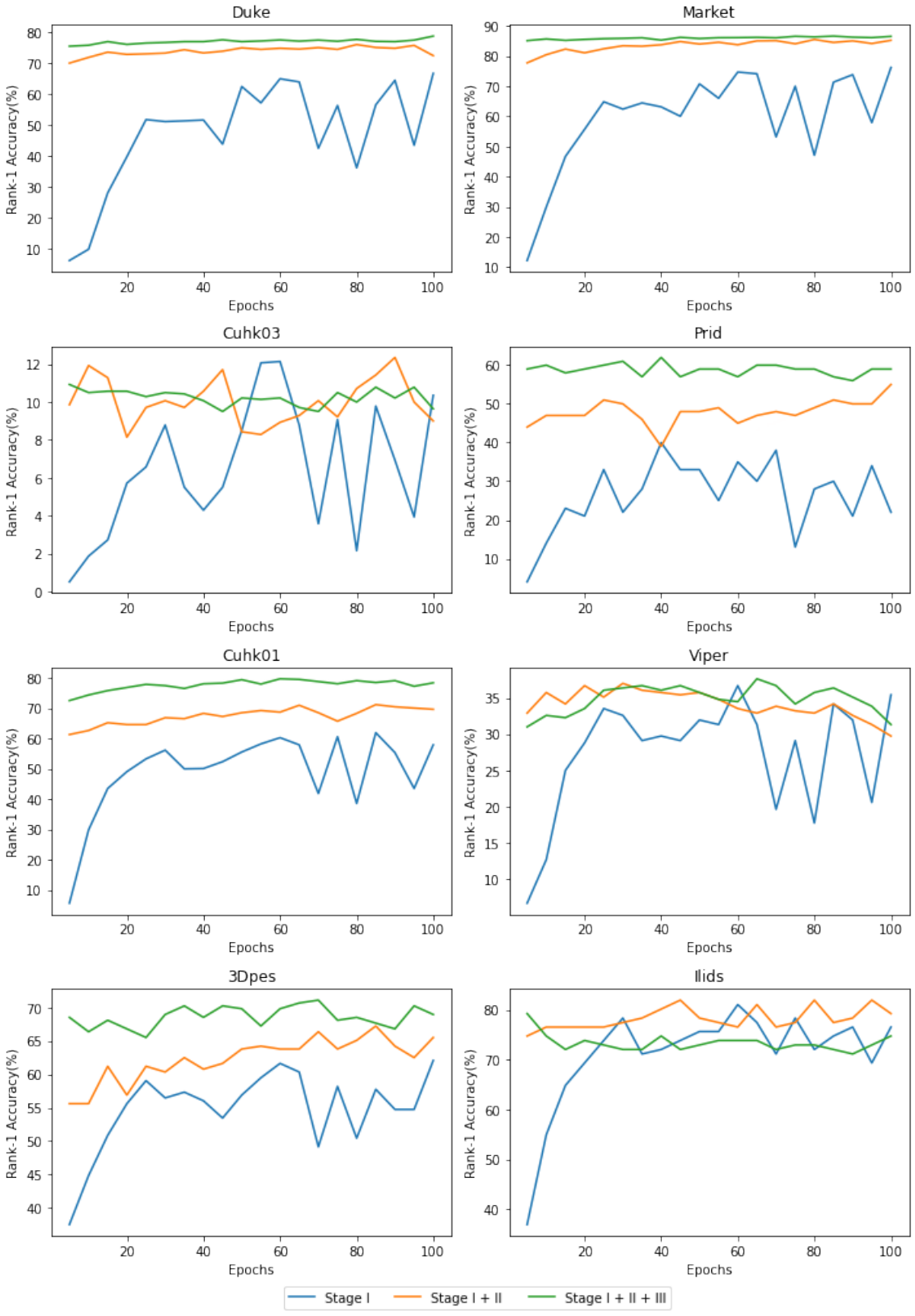


Figure 3.4: Rank-1 accuracy curves on all datasets

Chapter 4

CCIT: Context and Camera Invariant Transformer for Federated Unsupervised Person ReID

4.1 Introduction

Given the escalating concerns surrounding data privacy, federated learning has emerged as an increasingly appealing approach in the context of Person ReID. It enables the acquisition of a generalized central machine learning model through collaborative feature learning across multiple clients, all without the need for direct exchange of raw data. Federated learning shares abstract knowledge among the clients based on the weights of local machine learning models (e.g. neural networks). General federated learning schemes have been adopted for supervised person ReID [82, 81, 65], whilst the statistical heterogeneity between data domains of person ReID remains a challenge. Such domain-specific knowledge causes domain conflicts when sharing and aggregating the model weights, which results in

sub-optimal performance. Thus, several studies are in pursuit of consistent local modelling strategies among clients, such as synthesizing local images with cross-domain styles [43] and adversarially regularising the image representations to be domain-invariant [71]. Another approach introduces personalized models in addition to a global model to adapt it in accordance to the data distributions of local clients. For example, meta-learning is adopted in the local adaptation stage [13] and client-specific knowledge is encoded in auxiliary model weights [55]. However, federated unsupervised person ReID has been seldom investigated, until recently [81] explored conventional federated settings for this task.

In addition, these existing federated person ReID methods are all based on convolution neural networks (CNNs). They tend to rely more on local high-frequency patterns, which leads to domain-specific modeling and negatively impacts cross-domain global learning [48]. This is particularly the case for federated person ReID, where the input images captured in different datasets may vary significantly regarding their local patterns due to geographical variations and different camera settings, leading to serious non-IID (non-independent and identical distribution) problems for federated learning. Recently, vision transformers have been shown to learn image representations that are robust to domain variations [48], and therefore, various strategies have been devised to better utilise image tokens, such as token aggregation [39] and irrelevant tokens removal [70]. This motivates us to devise a transformer-based method in pursuit of invariant image representations across client domains for federated unsupervised person ReID.

In this chapter, a novel transformer architecture, namely context and camera invariant transformer (C^2IT), is proposed for federated unsupervised person ReID to deal with the non-IID problem. Overall, unsupervised contrastive learning is conducted to formulate image representations, during which two token-based mechanisms are introduced with the transformer to guide the representations to be discriminative by focusing on person identity patterns. First, a context-invariant learning

strategy regarding image contents is devised by associating semantic-similar tokens into a number of clusters. It distinguishes the less informative cluster tokens (e.g., background objects) from the more informative ones (e.g., body regional patterns). This enables disentanglement learning for the domain-specific irrelevant patterns to obtain high-quality image-level representations. Second, a camera-invariant learning strategy is further devised to reduce the intra-domain camera bias. A set of camera tokens are introduced to characterise the cameras within a client. The image-level representations are optimised to be of consistent similarities regarding all the cameras of the client based on a uniformity loss. Comprehensive experiments on eight public person ReID datasets demonstrate the effectiveness of the proposed method with the state-of-the-art performance.

In summary, the contributions of this chapter are four-fold:

- A novel Context and Camera Invariant Transformer architecture is proposed for the first time for federated unsupervised person ReID to alleviate the varied data distribution issue for non-IID scenarios.
- A context-invariant learning scheme is devised in pursuit of disentangling the irrelevant contents from image-level person ReID related representations.
- A camera-invariant learning scheme is devised to alleviate the impacts of various factors such as photography parameters and to obtain consistent image representations.
- Comprehensive experiments and analyses are conducted on eight public datasets to demonstrate the effectiveness and the state-of-the-art performance of our method.

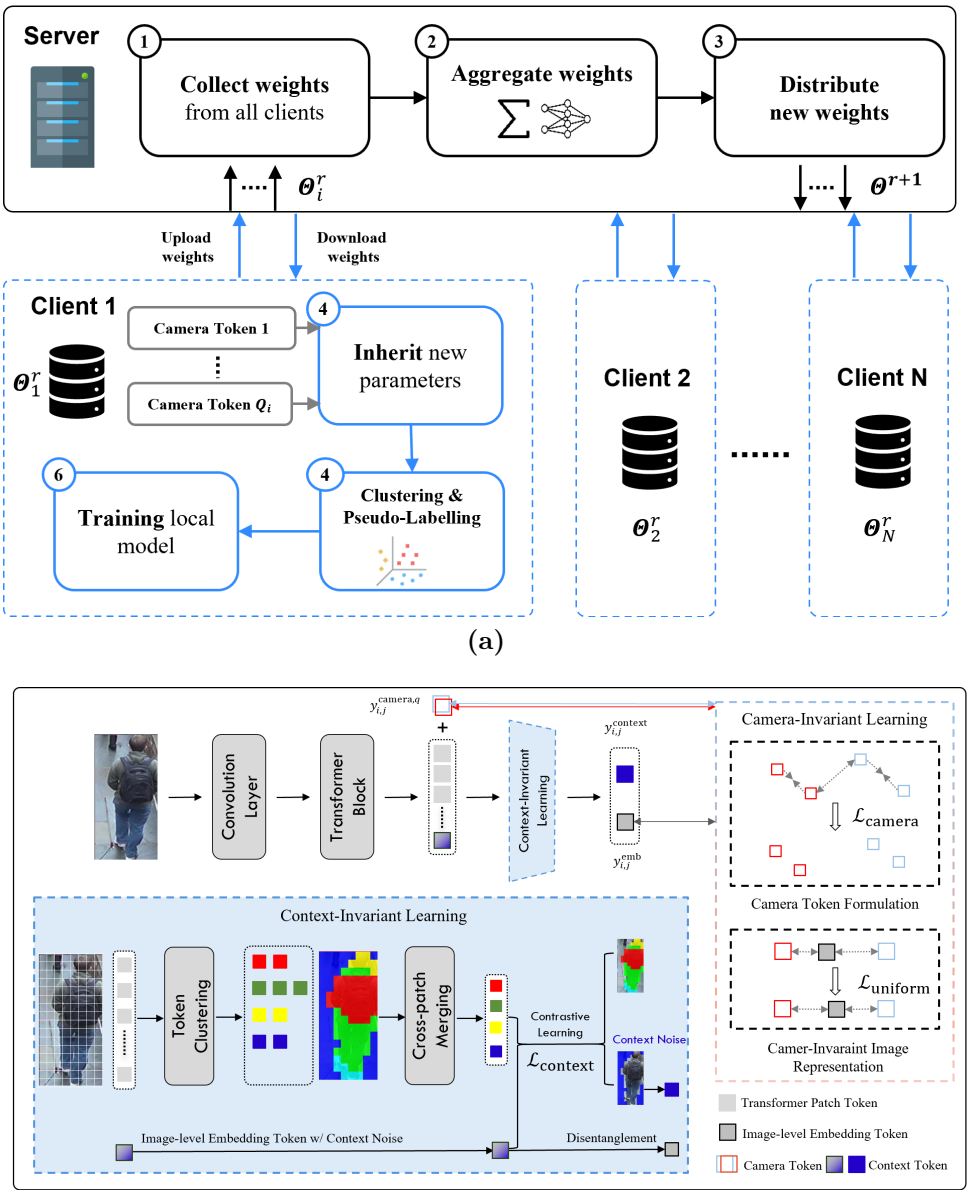


Figure 4.1: Illustration of the proposed C^{2}IT architecture for federated unsupervised person ReID. (a) Federated setting between the central server and the clients. (b) The proposed context-invariant and camera-invariant learning strategies with a vision transformer.

4.2 Methodology

4.3 Federated Unsupervised Person ReID

In our federated person re-identification task, N clients are involved, which are denoted as $C_1, \ldots, C_N$. Each client C_i has a local dataset $D_i = \{x_{i,j}\}$, where $x_{i,j}$, $j = 1, ..., |D_i|$, are the images. The objective is to obtain a deep learning model $y_{i,j} = f(x_{i,j}|\mathbf{\Theta})$, where $\mathbf{\Theta}$ contains learnable weights and $x_{i,j}$ is an input image to be represented with $y_{i,j}$ in a latent space. The latent representations can be used to identify a person in different images based on their similarity. This model is expected to work for all datasets $\mathcal{D} = \bigcup_i \mathcal{D}_i$, whilst the training procedure should not reveal the raw data on each client to others. Thus, a server-client paradigm is adopted to aggregate and exchange knowledge across clients in a federated learning setting. The exchange is set to be conducted in R rounds in this study.

Learning on Server

The responsibility of the central server is to coordinate the knowledge aggregation process. We follow the standard FedAvg [45] pipeline as illustrated in Figure 1 (a), which distributes the global model towards all the clients, collects optimized client models to the server side and aggregates the weights of client models to obtain a new global model. The total number of rounds R is based on the global model convergence regarding all the client datasets or predefined criteria. Particularly, for the r-th round, $r = 1, .., R$, the knowledge is the weights from the local models of all the clients: $f_i^r(x|\mathbf{\Theta}_i^r)$, $i = 1, ..., N$, where f_i^r is a local model updated with the dataset on the i-th client during the r-th round and $\mathbf{\Theta}_i^r$ contains the corresponding model weights. Mathematically, in accordance with the data volume per client, we have:

$$\mathbf{\Theta}^{r+1} = \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \mathbf{\Theta}_i^r, \tag{4.1}$$

where $\mathbf{\Theta}^{r+1}$ is the updated global model weight and to be distributed to local clients in the $(r+1)$-th round. Note that, due to the forbidden access to the client datasets, this server-side update can be sub-optimal when simply using the local model weights.

Learning on Clients

For the local model optimization on the client side, once the i-th client obtains the global model from the server, a local training on $f_i^r(x|\mathbf{\Theta}_i^r)$ is conducted by following the pipeline of a standard unsupervised learning procedure, i.e. FedAVG [45]. Initially, for images $x_{i,j}$, the pseudo labels are generated based on their representations $y_{i,j}$, which are clustered into K groups. The ones within the same cluster have the same pseudo-label. Next, a memory-based cluster contrastive learning algorithm is adopted to optimize the feature representations and their clusters iteratively [7].

First, cluster centroid representations are formulated as the average of the image features $y_{i,j}$ within each cluster, and a memory bank is formulated by storing the centroid representations of all the clusters: $c_{i,1}, ..., c_{i,k}, ..., c_{i.K}$. Particularly, we denote $k_{i,j}$ to represent the clustered index associated with $y_{i,j}$. Second, a cluster contrastive learning aligns image features $y_{i,j}$ to the corresponding cluster feature. In detail, given the image feature $y_{i,j}$, a cluster-contrastive loss optimises $f_i^r(x_{i,j}|\mathbf{\Theta}_i^r)$ to minimise the distance between $y_{i,j}$ and its cluster centroid, whilst maximising its distance to the rest cluster centroids:

$$\arg\min_{\mathbf{\Theta}_i^r} \mathcal{L}_{\text{unsupervised}} = -\log \frac{\exp\left(y_{i,j} \cdot c_{i,k_{i,j}}/\tau\right)}{\sum_k \exp\left(y_{i,j} \cdot c_{i,k}/\tau\right)}, \tag{4.2}$$

where τ is a temperature parameter.

Finally, based on the optimised image features, the corresponding cluster feature vectors $c_k(i,j)$ in the memory bank are updated as:

$$c_{i,k_{i,j}} \leftarrow \lambda c_{i,k_{i,j}} + (1-\lambda) y_{i,j}, \tag{4.3}$$

where $\lambda \in [0, 1]$ is a hyper-parameter controlling the update pace of the cluster feature.

Note that the variables such as $y_{i,j}$ are associated with federated rounds and iteration steps. For convenience, we omit these indications in the symbol definitions and similar practices are with the following discussions.

4.4 Context and Camera Invariant Transformer

Vision transformers [10] are adopted for Person ReID in this study, which involves patch-based token formulation for visual patterns. Specifically, an image $x_{i,j}$ is segmented into a set of P fixed-size patches. These patches are then forwarded through a linear projection mapping and positional encoding to obtain the initial patch tokens, which are denoted as $t_{i,j}^1, ..., t_{i,j}^p, ..., t_{i,j}^P$. In addition, a learnable image-level initial embedding token $t_{i,j}^{\text{emb}}$ is attached to the beginning of the patch-based image token sequence. The output corresponding to this token is used as the image-level embedded representation $y_{i,j}^{\text{emb}}$. In summary, the input of the transformer can be denoted as $T_{i,j} = \{t_{i,j}^{\text{emb}}, t_{i,j}^1, \ldots, t_{i,j}^p, \ldots, t_{i,j}^P\}$; and the output tokens from the transformer are denoted as $Y_{i,j} = \{y_{i,j}^{\text{emb}}, y_{i,j}^1, \ldots, y_{i,j}^p, \ldots, y_{i,j}^P\}$.

4.4.1 Context-Invariant Learning

Contextual information refers to various image contents that are not closely relevant to the Person ReID task, such as background objects, which could negatively impact the person ReID performance. Therefore, we propose a context disentangling strategy to remove such information with a context token in the transformer, whilst the remaining tokens are expected to focus on the discriminative hints for person ReID. The learning module consists of three steps: token clustering & cross-patch merging, context token identification and context noise disentanglement.

Token Clustering & Cross-Patch Merging

Token clustering is to associate similar tokens that contain similar semantic regions (e.g, background objects, body regions) and cross-patch merging is to collectively aggregate the tokens within the same region.

The token clusters are obtained with the commonly used Density Peaks Clustering method which is based on the $\mathcal{K}$ nearest neighbours (DPC-KNN) [11]. On the i-th client, all the patch-based tokens $t_{i,j}^p$ of the j-th image are first evaluated regarding the suitability to be the KNN centroids under two metrics: local density and distance indicator. The local density value $\rho_{i,j}$ is obtained according to its $\mathcal{K}$-nearest neighbours:

$$\rho_{i,j} = \exp\left(-\frac{1}{\mathcal{K}} \sum_{y_{i,j}^{p'} \in \mathrm{KNN}\left(y_{i,j}^{p}\right)} \left\| y_{i,j}^{p'} - y_{i,j}^{p} \right\|_2^2\right), \tag{4.4}$$

where $\mathrm{KNN}(y_{i,j}^p)$ denotes the $\mathcal{K}$-nearest neighbours of token $y_{i,j}$ regarding the distance between vectors.

Then, for a token $t_{i,j}^p$, a distance score $\delta_{i,j}$ is computed as the minimal distance to a token with higher local density. It indicates the isolation degree of a token, where a higher score suggests that it can form a cluster. Note that the score of the token with the highest density is the maximal distance between the token and any other tokens. In detail, we have:

$$\delta_{i,j} = \begin{cases} \min_{j':\rho^{j'} > \rho^{j}} \left\| y_{i,j} - y_{i,b} \right\|_2, \text{if } \exists\, b \text{ s.t. } \rho_{i,b} > \rho_{i,j}, \\ \max_{j'} \left\| y_{i,j} - y_{i,j'} \right\|_2, \text{ otherwise }. \end{cases} \tag{4.5}$$

The local density and the distance indicators are merged to evaluate the overall score of each token, via multiplication as $\rho_{i,j}\delta_{i,j}$. A higher final score indicates a higher potential as a cluster centroid.

Finally, V tokens are selected as cluster centroids with the top-V' highest scores and the rest tokens are assigned to their nearest cluster centroids. For the v-th

cluster, a cross-patch token $\bar{y}^v_{i,j}$ can be merged with the average of $y^p_{i,j}$ within the cluster.

Context Token Identification

The cross-patch token with the least person ReID related context is treated as the context token. It is identified by measuring the distance between the cross-patch tokens and the embedding token. Specifically, we have:

$$v^*_{i,j} = \arg\max_v \left\| y^{\text{emb}}_{i,j} - \bar{y}^v_{i,j} \right\|_2 , \tag{4.6}$$

and $y^{\text{context}}_{i,j} = \bar{y}^{v^*_{i,j}}_{i,j}$ is denoted as the context token.

Context Noise Disentanglement

With the obtained context tokens $y^{\text{context}}_{i,j}$, a disentanglement scheme is introduced to minimise the context's impacts on person ReID. In detail, a triplet loss is formulated to draw closer to the image-level representations of the images from the same person and push away them with the context tokens. Specifically, each triplet is composed of three samples: $y^{\text{emb}}_{i,j}$, $y^{\text{emb}}_{i,+}$ and $y^{\text{context}}_{i,-}$, where $y^{\text{emb}}_{i,+}$ is a positive image token with the same pseudo-label as $y^{\text{emb}}_{i,j}$ and $y^{\text{context}}_{i,j'}$ is any context token within a dataset or a batch:

$$\begin{aligned} \mathcal{L}_{\text{context}} = [\max_{y^{\text{emb}}_{i,+}} \left\| y^{\text{emb}}_{i,j} - y^{\text{emb}}_{i,+} \right\|_2 - \\ \min_{y^{\text{context}}_{i,j'}} \left\| y^{\text{emb}}_{i,j} - y^{\text{context}}_{i,j'} \right\|_2 + \alpha]_+, \end{aligned} \tag{4.7}$$

where α is a predefined margin and $[\cdot]_+$ denotes the function $\max(\cdot, 0)$.

4.4.2 Camera-Invariant Learning

Another challenge for person ReID is to address the visual variations between cameras caused by various factors such as view angles and photography parameters

and conditions. Such variations exist in both intra-domain and inter-domain cameras. To address this challenge, a camera-invariant learning strategy is proposed to explicitly formulate camera representations and reduce their impacts on obtaining effective image-level representations.

Camera Tokens

Camera tokens formulate camera specific information. Initial camera tokens are introduced for an image $x_{i,j}$: $t_{i,j}^{\text{camera},1}, \cdots, t_{i,j}^{\text{camera},Q_i}$. Each $t_{i,j}^{\text{camear},q}$ is an initial representation regarding the j-th camera of the total Q_i cameras on the i-th client. Note that the camera capturing image $x_{i,j}$ is denoted as $q_{i,j}$. The transformer encodes them as $y_{i,j}^{\text{camera},1}$,...,$y_{i,j}^{\text{camera},Q_i}$, which are further optimised to align the image representations of the same camera towards a shared representation and keep the inter-camera representations apart from each other. Specifically, a camera contrastive loss is introduced:

$$\begin{aligned}\mathcal{L}_{\text{camera}} = [\min_{q_{i,j}=q_{i,j'}} \left\| y_{i,j}^{\text{camera},q_{i,j}} - y_{i,j'}^{\text{camera},q_{i,j'}} \right\|_2 - \\ \max_{q_{i,j}\neq q_{i,j''}} \left\| y_{i,j}^{\text{camera},q_{i,j}} - y_{i,j''}^{\text{camera},q_{i,j''}} \right\|_2 + \beta \,]_+,\end{aligned} \tag{4.8}$$

where the first term minimises the distance between the given image and other images captured by the same camera, the second term maximises the distance between the given image and the images captured by different cameras, and β is a predefined margin.

Camera-invariant Image Representation

The camera tokens are used to guide the image-level embedding tokens to be camera agnostic. This is achieved by enforcing the embedding token with consistent similarity towards all camera tokens, which helps alleviate the issue of camera bias. First, the dot product similarity is adopted to measure the embedding token with

all camera tokens:

$$s_{i,j}^{\text{camera},q} = y_{i,j}^{\text{emb}} \cdot y_{i,j}^{\text{camera},q}, q = 1, ..., Q_i, \tag{4.9}$$

where $s_{i,j}^{\text{camera},q}$ indicates the similarity between the image representation $y_{i,j}$ and the q-th camera in the i-th client.

Then, in pursuit of consistent similarity, we introduce the following optimisation:

$$\arg\min_{\mathbf{\Theta}_i^r} \mathcal{L}_{\text{uniform}} = -\sum_j \sum_q \log(s_{i,j}^{\text{camera},q}), \tag{4.10}$$

which uniformly maximises the similarity regarding to the similarity values with every camera.

4.5 Experiments and Discussions

4.5.1 Experimental Settings

Datasets. We evaluated our proposed method with eight commonly used public datasets as shown in Table 3.1, including DukeMTMC-ReID [50], Market1501 [76], CUHK03-NP [35], PRID [26], CUHK01 [34], VIPeR [21], 3DPeS [2] and iLIDS [62]. Each of them is treated as an individual dataset held by a client in the federated learning paradigm. Note that these datasets were collected at different geographical locations, so they are statistically heterogeneous as in the general assumption of federated learning. Two commonly used metrics are adopted to evaluate our proposed method: Rank-1 accuracy and mean average precision (mAP) [81].

Implementation details. We constructed the client models based on ViT-S, which was pre-trained on LUPerson [16]. We resized all person images to 256 × 128 and conducted data augmentations on training images with random horizontal flipping, padding, cropping and erasing. Each image was then split into 128 patches and projected into feature tokens in a 386-dimensional space. We set the batch size to 64 with 4 images per pseudo-label. An Adam optimizer was adopted with an initial

learning rate 0.0065: a consistent learning rate was used for domain adaptation pre-training and a weight decaying parameter 0.1 for every 40 federated training rounds was used for the context-invariant learning. We empirically set the temperature hyper-parameter τ as 0.05.

Table 4.1: Overall performance with ablation studies on eight person ReID datasets.

Methods	Duke		Market		CUHK03		PRID	
Metric (%)	R1	mAP	R1	mAP	R1	mAP	R1	mAP
FedUReID	51.0	-	65.2	-	8.9	-	38.0	-
Baseline (ViT-S Backbone)	82.0	69.6	94.3	87.0	68.3	63.9	83.9	86.2
Baseline + Context-IL	81.8	69.8	**94.4**	86.8	69.2	65.2	**86.9**	**88.9**
Baseline + Camera-IL	81.3	69.1	94.3	86.1	67.5	63.9	85.9	87.7
C^2IT (Ours)	**82.3**	**70.0**	94.2	**87.1**	**71.2**	**66.3**	**86.9**	88.8

Methods	CUHK01		VIPeR		3Dpes		iLIDS		Avg	
Metric (%)	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
FedUReID	43.6	-	26.6	-	65.3	-	73.5	-	46.5	-
Baseline (ViT-S Backbone)	**91.6**	89.5	56.3	60.8	85.3	77.6	82.8	75.8	80.6	76.3
Baseline + Context-IL	91.1	88.8	**56.6**	61.3	**86.2**	77.1	85.5	77.2	81.5	76.9
Baseline + Camera-IL	90.6	88.8	55.6	60.7	**86.2**	77.5	85.5	77.2	80.9	76.4
C^2IT (Ours)	91.5	**89.5**	56.3	**61.4**	85.7	**77.9**	**87.3**	**78.3**	**81.9**	**77.4**

4.5.2 Overall Performance

In Table 4.1, our proposed method is compared with the existing unsupervised federated ReID method FedUReID [81]. On DukeMTMC-reID and Market1501, which are the two largest datasets, the proposed method achieves 81.8%, 94.4% Rank-1 accuracy, respectively. It significantly outperforms FedUReID with 51.0%, 65.2% Rank-1 accuracy on the two datasets, respectively. More importantly, more significant improvements can be observed on smaller datasets. For example, the performance on CUHK03 is boosted from 8.9% to 71.2% on Rank-1 accuracy when comparing C^2IT with FedUReID. Other datasets also demonstrate the effectiveness of the proposed method, which increases the average Rank-1 accuracy by 35.4% compared with FedUCC.

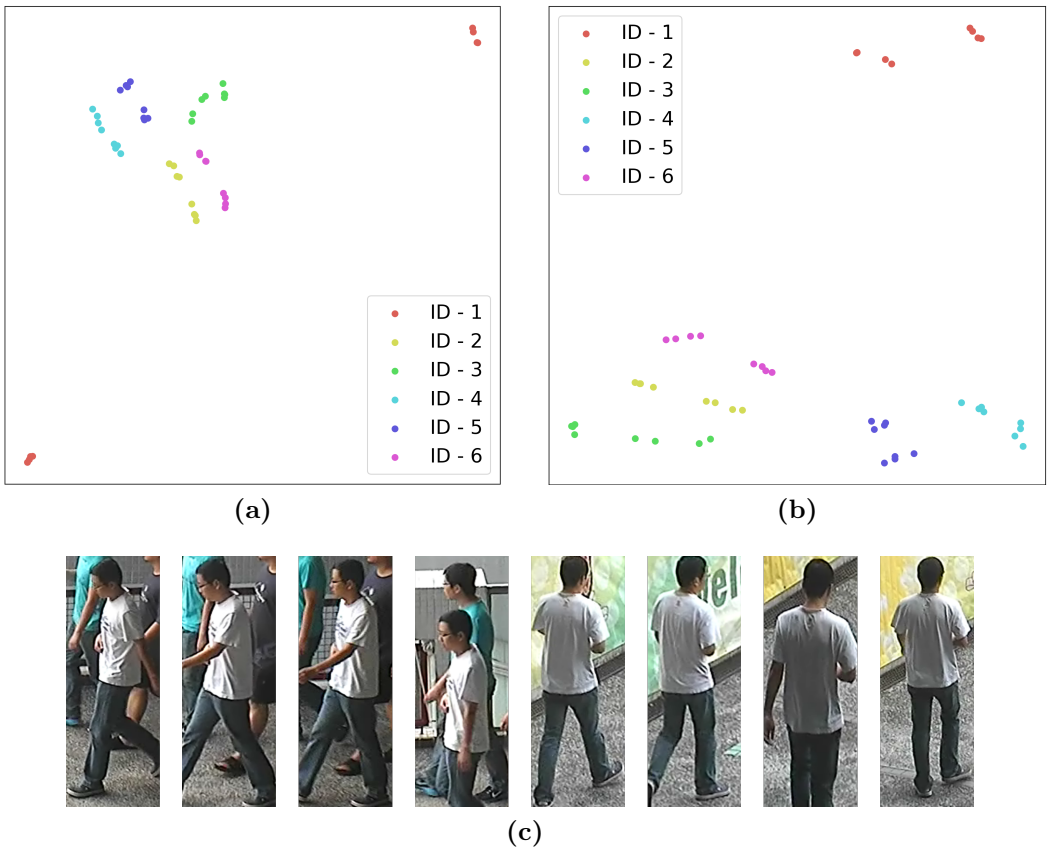


Figure 4.2: t-SNE analysis on (a) baseline and (b) C^2IT; and (c) the images of ID-1 (in red) contain diverse background contents and a general modelling strategy results in two distant clusters.

4.5.3 Ablation Study on Context-Invariant Learning

To demonstrate the effectiveness of the proposed context-invariant learning mechanism, the results of ablation studies are shown in Table 4.1. Compared with the baseline method (ViT-S) not involving the proposed two learning mechanisms, our proposed method increases the average Rank-1 accuracy over all eight datasets with an average from 80.6% to 76.9%. In particular, PRID has improved the Rank-1 accuracy from 83.9% to 86.9% and iLIDS from 82.8% to 85.5%. Performance improvement is also observed on other datasets, compared with the baseline. The different levels of performance improvement indicate the level of cluttered background information that influences the retrieval of a person's feature. On Duke and CUHK01, however, there was a slight decrease in terms of Rank-1 accuracy from 82.0% to

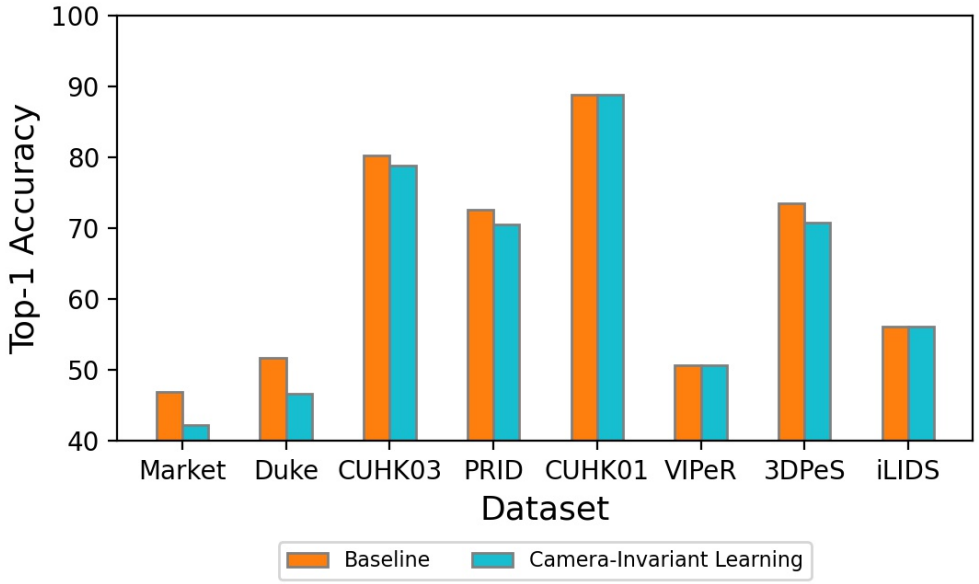


Figure 4.3: Multinomial logistic regressions for camera identification.

81.8% and 91.6% to 91.1%, respectively. This is due to the issue of optimization instability. That is, the unsupervised learning model tends to have fluctuated performance given the image label change across federated training rounds. Overall, the average accuracy increase indicates the effectiveness of our context-invariant learning strategy.

In addition, t-SNE [59] analysis is conducted to analyse the distributions of the obtained image-level representations. Image representations of 6 IDs are selected for visualisation as shown in Figure 4.2. Compared with the baseline model, context-invariant learning derives a more separable inter-ID distribution. Specifically, the ID-1 associated images (in red dots) are distributed closely under the context-invariant learning, whilst the baseline method represents them into two distant clusters. This is consistent with the ID-1 associated images as shown in Figure 4.2 (c), where two major visual discrepancies regarding the background contents can be observed: the crowd scenes contain irrelevant information about body parts of other people and the single-person scenes at another location contain a billboard. The proposed context-invariant learning mechanism successfully disentangles these irrelevant patterns.

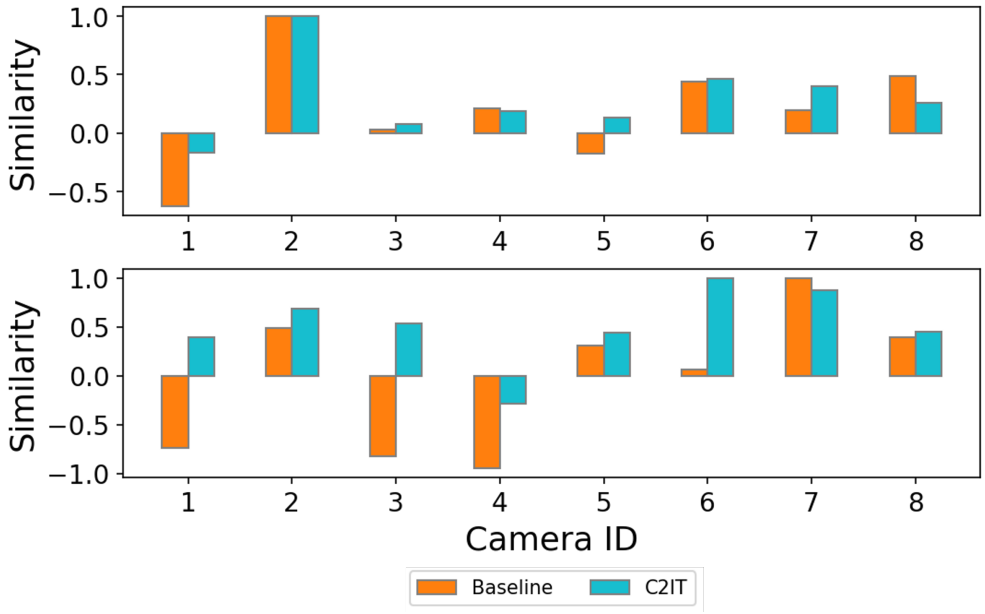
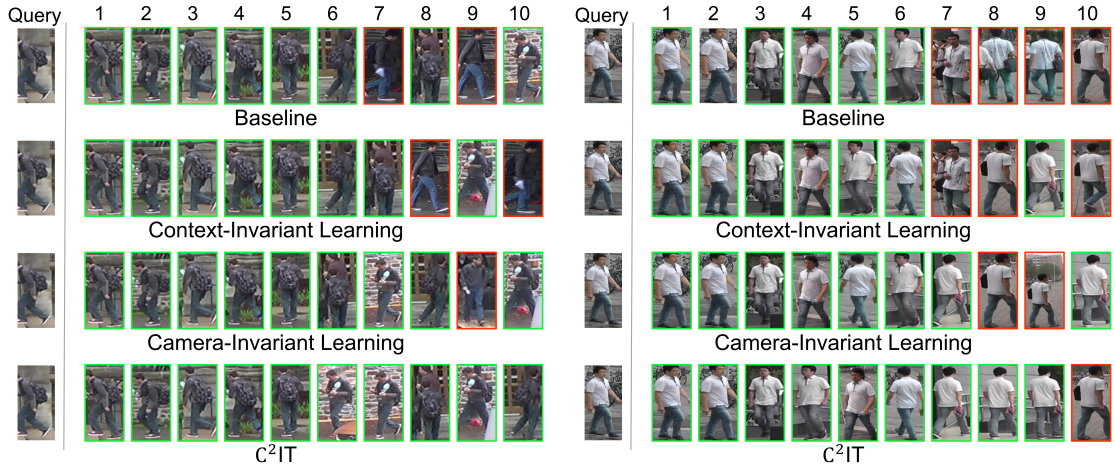


Figure 4.4: Selected two Duke images and their camera similarities.

4.5.4 Ablation Study on Camera-Invariant Learning

As shown Table 4.1, with the camera-invariant learning strategy, the average Rank-1 accuracy is effectively increased from 80.6% to 80.9%. Specifically, the performance on PRID is increased, where the rank-1 accuracy is from 83.9% to 85.9% and on iLIDS increased from 82.8 to 85.5%. To further demonstrate the effectiveness of camera-invariant learning, multinomial logistic regression is conducted to discriminate the cameras on the obtained image feature representations per the eight datasets for the baseline and C^{2}IT. As the top-1 accuracy illustrated in Figure 4.3, compared with the baseline, the camera-invariant representations have less capability to classify themselves with the camera labels as expected. Such effects are more obvious with the clients with more cameras (e.g., Market - 6 cameras and Duke - 8 cameras). Specifically, two examples in Duke dataset are illustrated in Figure 4 with their normalised similarity values between camera tokens and image representations. Figure 4 (a) is captured with camera 1 and (b) is with camera 7. Fewer similarities can be observed between the image representations and their associated cameras under C^{2}IT compared with the baseline.



(a) Duke **(b)** Market

(c) iLIDS

Figure 4.5: Qualitative analysis and the results of different methods for three queries from three types of datasets: large, medium and small.

4.5.5 Qualitative Analysis

In Figure 4.5, we demonstrate the retrieval results of our proposed methods with the baseline and ablation settings: 1) a conventional unsupervised federated baseline with transformer backbone, 2) the baseline with the context-invariant learning strategy and 3) the baseline with the camera-invariant learning strategy. Particularly, three sample queries were selected from the Duke, Market, and iLIDS datasets, of which the sizes are large, medium, and small, respectively.

Though the baseline method is able to formulate discriminative representations

to some extent, the environmental variations still impact its capability to identify the most informative features from the foreground pedestrian regions. For example, as shown in Figure 4.5 (a), the 10-th retrieved image of the same person has a lower confidence score than two negative images. This is largely due to the background difference, specifically, the color and texture of the bricks in the retrieved image being different from the query image. Similarly, for the query from Market, the retrieved images tend to have a similar chaotic background within the upper region, especially for the negative samples, which match the background visual patterns of the query image. Our proposed context-invariant learning strategy successfully addresses the issue of irrelevant context similarity by reducing the importance of these regions and focusing on the tokens associated with the pedestrian body. As a result, in Figure 4.5 (a), the negative samples have decreased priorities and the positive ones have diverse backgrounds. Similarly, the query from Market shown in Figure 4.5 (b) has successfully retrieved other back-view snapshots with a cleaner background. In Figure 4.5 (c) the results for the query from iLIDS also demonstrate a similar conclusion, where the partially occluded query image can successfully match with another occluded snapshot ranked at the 10-th.

The retrieved results with the camera-invariant learning strategy are also visualised, where the retrieval quality on the Market and Duke datasets has improved in terms of the quantity and ranking of the positive samples compared with the baseline. Lastly, the combination of the two strategies has further refined the image retrieval results for better performance. In all three retrieval examples, our proposed method obtains the most relevant results in terms of the number of positive samples and their ranking.

4.6 Summary

In this chapter, we present a novel transformer based deep learning method - C^2IT to address the cross-domain statistical heterogeneity in federated unsupervised person ReID. C^2IT consists of two key learning strategies in pursuit of context and camera invariant pedestrian image representations. Context-invariant learning helps reduce the irrelevant contents in image-level representations, and camera-invariant learning formulates consistent representations alleviating the impacts caused by different camera settings.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis investigates the challenging problem of unsupervised Person ReID by devising federated learning schemes to enable collaborative training on multi-source datasets to reduce the risks of privacy leakage. Federated Person ReID, compared with general federated learning, suffers from a more severe non-IID issue, resulting in the aggregated model not only fluctuating significantly in performance but also showing inferior performance than locallt trained models. The objective of this thesis is to improve the aggregated model's capacity to acquire general and robust features across diverse environments while preventing it from over-fitting with the domain-specific patterns. The contributions are summarised as follows:

1. A three-stage learning strategy is studied with coarse-to-fine feature learning, exploiting patch-level representation learning to enhance cross-client consensus learning.

2. A Context and Camera Invariant Transformer architecture is presented for learning context-disentangled Person ReID representations to alleviate the varied data distributions across clients.

5.2 Limitations & Future Directions

The scope of future studies can be expanded to include, but is not limited to, the following aspects:

- During clustering, a small number of images in each cluster are considered outliers and are discarded to reduce their negative impacts. However, this practice may ignore the important information contained in these images. In fact, outlier images can often exhibit unique features and characteristics that are not present in the clusters. By a proper strategy to further modelling with them, Person ReID is expected to have the potential to achieve a better representation diversity.

- The camera related information is still remained within the final image representations as indicated in Figure 4.4, though the proposed mechanism decreases them to some extent. The ideal case is that the cameras can not be differentiated with the obtained representations. Thus, an adversarial learning strategy can be studied for the camera-invariant learning in the future. The challenge for an adversarial design is that adversarial learning under the federated setting can be more difficult when exchanging local knowledge across clients. Moreover, the current practice of process the camera tokens is within their own clients, and explicit cross-client strategies can be further considered.

- Exchanging the model weights, which has been commonly used as the cross-domain knowledge under a federated setting, can lead to personal sensitive information leakage as suggested in some recent studies(e.g., [3]). Stricter learning protocols for privacy preserving purposes such as differentiable privacy [80] should be considered for federated person ReID.

Bibliography

[1] Manoj Ghuhan Arivazhagan et al. "Federated learning with personalization layers". In: *arXiv preprint arXiv:1912.00818* (2019).

[2] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. "3DPeS: 3D People Dataset for Surveillance and Forensics". In: *Joint ACM Workshop on Human Gesture and Behavior Understanding*. 2011.

[3] Nicholas Carlini et al. "Extracting training data from large language models". In: *USENIX*. 2021.

[4] Haoran Chen et al. "Deep transfer learning for person re-identification". In: *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2018, pp. 1–5.

[5] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. "Instance-guided context rendering for cross-domain person re-identification". In: *International Conference on Computer Vision*. 2019.

[6] Zuozhuo Dai et al. "Cluster Contrast for Unsupervised Person Re-Identification". In: *arXiv preprint arXiv:2103.11568* (2021).

[7] Zuozhuo Dai et al. "Cluster contrast for unsupervised person re-identification". In: *ACCV*. 2022.

[8] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition*. 2009.

[9] Weijian Deng et al. “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification”. In: *Computer Vision and Pattern Recognition*. 2018.

[10] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[11] Mingjing Du, Shifei Ding, and Hongjie Jia. “Study on density peaks clustering based on k-nearest neighbors and principal component analysis”. In: *Knowledge-Based Systems* 99 (2016), pp. 135–145.

[12] Martin Ester et al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: KDD’96. AAAI Press, 1996.

[13] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. “Personalized federated learning: A meta-learning approach”. In: *arXiv preprint arXiv:2002.07948* (2020).

[14] Hehe Fan et al. “Unsupervised Person Re-identification: Clustering and Fine-tuning”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications TOMM* 14.4 (2018), 83:1–83:18. DOI: 10.1145/3243316.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. “Model-agnostic meta-learning for fast adaptation of deep networks”. In: *International Conference on Machine Learning*. 2017.

[16] Dengpan Fu et al. “Unsupervised Pre-training for Person Re-identification”. In: *CVPR* (2021).

[17] Shang Gao et al. “Pose-guided visible part matching for occluded person ReID”. In: *CVPR*. 2020, pp. 11744–11752.

[18] Yixiao Ge, Dapeng Chen, and Hongsheng Li. “Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification”. In: *International Conference on Learning Representations*. 2020.

[19] Avishek Ghosh et al. “An efficient framework for clustered federated learning”. In: *arXiv preprint arXiv:2006.04088* (2020).

[20] Avishek Ghosh et al. “Robust federated learning in a heterogeneous environment”. In: *arXiv preprint arXiv:1906.06629* (2019).

[21] Douglas Gray and Hai Tao. “Viewpoint invariant pedestrian recognition with an ensemble of localized features”. In: *European Conference on Computer Vision*. 2008.

[22] Kaiming He et al. “Deep residual learning for image recognition”. In: *Computer Vision and Pattern Recognition*. 2016.

[23] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Computer Vision and Pattern Recognition*. 2020.

[24] Shuting He et al. “Transreid: Transformer-based object re-identification”. In: *ICCV*. 2021.

[25] Alexander Hermans, Lucas Beyer, and Bastian Leibe. “In defense of the triplet loss for person re-identification”. In: *arXiv preprint arXiv:1703.07737* (2017).

[26] Martin Hirzer et al. “Person re-identification by descriptive and discriminative classification”. In: *Scandinavian Conference on Image Analysis*. 2011.

[27] Takashi Isobe et al. “Towards discriminative representation learning for unsupervised person re-identification”. In: *International Conference on Computer Vision*. 2021.

[28] Zilong Ji et al. “An attention-driven two-stage clustering method for unsupervised person re-identification”. In: *European Conference on Computer Vision*. 2020.

[29] Peter Kairouz et al. “Advances and open problems in federated learning”. In: *Foundations and Trends® in Machine Learning* 14.1–2 (Jan. 2021), pp. 1–210.

[30] Daliang Li and Junpu Wang. "Fedmd: Heterogenous federated learning via model distillation". In: *arXiv preprint arXiv:1910.03581* (2019).

[31] Minxian Li, Xiatian Zhu, and Shaogang Gong. "Unsupervised tracklet person re-identification". In: *Tpami* 42.7 (2019), pp. 1770–1782.

[32] Qinbin Li, Bingsheng He, and Dawn Song. "Model-contrastive federated learning". In: *Computer Vision and Pattern Recognition*. 2021.

[33] Tian Li et al. "Federated optimization in heterogeneous networks". In: *Machine Learning and Systems* (2020).

[34] Wei Li, Rui Zhao, and Xiaogang Wang. "Human reidentification with transferred metric learning". In: *Asian Conference on Computer Vision*. 2012.

[35] Wei Li et al. "Deepreid: Deep filter pairing neural network for person re-identification". In: *Computer Vision and Pattern Recognition*. 2014.

[36] Xiaoxiao Li et al. "FedBN: Federated Learning on Non-IID Features via Local Batch Normalization". In: *International Conference on Learning Representations*. 2021.

[37] Xulin Li et al. "Counterfactual Intervention Feature Transfer for Visible-Infrared Person Re-identification". In: *ECCV*. 2022, pp. 381–398.

[38] Paul Pu Liang et al. "Think locally, act globally: Federated learning with local and global representations". In: *arXiv preprint arXiv:2001.01523* (2020).

[39] Youwei Liang et al. "Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations". In: *ICLR*. 2022.

[40] Xiangtan Lin et al. "Unsupervised person re-identification: A systematic survey of challenges and solutions". In: *arXiv preprint arXiv:2109.06057* (2021).

[41] Yutian Lin et al. "A bottom-up clustering approach to unsupervised person re-identification". In: *AAAI Conference on Artificial Intelligence*. 2019.

[42] Jiawei Liu et al. “Ca3net: Contextual-attentional attribute-appearance network for person re-identification”. In: *ACM international conference on Multimedia*. 2018.

[43] Quande Liu et al. “Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space”. In: *Computer Vision and Pattern Recognition*. 2021.

[44] Hao Luo et al. “Self-Supervised Pre-Training for Transformer-Based Person Re-Identification”. In: *arXiv preprint arXiv:2111.12084* (2021).

[45] Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial Intelligence and Statistics*. 2017.

[46] Zhangqiang Ming et al. “Deep learning-based person re-identification methods: A survey and outlook of recent works”. In: *Image and Vision Computing* 119 (2022), p. 104394.

[47] Henri J Nussbaumer. “The fast Fourier transform”. In: *Fast Fourier Transform and Convolution Algorithms*. 1981.

[48] Liangqiong Qu et al. “Rethinking architecture design for tackling data heterogeneity in federated learning”. In: *CVPR*. 2022.

[49] Ergys Ristani and Carlo Tomasi. “Features for multi-target multi-camera tracking and re-identification”. In: *CVPR*. 2018, pp. 6036–6046.

[50] Ergys Ristani et al. “Performance measures and a data set for multi-target, multi-camera tracking”. In: *European Conference on Computer Vision*. 2016.

[51] Yiqing Shen, Yuyin Zhou, and Lequan Yu. “CD2-pFed: Cyclic Distillation-guided Channel Decoupling for Model Personalization in Federated Learning”. In: *Computer Vision and Pattern Recognition*. 2022.

[52] Debaditya Shome and Tejaswini Kar. “FedAffect: Few-shot federated learning for facial expression recognition”. In: *International Conference on Computer Vision*. 2021.

[53] Chunfeng Song et al. “Mask-guided contrastive attention model for person re-identification”. In: *CVPR*. 2018, pp. 1179–1188.

[54] Chi Su et al. “Pose-driven deep convolutional model for person re-identification”. In: *ICCV*. 2017.

[55] Shitong Sun, Guile Wu, and Shaogang Gong. “Decentralised Person Re-Identification with Selective Knowledge Aggregation”. In: *British Machine Vision Conference* (2021).

[56] Yifan Sun et al. “Beyond part models: Person retrieval with refined part pooling”. In: *European Conference on Computer Vision*. 2018.

[57] Yifan Sun et al. “Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification”. In: *CVPR*. 2019.

[58] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: *arXiv preprint arXiv:1807.03748* (2018).

[59] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[60] Dongkai Wang and Shiliang Zhang. “Unsupervised person re-identification via multi-label classification”. In: *Computer Vision and Pattern Recognition*. 2020.

[61] Menglin Wang et al. “Camera-aware proxies for unsupervised person re-identification”. In: *AAAI Conference on Artificial Intelligence*. 2021.

[62] Taiqing Wang et al. “Person re-identification by video ranking”. In: *European Conference on Computer Vision*. 2014.

[63] Yicheng Wang et al. “Person re-identification with cascaded pairwise convolutions”. In: *CVPR*. 2018, pp. 1470–1478.

[64] Longhui Wei et al. “Person transfer gan to bridge domain gap for person re-identification”. In: *Computer Vision and Pattern Recognition*. 2018.

[65] Guile Wu and Shaogang Gong. “Decentralised learning from independent multi-domain labels for person re-identification”. In: *AAAI Conference on Artificial Intelligence*. 2021.

[66] Shiyu Xuan and Shiliang Zhang. “Intra-inter camera similarity for unsupervised person re-identification”. In: *Computer Vision and Pattern Recognition*. 2021.

[67] Cheng Yan et al. “BV-person: a large-scale dataset for bird-view person re-identification”. In: *ICCV*. 2021, pp. 10943–10952.

[68] Fan Yang et al. “Attention driven person re-identification”. In: *Pattern Recognition* 86 (2019), pp. 143–155.

[69] Qize Yang et al. “Patch-based discriminative feature learning for unsupervised person re-identification”. In: *Computer Vision and Pattern Recognition*. 2019.

[70] Wang Zeng et al. “Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer”. In: *CVPR*. 2022.

[71] Lin Zhang et al. “Federated learning for Non-IID data via Unified Feature learning and Optimization objective alignment”. In: *International Conference on Computer Vision*. 2021.

[72] Pengyi Zhang et al. “Adaptive Cross-domain Learning for Generalizable Person Re-identification”. In: *ECCV*. Springer. 2022, pp. 215–232.

[73] Xiaokang Zhang et al. “Semantic-aware occlusion-robust network for occluded person re-identification”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.7 (2020), pp. 2764–2778.

[74] Yuyang Zhao et al. “Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification”. In: *Computer Vision and Pattern Recognition*. 2021.

[75] Liang Zheng, Yi Yang, and Alexander G Hauptmann. “Person re-identification: Past, present and future”. In: *arXiv preprint arXiv:1610.02984* (2016).

[76] Liang Zheng et al. “Scalable person re-identification: A benchmark”. In: *International Conference on Computer Vision*. 2015.

[77] Meng Zheng et al. “Re-identification with consistent attentive siamese networks”. In: *CVPR*. 2019, pp. 5735–5744.

[78] Zhun Zhong et al. “Generalizing a person retrieval model hetero-and homogeneously”. In: *European Conference on Computer Vision*. 2018.

[79] Zhun Zhong et al. “Re-ranking person re-identification with k-reciprocal encoding”. In: *Computer Vision and Pattern Recognition*. 2017.

[80] Tianqing Zhu et al. “More than privacy: Applying differential privacy in key areas of artificial intelligence”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.6 (2020), pp. 2824–2843.

[81] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. “Joint optimization in edge-cloud continuum for federated unsupervised person re-identification”. In: *ACM International Conference on Multimedia*. 2021.

[82] Weiming Zhuang et al. “Performance optimization of federated person re-identification via benchmark analysis”. In: *ACM International Conference on Multimedia*. 2020.