FedCAPR:Federated Camera-Aware Unsupervised Person Re-Identification with Identity-Distributed Equalization for Decentralized Data Clustering

Yu-Syuan Tseng[1†], Tzu-Ching Hsu[1†], Chih-Ting Liu[2], Shao-Yi Chien[1]
[1]Graduate Institute of Electronics Engineering, National Taiwan University, [2]Amazon († Equal contribution.)
{ystseng, hru059026}@media.ee.ntu.edu.tw, liuchiht@amazon.com, sychien@ntu.edu.tw

Abstract

With the emergence of smart cities, person re-identification (re-ID) has become crucial for tracking individuals across surveillance footage. However, the acquisition of large-scale datasets with human annotations is not practical for real-world scenarios. Thus, purely unsupervised learning methods are widely employed in re-ID application. On the other hand, handling of sensitive personal data causes significant privacy concern. Fortunately, federated learning (FL) has been introduced to address privacy problems without exchanging data collected by each local client. To tackle both issues at the same time, previous works integrate unsupervised learning with FL on numerous re-ID benchmarks. Nevertheless, these existing methods face the difficulties due to the variance and complexity of data among clients, as well as the issue of information imbalance caused by traditional aggregation strategies such as FedAvg [18]. Therefore, we proposed FedCAPR, a novel federated unsupervised re-ID framework, which contains four principle features. First, we introduced a novel camera-aware contrastive loss to leverage camera information, significantly improving re-ID accuracy. Second, we incorporated a regularization penalty to ensure consistency and stability in the local federated optimization. Third, we developed the identity-distributed equalization (IDE) mechanism to address data heterogeneity across clients. Last, for FL aggregation, we introduced the cosine-distance weighted (CDW) method to guide the global model with a more effective aggregation during each round. Compared to state-of-the-art methods, FedCAPR achieves superior performance, up to a 23.4% improvement, on eight benchmark datasets. The source code is made publicly available: https://github.com/maho666/FedCAPR.

1. Introduction

Person re-identification (re-ID) is vital for urban security and surveillance systems, enabling the identification of individuals across multiple camera views. Traditional re-ID methods rely on supervised learning, requiring extensive labeled datasets, which are often impractical for large-scale applications. Unsupervised learning presents a promising alternative, eliminating the need for ground truth identity labels. Unsupervised person re-ID models can be divided into two categories, domain adaptive and purely unsupervised person re-ID. Domain adaptive methods use labeled data from source domain to train a model capable of aligning features from source to unlabeled target domain [1–5]; purely unsupervised techniques utilize clustering-based [6, 7, 12–14, 28], graph-based [8, 9] or k-NN [10, 11] methods to generate pseudo-labels for data and perform supervised training to improve performance. Although unsupervised person re-ID methods address the challenge of obtaining annotated datasets and show promising accuracy improvement, this typical training pipeline, which is conducted by collecting large-scale unsupervised pedestrian images and optimizing the model in a centralized way, causes the privacy concern since personal sensitive data are utilized directly.

Federated learning (FL), a distributed machine learning framework, has emerged as a promising solution to address data privacy concerns. By allowing multiple clients to train local models on their private data and aggregating these models into a shared global model without exchanging raw data, FL ensures that sensitive information remains localized and secure [18–27, 29–31]. This approach eliminates the need for direct data sharing among participants, thereby preserving privacy while enabling collaborative learning through model aggregation.

Hence, the recent works integrate unsupervised and federated learning to form a robust framework for person re-ID, effectively tackling challenges related to labeling and privacy [32, 33]. These works assign eight benchmark datasets as eight local clients, including DukeMTMC (abbr. Duke) [43], Market-1501 (abbr. Market) [44], iLIDS-VID (abbr. iLIDS) [45], CUHK03 [35], PRID2011 (abbr. PRID) [46], VIPeR [48], CUHK01 [47], and 3DPeS [49], to simulate the cross-clients person re-ID system (issues related to Duke are discussed in Sec. 4.1). Nevertheless, the accuracy of recent federated unsupervised re-ID works suffers from substantial data variances, such as shooting angle, lighting, occlusion and color, caused by different camera positions and data distributions across these datasets. These variations make building the initial clustering in unsupervised learning more challenging. The error propagation of clustering results throughout the whole training processing will significantly degrade the performance of models on each client. Besides, there are significant disparities in the number of images and identities between the datasets. The imbalance sizes of datasets also lead to the reduction of performance of model since larger dataset, like Market, dominates the performance of global model and the clients with smaller datasets fail to produce valuable updates for the global model. It leads to the federated process becoming ineffective and disrupting the learning for all clients.

To solve the challenges which are mentioned above, we proposed FedCAPR, a novel federated unsupervised re-ID framework. In each local re-ID training, in additional to typical unsupervised loss [33], we leverage the inherent camera information of each image to design a camera-aware contrastive loss. This loss function enhances the model's ability to distinguish instances of the same identity across different camera views, effectively addressing the challenge of high variance in data caused by camera differences. Meanwhile, to address model divergence in federated learning, we introduce a regularization loss that constrains the deviation between local and aggregated global models during training. We found that the disparity of data quantity among clients influence the performance a lot. Thus, for local training, we proposed a preprocessing method that based on the number of images per identity to adjust the dataset size, which can facilitate the pseudo-labels generation. For the federated framework, instead of FedAvg [18], we propose a unsupervised cosine-distance weighted aggregation method that assigns weights based on the knowledge gained by local models. This approach addresses the diminished contributions of smaller datasets, ensuring that smaller datasets contribute meaningfully to the global model.

In summary, there are four principle contributions of FedCAPR as follows:

- **Camera-Aware Contrastive Learning Loss**: We propose a camera-aware contrastive loss to improve identity distinction across different cameras in unsupervised re-ID.
- **Cosine Similarity Regularization**: We introduce a regularization penalty to reduce deviation between local and global models, ensuring consistency during training.
- **Identity-Distributed Equalization**: Our pre-processing technique balances the average number of images per identity to adjust the dataset size, improving clustering stability and mitigating performance degradation.
- **Cosine-Distance Weighted Aggregation**: We propose a cosine-distance weighted aggregation to address the diminished contributions of smaller datasets in FedAvg.

2. Related work

2.1. Unsupervised Person Re-ID

Unsupervised Person re-ID includes domain adaptive and purely unsupervised approaches. The domain adaptive person re-ID methods [1–3] utilize labeled data from a source domain to train a model that generalize to an unlabeled target domain. As demonstrated in [4], transfer learning techniques help mitigate domain gaps by transferring spatial-temporal features. Purely unsupervised person re-ID methods [6, 7] rely entirely on unlabeled data, employing clustering-based, graph-based [8, 9], or k-NN approaches [10, 11] to generate pseudo-labels. The prevalent clustering methods include Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [12] and Bottom-Up Clustering (BUC) [13], which iteratively refine pseudo-labels through soft multilabel learning [14] to enhance accuracy. Furthermore, CAP [28] incorporates a camera-aware proxy loss to address intra-camera variance and a contrastive loss for inter-camera variance, effectively mitigating noise from camera variances during training.

2.2. Federated Learning

Federated learning (FL) enables decentralized model training while preserving data privacy by allowing participants to train locally and share only model updates. Traditional federated learning aims to build a global model that generalizes across different clients, with FedAvg [18] as the main algorithm for model aggregation. However, statistical heterogeneity hinder performance, prompting methods like FedProx [19] to address this issue.

2.3. Unsupervised Federated Person Re-ID.

This approach addresses the challenges of data privacy and the need for diverse datasets. Recent contributions include FedUCC(Weng et al., 2023) [32], which utilizes a coarse-to-fine framework comprising four stages: starting with broad information, preserving personalized data through BatchNorm[51], refining features from image patches, and decomposing features to mitigate cross-camera divergence. This process demands significant communication, requiring up to 100 rounds. Another key work is FedUCA [33], which integrates the CAP [28] method with FL via EasyFL [34]. It emphasizes the importance of addressing camera variance in with the unsupervised manner under the federated scenario.

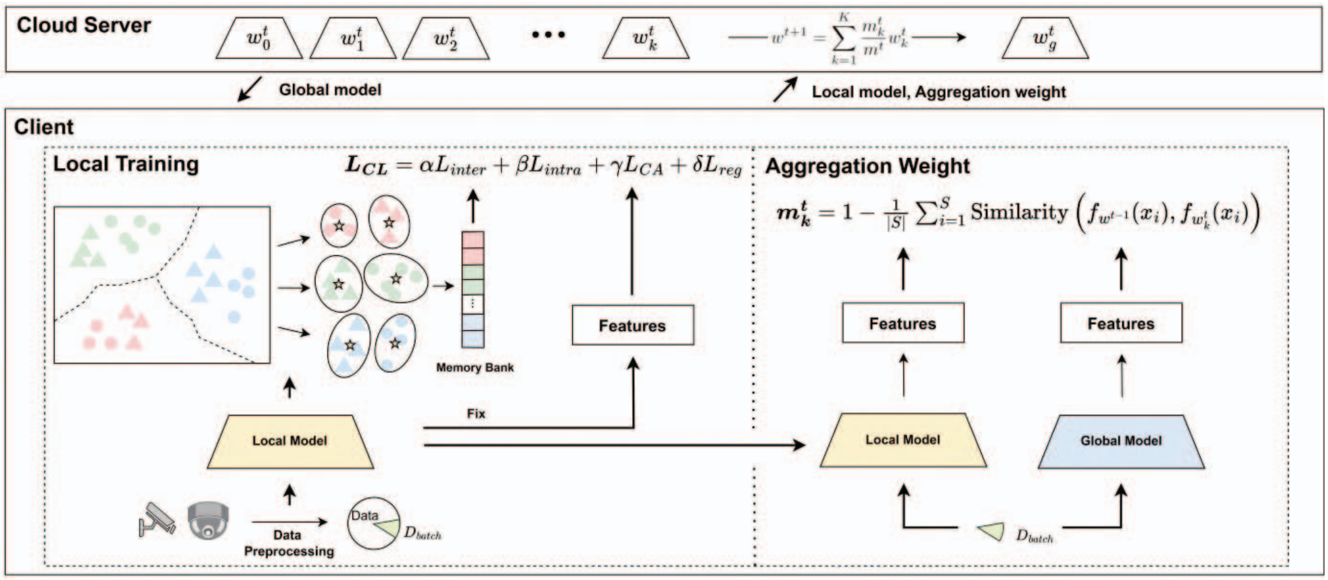

Figure 1. **Overall architecture of FedCAPR.** Each client generates pseudo-labels for preprocessing unlabeled data by clustering features with camera information, storing them in the memory bank that updates to calculate the contrastive loss. The server aggregates the client models with aggregation weight. In the cluster diagram, each shape represents pid, and each color represents a camera.

3. Methodology

This section outlines our unsupervised federated person re-ID approach, and the overall architecture is shown in Fig. 1. Our federated learning framework involves a server and K clients, each corresponding to a distinct dataset k. To enhance performance, we will introduce a pre-processing step to adjust each local dataset for further clustering step. During local training, each client performs unsupervised learning using pseudo-labels generated via DBSCAN[12] clustering respectively. The training objective combines three contrastive losses, consisting of inter-camera, intra-camera, and camera-aware loss terms, designed to enhance unsupervised learning performance. Additionally, a regularization loss term is incorporated to improve model consistency during aggregation. After local model updates, clients upload their models to the server, where aggregation is performed using a cosine-distance weighted mechanism. The above mechanism will be conducted for many rounds until convergence. Algorithms 1 illustrate the overall process of our proposed approach.

The remainder of this section is organized as follows: Sec. 3.1, 3.2, and 3.4 focus on improving local unsupervised learning, while Sec. 3.3 and 3.5 targets challenges in federated learning.

3.1. Proxy Creation and the Memory Bank

In person re-ID, following CAP [28], the training process alternates between a clustering step and a model updating step. The clustering step generates pseudo-labels for the unlabeled data, while the model updating step trains the model using these pseudo-labels.

For unsupervised training on each local client, we employ a CNN backbone $f_{w_k^t}$ to extract features from the unlabeled dataset $D_k = \{x_i\}_{i=1}^N$, where x_i indicates the images in the dataset and N is the total number of the images. Note that the local k^{th} dataset won't be shared with other clients. We then utilize DBSCAN to assign pseudo-labels $\hat{y}_i$ to each sample x_i. Samples identified as outliers by DBSCAN are removed from the dataset and do not participate in subsequent training of the same epoch. Next, we assign a proxy label $\hat{z}_i$ by dividing the pseudo-labels into smaller clusters based on camera information, which means each sub-cluster contains only samples from the same camera with the same pseudo-label (pid). Consequently, we derive a labeled dataset $D'_k = \{(x_i, \hat{y}_i, \hat{z}_i)\}_{i=1}^{N'}$, where N' represents the total number of images in the dataset after removing outliers. $\hat{y}_i \in \{1, 2, \ldots, Y\}$ represents the generated pseudo-label and $\hat{z}_i \in \{1, 2, \ldots, Z\}$ represents the proxy label. For each sub-cluster $\hat{z}$, we calculate the average feature representation from CNN with d dimension to generate a proxy stored in the memory bank $M \in \mathbb{R}^{d \times Z}$, where $Z = \sum_{c=1}^{C} Z_c$ is the total number of proxies across all cameras, which follows the approach in [28], corresponding to the averaged feature of instances from the same camera and the same pid. These proxies are then used as samples for the contrastive learning loss during back-propagation, and the memory bank is updated as described by the following equation:

$$M[\hat{z}_i] \leftarrow \mu M[\hat{z}_i] + (1-\mu) f_{w_k}(x_i), \tag{1}$$

where $M[\hat{z}_i]$ represents the memory bank entry for proxy $\hat{z}_i$, sample x_i belongs to $\hat{z}_i$, and $\mu \in [0,1]$ is the updating rate. Dynamic adjustments ensure the memory bank's features reflect the current model state.

3.2. Camera-Aware Contrastive Learning Loss

In this section, we present the contrastive losses in our framework. We utilize three types of contrastive learning loss: intra-camera loss L_{intra}, inter-camera loss L_{inter} and our proposed camera-aware loss L_{CA}, where the intra- and inter-camera losses are proposed by CAP [28]. The inter-camera loss explicitly leverages the correlations between cameras to enhance the model's ability. On the contrary, an intra-camera loss reduces the label noise and obtains a better discrimination ability within a single camera. Nonetheless, the model with those two losses still suffer from data with the considerable camera variances between the same identity. Thus, we propose a novel camera-aware loss to confront the significant camera variance. The combined loss function for the proposed contrastive learning approach is defined as follows:

$$L_{CL} = \alpha L_{inter} + \beta L_{intra} + \gamma L_{CA}, \tag{2}$$

where α, β, and γ are coefficients that control the relative contributions of the three losses.

Inter-Camera Loss. In contrastive learning, the goal is to bring positive pairs closer together and push negative pairs further apart. The inter-camera loss is constructed based on the generated pseudo-labels (pid). For each anchor x_i, the proxy stored in the memory bank with the same pid $\hat{y}_i$ are considered as its positive pairs; thus, samples with the same pid but under different cameras will all be pulled closed. On the other hand, proxies with $\hat{y} \neq \hat{y}_i$ are treated as negative samples. Note that following [28], we employ the K-nearest strategy to select only hard negative proxies for pushing apart. The inter-camera loss, L_{inter}, is defined as:

$$L_{\text{inter}} = -\sum_{i=1}^{N'} \frac{1}{|P|} \sum_{p \in P} \log \frac{S(p, x_i)}{Z(x_i)}, \tag{3}$$

$$Z(x_i) = \sum_{u \in P} S(u, x_i) + \sum_{q \in Q} S(q, x_i), \tag{4}$$

where $S(p, x_i) = e^{(M[p]^T f_{w_k}(x_i)/\tau)}$, and the P and Q represents the sets of the positive and hard negative proxies, respectively.

Intra-Camera Loss. This loss aims to distinguish the different identities under the same camera, thereby significantly reducing the noise introduced by the variation

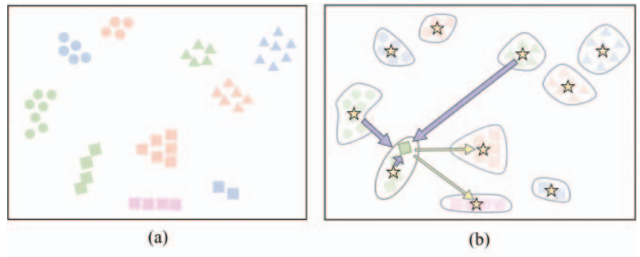

Figure 2. **Concept of camera-aware loss L_{CA}.** Same-shape samples come from the same camera, and same color represent same pid. (a) Applying L_{intra} , clustering results remain suboptimal due to the significant camera variance (b) From the proxies of each camera, choose the closest proxies to the green square anchor x_i as the positive, while the negative set consists of the K_2-nearest proxies, excluding the positive one, with purple arrows pulling similar proxies closer and yellow arrows pushing the hard negatives away.

between cameras. To achieve that, we denote a newly constructed dataset for each camera, for instance, $D_k^{a\prime} = \{x_i, \hat{y}_i, \tilde{z}_i, a\}_{i=1}^{N_a}$ indicates the sub-dataset from camera c_a, where N_a represents the total number of images captured by camera a. The variable $\tilde{z}_i$ indicates the proxy labels derived from camera a, $\tilde{z}_i \in \{1, 2, \ldots, Z_a\}$, where Z_a represents the total number of unique proxy labels in camera a. The formulated loss function is then expressed as follows:

$$L_{\text{intra}} = -\sum_{j=1}^{C} \frac{1}{N_c} \sum_{x_i \in D_k^{a\prime}} \log \left(\frac{S(m, x_i)}{\sum_{k=A+1}^{A+Z_{ci}} S(k, x_i)} \right), \tag{5}$$

For an image x_i labeled with camera c_j, let $A = \sum_{c=1}^{c_j - 1} Z_c$ denote the total number of proxies associated with camera IDs that are different from c_j. The corresponding entry index m in the memory bank is then calculated as $m = A + \tilde{z}_i$.

Camera-Aware Loss. Within a single camera, we use L_{intra} to address the classification problem, but camera variance between the same identity still poses a challenge, as shown in Fig. 2 (a). When the same person ID is captured by different cameras, the anchor may not cluster correctly due to these variations. For example, if camera c_a captures images in a dimly lit room, while camera c_b takes photos in bright sunlight, the model might rely more on the lighting conditions rather than learning the distinct features of the person's appearance. Images of the same identity from cameras c_a and c_b are placed far apart in the feature space. Additionally, when pseudo labels $\hat{y}_i$ are split into camera-aware proxies $\hat{z}_i$, the error is reinforced by assigning them to the same proxy. To solve this, we propose a new loss function called camera-aware loss L_{CA}, which significantly improves the clustering quality by associating a positive sample from each camera with the anchor x_i.

The issue of camera variance arises only when the same identity appears across multiple cameras. Therefore, we assume that each camera contains at least one positive proxy. For each anchor, we locate the nearest proxy from each camera to form a positive pair, as illustrated in Fig. 2 (b). Positive proxies are chosen based on the similarity of the features between different cameras, preventing the model from mistakenly considering samples from the same camera as positive. Additionally, we select the hard negative proxies from the remaining ones to push apart. The resulting loss function is defined as follows:

$$L_{\text{CA}} = -\sum_{i=1}^{B} \frac{1}{|P|} \sum_{p \in P} \log \left(\frac{S(p, x_i)}{Z(x_i)} \right), \quad (6)$$

$$Z(x_i) = \sum_{u \in P} S(u, x_i) + \sum_{q \in Q} S(q, x_i), \quad (7)$$

where B represents the batch size, while P and Q denote the positive and hard negative sets, respectively.

In summary, this approach incorporates the inherent camera information into the training process to enhance the model's discriminative performance. Incorporating the camera ID during proxy selection allows the camera-aware loss to guide the model in acquiring more refined and generalized features, which in turn minimizes the noise from camera variance and boosts clustering performance.

3.3. Cosine Similarity Regularization

During local training, each client's model is optimized based on its own dataset, which may cause the trained models to move in different directions, leading to divergence from the global model. If this divergence becomes too large, the global model may struggle to effectively integrate knowledge from different clients, resulting in suboptimal performance across the entire federated learning process.

To mitigate this issue, we introduce a regularization loss that constrains the deviation between the global model and the local model during training, as described in Eq. 8. Specifically, during local training, we pass a fixed batch of data D_{batch} through the aggregated global model of $(t-1)^{th}$ round and the current local model in t^{th} round, then compare the resulting feature representations. The regularization term is designed to limit excessive feature shifts by penalizing large cosine distance changes, ensuring that local updates remain aligned with the global model's optimization trajectory.

$$L_{reg} = 1 - \text{ Similarity} \left(f_{w_g^{t-1}}(x_i), f_{w_k^t}(x_i) \right), \quad (8)$$

In summary, our proposed regularization loss ensures that local models remain close to the global model during training, making it more effective in maintaining consistency across clients. This method not only addresses the challenge of model divergence but also enhances the overall efficiency and reliability of the federated learning process. The complete loss function is defined as follows:

$$L_{all} = L_{CL} + \delta L_{reg}, \quad (9)$$

3.4. Identity-Distributed Equalization

Typically, effective clustering is the critical factor for the success of unsupervised learning. In person re-ID, clustering algorithms like DBSCAN are widely used to assign pseudo-labels before training. However, we observed that for many re-ID datasets, the initial pseudo-labels obtained from clustering often have a high error rate. This issue primarily stems from the choice of DBSCAN hyperparameters, which are crucial for computing pairwise distances, reranking [38], and determining the appropriate neighborhood radius for clustering.

Existing works [21, 32, 33, 37] typically adopt a fixed set of hyperparameters optimized for large-scale datasets such as Duke [43] and Market [44]. These hyperparameters are dataset-dependent and tuned based on the specific feature distributions of those two datasets. However, applying the same parameters to other re-ID datasets often leads to performance degradation, as different datasets exhibit diverse distribution characteristics. On the other hand, manually tuning hyperparameters for each dataset is prohibitively time-consuming and impractical.

To address this challenge, we propose an innovative preprocessing approach, Identity-Distributed Equalization (IDE), which adjusts the average number of images per person ID to align with large datasets like Duke and Market, formulated as:

$$N_{avg} = \frac{N_{\text{img}}}{N_{\text{pid}}}, \quad (10)$$

where N_{img} represents the total number of images in the training set, and N_{pid} denotes the number of person identities. By ensuring a more consistent distribution across datasets, IDE improves clustering stability and mitigates the performance degradation caused by fixed DBSCAN parameters. The calculated values of N_{avg} for eight datasets are presented in Tab. 1.

As mentioned above, IDE enlarges datasets by multiplying each image by λ, ensuring the image number per identity is comparable to datasets like Duke and Market (approximately 20). For example, when aligning iLIDS ($N_{\text{avg}} = 4$ images per identity) with Duke, λ is computed as $\lambda = \frac{20}{4} = 5$. This adjustment scales iLIDS to align with large-scale datasets, enabling effective federated training.

Moreover, IDE can easily adapt to various re-ID datasets, even without ground truth. It is effortless to obtain the number of images per pid since the timestamp of video, or cropping frequency are known during converting video to image datasets, which is a method commonly used in the Re-ID field [4]. Therefore, compared to the traditional preprocesses, the IDE is not only easier to implemented to improve the quality of data clustering, but also more appropriate to apply in various datasets.

Algorithm 1 Camera-aware unsupervised person Re-id

Server:
Initialize the server model with w_g
for *round* $t = 1, 2, \ldots$ **do**
  Receive local model parameters from each client $k \in K$
  Aggregate local models with Eq. 13
  Distribute updated global model parameters to each client $w_k^t \leftarrow w_g^{t-1}$

Client:
Initialize memory bank M with size $M \times d$ (where d is the feature dimension)
for *each round* $t = 1, 2, \ldots$ **do**
  for *each client* k ***in parallel*** **do**
    Download global model w_k^t
    Extract features for dataset D_k using w_k^t
    Apply DBSCAN to cluster features and assign pseudo labels $\hat{y}_i$ and get $\hat{z}_i$
    for *epoch* $e = 1, 2, \ldots$ **do**
      Compute local contrastive loss:

$$L_{cl} = \alpha L_{inter} + \beta L_{intra} + \gamma L_{CA} + \delta L_{reg}$$

      Update local model using Adam optimizer θ_k^t
      Update Memory Bank:

$$M[\hat{z}_i] \leftarrow \mu M[\hat{z}_i] + (1 - \mu) f_{w_k^t}(x_i)$$

  Calculate similarity between $f_{w_g^{t-1}}(x_i)$ and $f_{w_k^t}(x_i)$
  Upload updated model parameters w_k^t to the server

Table 1. Statistic of Datasets.

Datasets	Cam IDs	N_{img}	N_{pid}	$\frac{N_{\text{img}}}{N_{\text{pid}}}$	λ
Duke [43]	8	16522	702	23	1
Market [44]	6	12936	751	18	1
iLIDS [45]	2	248	59	4	5
CUHK03 [35]	2	7365	767	10	2
PRID [46]	2	3744	285	13	2
VIPeR [47]	2	632	316	2	10
CUHK01 [48]	2	1940	485	4	5
3DPeS [49]	2	450	93	5	4

3.5. Cosine-Distance Weighted Aggregation

During the aggregation stage, FedAvg [18] is the commonly used aggregation method in most existing studies, especially those focusing on unsupervised federated person re-ID [21, 32, 33, 37]. This approach determines the aggregation weights for merging local models into a global model by accounting for the relative sizes of the individual datasets. The parameters of the global model can be formulated as follows:

$$w^t = \frac{1}{K} \sum_{k=1}^{K} \frac{n_k}{n} w_k^t, \tag{11}$$

where, w_k^t represents the model parameters from the k-th client at the t-th round. The term n_k indicates the size of the dataset belonging to the k-th client, while n denotes the total data size.

We observed that the eight benchmark datasets utilized in our study exhibit significant variability in size. Consequently, the application of FedAvg results in diminished contributions from smaller datasets to the global model, especially with significant size discrepancies between datasets. Such outcomes do not align with the intended objectives of FL.

To address this issue, we propose Cosine-Distance Weighted Aggregation as an alternative method for calculating aggregation weight, leveraging cosine distance to measure the extent of knowledge gained by a local model in a training round. Specifically, before the local training begins, we pass a batch of data D_{batch} through the untrained local model, which is the aggregated global model of the last round, and store its output features ($f_{w_g^{t-1}}(D_{\text{batch}})$). After the training process is completed within this round, we obtain a new set of output features by passing the same D_{batch} through the updated local model ($f_{w_k^t}(D_{\text{batch}})$. The aggregation weight is then determined by computing the cosine distance between them, where a larger distance indicates that the model has acquired more new knowledge during this round, justifying a higher aggregation weight in the overall global model.It can be expressed mathematically as follows:

$$m_k^t = 1 - \frac{1}{S} \sum_{i=1}^{S} \text{Similarity}\left(f_{w_g^{t-1}}(x_i), f_{w_k^t}(x_i)\right), \tag{12}$$

where m_k^t denotes the aggregation weight for the k-th client during round t. The variable S denotes the number of images in D_{batch}, and D_{batch} is randomly selected from the training dataset at the start of each round. Ultimately, the aggregation equation for local models can be formulated the as follows:

$$w^t = \sum_{k=1}^{K} \frac{m_k^t}{m^t} w_k^t, \tag{13}$$

where $m^t = \sum_{k=1}^{K} m_k^t$. Utilizing this approach guarantees that all clients contribute fairly, regardless of the size of their datasets, ultimately resulting in a more efficient FL system.

Table 2. Comparison of the accuracy(%) with Existed Unsupervised Federated Person Re-ID.

Methods	Duke		Market		iLIDS		CUHK03		PRID		VIPeR		CUHK01		3DPeS	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
FedUReID [21]	51.0	-	65.2	-	73.5	-	8.9	-	38.0	-	26.6	-	43.6	-	65.5	-
FedUCC [37]	78.8	60.5	86.5	65.5	74.7	59.7	9.6	9.7	58.9	63.1	31.3	36.7	78.3	75.3	68.9	50.9
FedUCC [32]	-	-	90.3	75.2	**82.8**	72.0	38.7	35.5	69.0	72.0	43.0	48.6	80.1	76.6	73.2	57.8
FedUCA [33]	81.0	66.5	92.5	79.4	80.5	-	50.0	-	75.0	-	51.0	-	86.0	-	85.0	-
FedCAPR(Ours)	**83.6**	**69.6**	**92.7**	**82.5**	81.6	**77.8**	**73.4**	**69.7**	**85.0**	**88.4**	**68.7**	**75.6**	**95.6**	**95.1**	**87.4**	**79.2**

Table 3. Comparison of accuracy(%) to the approaches of domain adaptation and purely unsupervised methods.

Methods	Types	Duke		Market	
		R1	mAP	R1	mAP
NRMT [40]	Domain adaptation	77.8	62.2	87.8	71.7
MMT [3]	Domain adaptation	78.0	65.1	87.7	71.2
IICS [7]	Purely unsupervised	80.0	64.4	89.5	72.9
JVTC+ [41]	Purely unsupervised	80.4	66.5	86.8	67.2
MMCL [42]	Purely unsupervised	65.2	40.2	80.3	45.5
FedCAPR(Ours)	Federated purely unsupervised	**83.6**	**69.6**	**92.7**	**82.5**

Figure 3. **Sample images of eight datasets.** Example images were randomly selected from eight benchmark datasets.

4. Experiments

4.1. Setup

Datasets and Models. We follow [39] to utilize eight person re-ID benchmark datasets, each is treated as a distinct client, to evaluate the performance of our FedCAPR, where ResNet-50 [54] is the backbone architecture. Among these datasets, the two large-scale datasets, Duke [43] and Market[44], were captured with eight and six cameras, respectively. The remaining datasets, iLIDS [45], CUHK03 [35], PRID [46], VIPeR [48], CUHK01 [47], and 3DPeS [49], were recorded with only two cameras. Tab. 1 provides detailed statistics for these datasets. Notably, the datasets vary significantly in terms of location, image count, identities, and appearance diversity, as illustrated in Fig. 3. This diversity reflects real-world conditions, simulating the heterogeneity and imbalance inherent in federated learning.

Duke Issues. Although the Duke dataset was widely used in re-ID tasks, including supervised, unsupervised, video-based, occluded-based, and so on, it has been shut down due to privacy and human rights concerns. Despite this, most state-of-the-art methods [21, 32, 33, 37] rely on Duke for superior results. To ensure fair comparison, we temporarily adopt Duke in FedCAPR. We have replaced Duke with MSMT17 [52], a large-scale but noisier dataset, with results shown in the supplementary Tab. 9.

Evaluation Metrics. Re-identification (re-ID) retrieves images with the same ID from a gallery. Performance is evaluated using two metrics: Cumulative Matching Curve (CMC), measuring the likelihood of finding the correct match within the top k results, and Rank-1 accuracy, indicating if the first match is correct. Mean Average Precision (mAP) averages precision across queries, rewarding correct matches at various ranks even if the top result is incorrect. Both metrics assess immediate and overall retrieval quality.

Implementation Details. Our framework was implemented using EasyFL [34] and PyTorch [53], with training on NVIDIA GeForce RTX 3090 GPUs. We used ResNet-50 [54] (ImageNet pre-trained) as the backbone, omitting the fully connected layer to enhance clustering and contrastive learning. Feature representations were refined with L_2 regularization to improve camera-specific clusters. The federated setup included eight clients (one per dataset), with 20 communication rounds of five local epochs each. In the final round, local epochs increased to 50 for better personalization. Hyperparameters were set as follows: $\alpha = 0.6$ and $\beta = 0.7$ from FedUCA [33], and $\gamma = 0.7$ and $\delta = 0.25$ empirically tuned for optimal performance.

4.2. Comparisons with the State-of-the-Arts

In this section, we present a performance comparison between our FedCAPR and several state-of-the-art techniques, which are shown in Tab. 2 and Tab. 3.

The comparison with state-of-the-art works that combine unsupervised and federated learning shows that the results on eight benchmark datasets, as shown in Tab. 2, demonstrate that our method achieves superior performance across nearly all datasets. Notably, significant improvements are observed on smaller datasets. In comparison to FedUCC(Weng et al., 2023) [32], FedCAPR improves Rank-1 accuracy by 34.7%, 16%, and 15.5%, and mAP by 34.2%, 16.4%, and 18.5% on the CUHK03, PRID, and CUHK01 datasets, respectively.Moreover, FedCAPR outperforms FedUCA [33] on eight benchmark datasets, with improvements of 23.4%, 10%, 17%, and 9% in Rank-1 on CUHK03, PRID, VIPeR, and CUHK01, respectively.

Table 4. Ablation study of our proposed methods. The value in the table is Rank-1 accuracy (%).

Methods	Duke	Market	iLIDS	CUHK03	PRID	VIPeR	CUHK01	3DPeS
Baseline	81.0	92.5	77.6	50.0	75.0	50.3	86.2	82.3
$+L_{CL}$	83.1	92.6	78.6	63.7	77.0	50.0	88.3	82.9
+ CDW + $+L_{reg}$	83.4	**92.7**	80.6	66.9	**86.0**	53.8	89.2	86.5
+ IDE	**83.6**	**92.7**	**81.6**	**73.4**	85.0	**68.7**	**95.6**	**87.4**

Table 5. Comparison of V-measure (V) and Silhouette (S) scores with and without IDE.

Methods	DukeMTMC		Market-1501		iLIDS-VID		CUHK03		Prid2011		VIPeR		CUHK01		3DPeS	
	V	S	V	S	V	S	V	S	V	S	V	S	V	S	V	S
w/o IDE	**0.369**	**0.021**	**0.218**	**-0.074**	0.266	0.054	0.297	-0.094	0.865	0.548	0.385	0.022	0.234	-0.144	0.456	0.041
w/ IDE	**0.369**	**0.021**	**0.218**	**-0.074**	**0.848**	**0.881**	**0.71**	**0.312**	**0.865**	**0.548**	**0.929**	**0.957**	**0.912**	**0.878**	**0.83**	**0.863**

However, FedUCC (Weng et al., 2023) outperforms on iLIDS, the smallest dataset, by leveraging extensive inter-client information sharing (see Sec. 2.3). While our method shows slightly lower accuracy on iLIDS, it significantly reduces computational costs, offering a more efficient and practical alternative.

Also, Tab. 3 provides a performance comparison between our method and other existing unsupervised re-ID approaches, including domain adaptation and purely unsupervised methods [3, 7, 40–42], and demonstrates our method consistently achieves the best results. Compared to the works with domain adaption, FedCAPR improves Rank-1 accuracy by 5.6% and 5%, and mAP by 4.5% and 11.3% on the DukeMTMC and Market-1501 datasets, respectively.Moreover, FedCAPR outperforms purely unsupervised methods with improvements of 3.2% and 3% in Rank-1, and 3.1% and 9.6% in mAP on the DukeMTMC and Market-1501 datasets, respectively.

The superior performance of our method stems from the federated learning framework, which enables knowledge sharing across clients (i.e., datasets), enriching the training process. We focus on Duke and Market datasets, as smaller datasets are challenging to train without federated learning, and most unsupervised methods do not report results on them.

4.3. Ablation Study

In this section, we conduct an ablation study across eight datasets to assess the impact of our proposed methods: camera-aware loss, cosine similarity regularization, CDW aggregation, and IDE techniques, as presented in Tab. 4. In Tab. 4, camera-aware loss significantly improves accuracy on datasets like Duke and CUHK03, where camera variance is substantial, increasing Rank-1 accuracy from 50% to 63.7% on CUHK03. This demonstrates its effectiveness in mitigating the noise introduced by camera discrepancies.

Moreover, the improvements brought by the proposed methods in the federated learning framework are especially evident on smaller datasets, such as iLIDS and 3DPeS, or inherently challenging datasets like PRID. These datasets benefit from the novel techniques we introduced, such as CDW aggregation, which enables more effective information exchange between clients. Regularization loss, L_{reg}, further reduces the disparity among local models, ensuring that the aggregated model maintains robust and consistent performance. This allows for more effective training outcomes compared to standalone training, overcoming challenges posed by limited data.

Finally, IDE proves particularly beneficial for datasets like VIPeR, which have distributions differing from large-scale datasets. Tab. 5 shows initial clustering performance using V-measure (0 to 1, higher is better) and Silhouette score (-1 to 1, higher is better). IDE significantly improves clustering quality, reducing errors in the initial phase and ensuring reliable pseudo-labels for subsequent training. By aligning data distribution with DBSCAN parameters tuned on large datasets, IDE mitigates distribution differences, enhancing training outcomes.

5. Conclusion

In this paper, we introduced FedCAPR, a novel decentralized camera-aware unsupervised learning method within the federated learning (FL) framework. FedCAPR integrates a camera-aware loss function to address noise caused by camera variance; cosine-distance weighted aggregation enables to mitigate the information imbalance between large and small datasets; identity-distributed equalization ensures that each dataset can facilitate more effective and consistent training across heterogeneous datasets. Our FedCAPR achieves superior performance to state-of-the-art results, particularly up to 23.4% on smaller datasets. Hence, FedCAPR offers a highly secure, accurate and efficient solution for smart city surveillance systems.

References

[1] L. Wei, S. Zhang, W. Gao and Q. Tian, ”Person Transfer GAN to Bridge Domain Gap for Person Re-identification,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 79-88. 1, 2

[2] Y. Chen, X. Zhu and S. Gong, ”Instance-Guided Context Rendering for Cross-Domain Person Re-Identification,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 232-242.

[3] Y. Ge, D. Chen, and H. Li, “Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification,” in Proceedings of International Conference on Learning Representations (ICLR), 2020. 2, 7, 8

[4] J. Lv, W. Chen, Q. Li, and C. Yang, “Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 7948–7956. 2, 5

[5] J. Liu, Z. J. Zha, D. Chen, R. Hong, and M. Wang, “Adaptive transfer network for cross-domain person re-identification,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7202–7211. 1

[6] T. Isobe, D. Li, L. Tian, W. Chen, Y. Shan, and S. Wang, “Towards discriminative representation learning for unsupervised person re-identification,” in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2021, pp. 8526–8536. 1, 2

[7] S. Xuan and S. Zhang, “Intra-inter camera similarity for unsupervised person re-identification,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 1, 2, 7, 8

[8] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, “Dynamic label graph matching for unsupervised video re-identification,” in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5142–5150. 1, 2

[9] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li, “Unsupervised graph association for person re-identification,” in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019, pp. 8321–8330. 1, 2

[10] M. Li, X. Zhu, and S. Gong, “Unsupervised person re-identification by deep learning tracklet association,” in Proceedings of European Conference on Computer Vision (ECCV), 2018, pp. 737–753. 1, 2

[11] Y. Chen, X. Zhu, and S. Gong, “Deep association learning for unsupervised video person re-identification,” arXiv preprint arXiv:1808.07301, 2018. 1, 2

[12] M. Ester, H.-P. Kriegal, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI), 1996, pp. 226–231. 1, 2, 3

[13] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, “A bottom-up clustering approach to unsupervised person re-identification,” in Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI), 2019, pp. 8738–8745. 2

[14] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, “Unsupervised person re-identification by soft multilabel learning,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2148–2157. 1, 2

[15] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3733–3742.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning (ICML). PMLR, 2020, pp. 1597–1607.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9729–9738.

[18] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Artificial Intelligence and Statistics, 2017, pp. 1273–1282. 1, 2, 6

[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” Proceedings of Machine learning and systems, vol. 2, pp. 429–450, 2020. 2

[20] W. Zhuang, Y. Wen, W. Zhang, X. Gan, D. Yin, D. Zhou, S. Zhang, and S. Yi, “Performance optimization of federated person re-identification via benchmark analysis,” in Proceedings of 28th ACM Int. Conf. Multimedia (MM’20), 2020, pp. 955–963.

[21] W. Zhuand, Y. Wen, and S. Zhang, “Joint optimization in edge-cloud continuum for federated unsupervised person re-identification,” in 29th ACM Int. Conf. Multimedia (MM’21), 2021, p. 433–441. 5, 6, 7, 1, 2

[22] G. Wu and S. Gong, “Decentralised learning from independent multi-domain labels for person re-identification,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, 2021, pp. 2898–2906.

[23] S. Sun, G. Wu, and S. Gong, “Decentralised person re-identification with selective knowledge aggregation,” arXiv preprint arXiv:2110.11384, 2021.

[24] A. Fallah, A. Mokhtari, and A. Ozdaglar, “Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach,” Advances in neural information processing systems, vol. 33, pp. 3557–3568, 2020.

[25] C. T Dinh, N. Tran, and J. Nguyen, “Personalized federated learning with moreau envelopes,” Advances in neural information processing systems, vol. 33, pp. 21 394–21 405, 2020.

[26] T. Li, S. Hu, A. Beirami, and V. Smith, “Ditto: Fair and robust federated learning through personalization,” in International conference on machine learning. PMLR, 2021, pp. 6357–6368.

[27] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, “Personalized cross-silo federated learning on non-

iid data," in Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 9, 2021, pp. 7865–7873. 1

[28] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," in Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI), 2021, pp. 2764–27 772. 1, 2, 3, 4

[29] X.-C. Li, D.-C. Zhan, Y. Shao, B. Li, and S. Song, "Fedphp: Federated personalization with inherited private models," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2021, pp. 587–602. 1

[30] M. Zhang, K. Sapra, S. Fidler, S. Yeung, and J. M. Alvarez, "Personalized federated learning with first order model optimization," arXiv preprint arXiv:2012.08565, 2020.

[31] J. Luo and S. Wu, "Adapt to adaptation: Learning personalization for crosssilo federated learning," in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), vol. 2022. NIH Public Access, 2022, p. 2166. 1

[32] J. Weng, K. Hu, T. Yao, J. Wang, and Z. Wang, "Federated unsupervised cluster-contrastive learning for person reidentification: A coarse-to-fine approach," Computer Vision and Image Understanding (CVIU), 2023. 1, 2, 5, 6, 7

[33] J. Liu, W. Zhuang, Y. Wen, J. Huang, and W. Lin, "Optimizing federated unsupervised person re-identification via camera-aware clustering," in Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP), 2022, pp. 1–6. 1, 2, 5, 6, 7

[34] W. Zhuang, X. Gan, Y. Wen, and S. Zhang, "Easyfl: A low-code federated learning platform for dummies," pp. 13 740–13 754, 2022. 2, 7

[35] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 1, 6, 7

[36] H. Yu, W. Zheng, A. Yu, X. Guo, S. Gong, and J. Lai, "Unsupervised person re-identification by soft multilabel learning," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2143–2152.

[37] J.Weng, K. Hu, T. Yao, J.Wang, and Z.Wang, "Robust knowledge adaptation for federated unsupervised person reid," in International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2022, pp. 1–8. 5, 6, 7, 1, 2

[38] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 5

[39] W. Zhuang, Y. Wen, X. Zhang, X. Gan, D. Yin, D. Zhou, S. Zhang, and S. Yi, "Performance optimization of federated person re-identification via benchmark analysis," in Proceedings of the 28th ACM International Conference on Multimedia, 2022, pp. 955–963. 7

[40] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, and L. Shao, "Unsupervised domain adaptation with noise resistible mutual-training for person re-identification." in Proceedings of European Conference on Computer Vision (ECCV), 2021, pp. 4420–4428. 7, 8

[41] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in Proceedings of European Conference on Computer Vision (ECCV), 2020. 7

[42] D.Wang and S. Zhang, "Unsupervised person reidentification via multi-label classification." in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 7, 8

[43] Z. Zheng, L. Zhen, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro." in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017. 1, 5, 6, 7

[44] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015, p. 1116–1124. 1, 5, 6, 7

[45] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person reidentification by video ranking," in Proceedings of European Conference on Computer Vision (ECCV), 2014, p. 688–703. 1, 6, 7

[46] M. Hirzer, P. Roth, C. Beleznai, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in Proceedings of the Scandinavian Conference on Image Analysis (SCIA), 2011, p. 91–102. 1, 6, 7

[47] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in Proceedings of European Conference on Computer Vision (ECCV), 2008, p. 262–275. 1, 6, 7

[48] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in Proceedings of European Conference on Computer Vision (ECCV), 2008, p. 262–275. 1, 6, 7

[49] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in Joint ACM Workshop on Human Gesture and Behavior Understanding, 2011, p. 59–64. 1, 6, 7

[50] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 598–607.

[51] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456. 2

[52] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-identification," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 79-88. 7, 2

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer,James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch:An imperative style, high-performance deep learning library. NeurIPS, 2019 7

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. 7