2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Intra-Person Camera Adversarial for Intra-Camera Supervised Person Re-identification

Ruochen Tang[1] and Xun Gong[1,2,3,4,*]
[1]School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[2]Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 611756, China
[3]National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China
[4]Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory of Sichuan Province, Southwest Jiaotong University, Chengdu 611756, China

***Abstract*— Re-identification (ReID) of individuals across different cameras is a challenging task due to the high-quality large-scale datasets required for the model. Intra-camera supervised (ICS) person re-identification has been proposed to address the high cost of annotating large-scale datasets, but reducing the gap between camera domains remains a major challenge. Current approaches use intra- and inter-camera learning with contrastive learning performed separately in both phases. However, the effect of features from the same person under different cameras on the model during inter-camera learning, and the fine-grained characteristic where the same person can be classified into multiple classes based on the camera labels, still require further research. To address this issue, we propose a Camera-Based Contrastive Learning (CBCL) method that moves features away from their respective cameras and closer to other cameras to reduce domain gaps. We also introduce an Intra-Person Camera Adversarial (IPCA) loss that effectively utilizes fine-grained characteristics of person re-identification and improve IPCA by introducing camera labels to obtain IPCA_2 which achieves better model recognition performance than IPCA alone. Extensive experiments on multiple datasets demonstrate that our method outperforms existing methods and is comparable to fully-supervised methods.**

I. INTRODUCTION

Person re-identification (ReID) aims to identify individuals across multiple non-overlapping cameras. In recent years, there has been significant advancement in supervised person re-identification [26], [19], [40], [3], [28], [24], [18], [38], resulting in its increasing practical applications. However, with the widespread adoption of supervised person re-identification, researchers have found the data annotation process to be excessively cumbersome. In practical applications, in order to meet the delivery standards of the model, it is necessary to collect real surveillance data from the deployment area of the project and annotate the obtained large-scale data. However, annotating large-scale data for ReID is a laborious and time-consuming process that can significantly impede project progress. Consequently, researchers have shifted their attention to unsupervised and domain adaptation ReID [2], [6], [7], [9], [11], [27], [39], [17], [31], [35], [36], [37].

*Corresponding author: xgong@swjtu.edu.cn

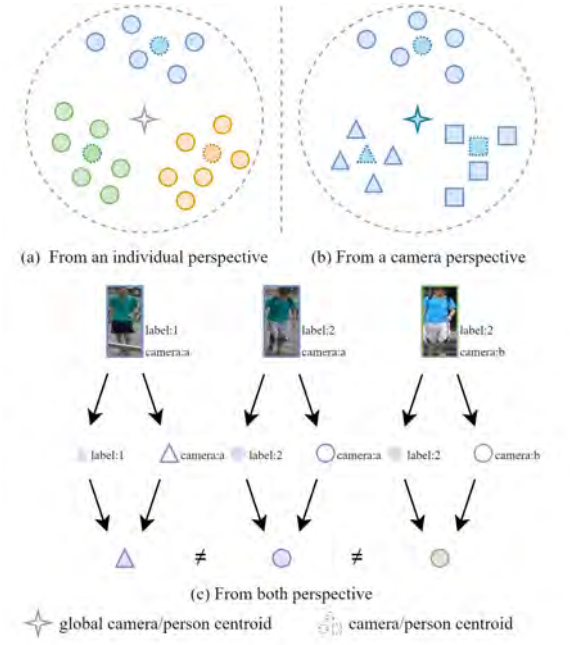

Fig. 1: (a). From an individual perspective, the global person centroid contains multiple camera centroids. (b). From a camera perspective, the global camera centroid also contains multiple person centroids. Different shapes represent different individuals, while the different colors represent different cameras. (c). From both perspective. The first image and the second image are captured by the same camera but from different persons, whereas the second and third images are from the same person but from different cameras, resulting in different colors of the clothes displayed in the two pictures. Thus they can be viewed as three different identities.

The former training directly on unlabeled data, while the latter pre-training the model on the source dataset and fine-tuning it on the unlabeled target dataset. However, both methods overlook the unique characteristics of data annotation in ReID.

The annotation of ReID data typically begins by annotating each camera's data separately and then further annotating the data from different cameras to assign the same label

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

to images of the same individual across different cameras. During this process, annotations within each camera can be obtained easily through tracking algorithms. Therefore, based on this characteristic, Intra-Camera Supervised (ICS) ReID has been proposed, which differs from unsupervised and domain adaptation methods in that it utilizes data that lacks cross-camera annotations.

The difficulty of Intra-Camera Supervised ReID lies in the fact that each camera, due to its lighting and viewing angles, can result in significant feature differences for the same individual across different cameras. Therefore, reducing the style differences between cameras has become the main focus of Intra-Camera Supervised (ICS) ReID. Current ICS methods utilize the labeled data within each camera for supervised training separately and then conduct inter-camera training using assigned pseudo-labels, which has proven effective [21], [41], [30], [32], [29], [15]. However, these methods only utilize the features from different cameras (named as camera centroid) of individuals without further investigate the impact of camera centroids of the same individual under different cameras on model, and overlook the fact that the same individual can actually be classified into multiple classes according to camera labels.

To address these limitations, we build our model on the intra-camera and inter-camera learning framework used in existing ICS while introducing a novel camera-based contrastive learning during inter-camera learning. Due to the nature of ICS problem, we can obtain camera centroids in other cameras with the same identity as the query during inter-camera learning. However, during intra-camera learning, the camera centroid of the same camera as the query already contains information about the query. Therefore, using it as a positive sample for contrastive learning may impede the model's ability to capture the intra-camera diversity of a person. To prevent overfitting to one specific camera, we only use the camera centroids with different camera labels from those of the query as positive samples during inter-camera learning.

At the same time, ReID can be regarded as a fine-grained problem. Based on the camera, there are multiple different person classes exist in the same camera. Based on the person, there are also multiple camera classes exist in the same person. In summary, as long as one of the person class and camera class of the two pictures is different, it can be regarded as a different class, as shown in Fig.1. Taking advantage of the fact that the same person can be classified into multiple classes based on camera labels, we designed an Intra-Person Camera Adversaria Loss that involves adversarial learning between different cameras for the same person. This encourages the backbone network to learn features that are invariant to the camera and specific to the person. Compared to traditional domain adversarial methods, this approach avoids interference from other indentities in the same camera and effectively eliminates camera style differences.

II. Related Work

A. Intra-Camera Supervised Person ReID

The mainstream method of existing Intra-Camera Supervised Person ReID is to divide the training into intra-camera learning and inter-camera learning. During inter-camera learning, previous methods mostly used simple triplet loss or parameter classifiers to learn from labeled data between cameras, such as PCSL [21], ACAN [22], MATE [41], but due to the limitations of triplet loss and parameter classifiers, the results are not ideal. Precise-ICS [30] proposes assigning a non-parametric classifier to each camera to address the issue of imbalanced numbers of images for different identities within each camera, and has achieved good results. This setting is also adopted by DCL [15] and GCL [32]. During inter-camera learning, SPCL [11] uses a soft label approach and employs triplet loss, while MATE [41] utilize a multi-task framework and employ multi-label learning on that framework. ACAN [22] eliminates style differences in cross-camera data by conducting adversarial learning between different cameras. However, since person re-identification is actually a fine-grained problem, only considering camera labels and conducting adversarial learning between images from different cameras of different pedestrians can eliminate camera style differences, but it can also affect the model's ability to distinguish person. Precise-ICS [30], on the other hand, proposes an effective ID association method based on the characteristics of the ICS problem, which greatly reduces the noise of pseudo-labels in the ICS problem and far exceeds previous methods in terms of mAP. We also use adversarial methods to make the model focus on camera-irrelevant features, but unlike ACAN, we introduce the pseudo-label association method in Precise-ICS [30] to obtain more accurate pseudo-labels and design an adversarial loss between the same person in different cameras.

B. Contrastive Learning

With the widespread application of contrastive learning [13], [4], [5] in various fields, it has also shown good results in person re-identification. Cluster Contrast [9] changes the memory dictionary in person re-identification contrastive learning from instance-level to cluster-level, and uses hard samples from each batch to update the dictionary, in order to solve the problem of different numbers of instances for each class in training data and pseudo-label noise. Wang et al. [30] and others used a contrastive learning method within and between cameras to effectively learn identity recognition capabilities within and across cameras. MCRN [34] created a Multi-Centroid Memory structure by dividing each class obtained by clustering into multiple subclasses, and then selecting positive and negative samples from these subclasses to remove noise from pseudo-labels and ensure model diversity. This paper starts from the characteristic that images of the same person come from multiple cameras, and selects positive and negative samples according to the camera to which the sample belongs on the basis of intra-

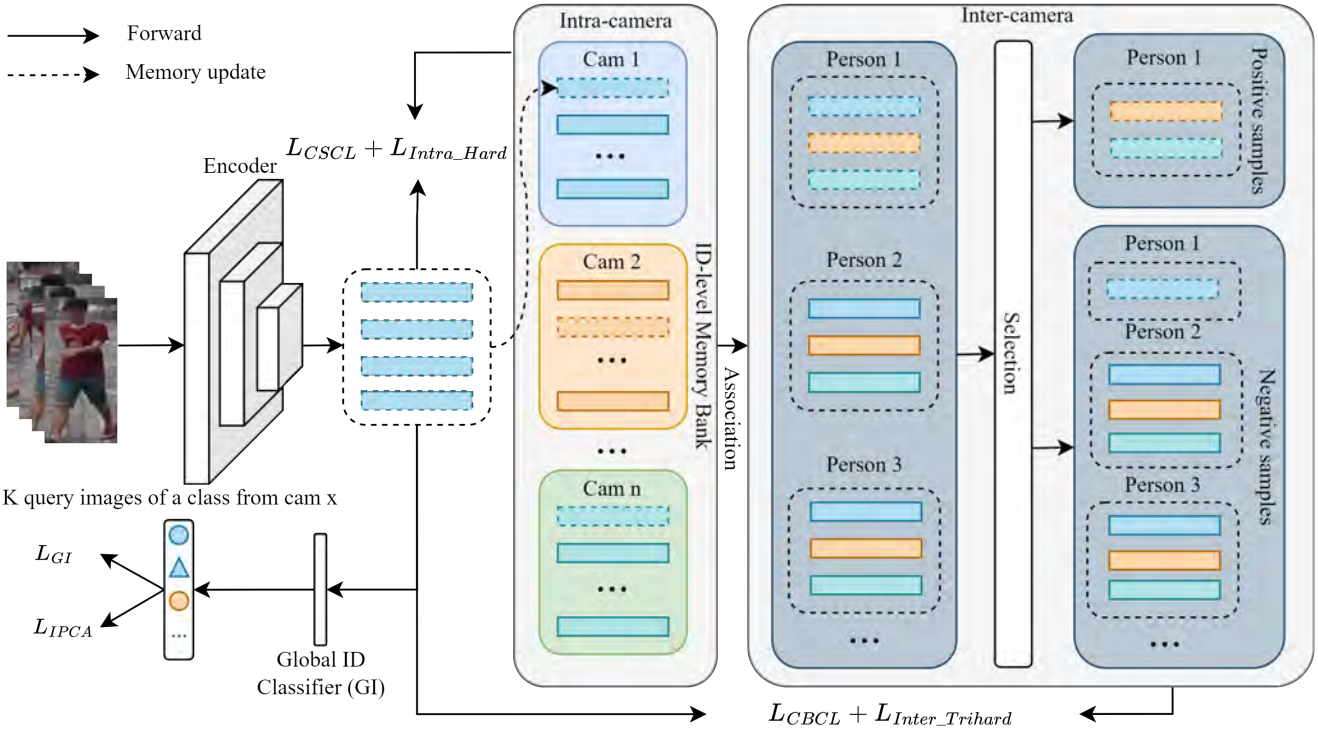

Fig. 2: The framework of the final model. The MCCN does not include the adversarial module and the selection module. Here, different colored blocks represent features of individuals from different cameras, while dashed blocks represent features of individuals who are the same as the query.

camera learning and inter-camera learning, avoiding overfitting within the same camera.

C. Adversarial Learning

Adversarial learning was first applied in GANs, and its ideas have been widely used in various fields. Domain adversarial learning has shown good results in eliminating style differences between domains, and has been widely applied in ReID, with some people applying it to eliminate style differences between cameras [22]. However, using traditional domain adversarial learning to eliminate camera styles can also affect the model's ability to distinguish pedestrians. Gu [12] and others proposed a clothing-based adversarial loss (CAL) to decouple clothing-independent features. This paper applies the idea of CAL to eliminate camera styles by employing adversarial training between images captured by different cameras of the same individual. This enables the model to extract camera-agnostic features and, compared to existing traditional domain adversarial methods, it also avoids reducing the model's pedestrian recognition capability.

III. Method

A. Multi-Camera Centroid Network

By combining the advantages of existing ICS models[21], [41], [30], [32], [29], [15], we have built a high-performance model as our baseline, which we refer to as Multi-Camera Centroid Network (MCCN). Unlike fully supervised ReID, the data set of intra-camera supervised ReID absent inter-camera annotations, only the same label under one camera can be guaranteed to be the same person. Therefore, we first look at each camera's data separately, assuming that there are C cameras in total, each has N_c images and Y_c person identities, then each camera's image can be represented as $D_c\{x_i, y_i, c_i\}_{i=1}^{Nc}$, where x_i, y_i, c_i represent the image ID, identity ID and camera ID of a person respectively. Then we integrate all the cameras together and set a global ID by adding up the identity IDs of each camera, $G\{1, 2, \ldots, N\}$ where $N = \sum_{i=1}^{C} N_i$. For convenience, we use q to represent a query in a batch, with its global ID denoted as g, C_g represents to the camera that the global ID g belong to.

1) Intra-camera Learning: Since there is a significant difference in the data between the cameras, the model can easily identify negative samples from different cameras. As a result, the model will learn the camera style difference, which is not conducive to eliminating the camera style difference. To solve this problem, each camera is regarded as a fully supervised task separately, and Camera-Specific Contrastive Learning is used for intra-camera learning. That is, each camera is assigned a memory bank, denoted as M_c. At the beginning of each epoch, the data in the memory bank is initialized by creating C tensors of shape $Y_c \times F$ based on the number of cameras C, where F represents the feature dimension. After that we initialize each row of each tensor separately to the average feature of all instance images of each class in each camera.

During training, each camera's memory bank is updated by an exponential moving average during backpropagation, as follows:

$$M_c[i] = mM_c[i] + (1-m)q[i], \tag{1}$$

where $i \in \{1, 2...Y_c\}$, m is the momentum coefficient, we set it to 0.1 to make the memory bank get more new features.

The intra-camera loss are the same as Precise-ICS [30], using the data of each camera individually for the supervised training of the model. The camera-specific contrastive learning (CSCL) loss function is as follows.

$$L_{CSCL} = -\sum_{c=1}^{C} \frac{1}{n_c} \sum_{q_c \in Q_c} log \frac{exp(\frac{1}{\tau_1} q_c \cdot M_c^+)}{\sum_{i=1}^{Y_c} exp(\frac{1}{\tau_1} q_c \cdot M_c[i])}, \tag{2}$$

where q_c represents the query from camera c. Each mini-batch consists of n encoded queries $Q = \{q_i|_{i=1}^{n}\}$, and can be further divided into multiple subsets $Q_c = \left\{q_c^i|_{i=1}^{n_c}\right\}$ based on the camera. M_c^+ is the positive centroid in camera c of the positive class and τ_1 indicates the temperature coefficient.

And the hard sample loss is used to enhance the recognition ability of the model. The loss function is as follows:

$$\begin{aligned} &L_{Intra\_Hard} = \\ &\sum_{c=1}^{C} \sum_{q_c \in Q_c} [\max d(q_c, p_c) - \min d(q_c, n_c) + m_1]_+ \\ &+ \ [d(q_c, M_c^+) - \min_{i \neq t} d(q_c, M_c[i]) + m_1]_+, \end{aligned} \tag{3}$$

where, $d(x, y)$ represents the cosine distance between x and y. p_c represents positive samples from camera c in the batch, while n_c represents negative samples from camera c in the batch. t is the label for the q_c.

The loss during intra-camera learning can be written as:

$$L_{Intra} = L_{CSCL} + L_{Intra\_Hard}. \tag{4}$$

2) Inter-camera Learning: In order for the model to learn the features under different cameras, inter-camera learning is also performed. Identity association should be performed first before inter-camera learning, because there is no cross-camera identity annotation in ICS. For cross-identity association, we also adopt the method in Precise-ics [30], but the difference is that we do the association in each epoch to facilitate the application of pseudo-labels for subsequent inter-camera learning.

In intra-camera learning, we have obtained the memory bank under different cameras. As mentioned before, there will be multiple different camera centroids in the same person. Therefore, for inter-camera learning, we concatenate the memory bank tensors of different cameras to obtain the tensor M of $N \times F$, and use the positive centroids of the query as positive samples, while negative centroids as negative samples for inter-camera contrastive learning (ICCL).

$$\begin{aligned} &L_{ICCL} = \\ &-\frac{1}{S} \sum_{i \in G^+} log \frac{exp(\frac{1}{\tau_1} q \cdot M[i])}{exp(\frac{1}{\tau_1} q \cdot M[i]) + \sum_{j \in G^-} exp(\frac{1}{\tau_1} q \cdot M[j])}, \end{aligned} \tag{5}$$

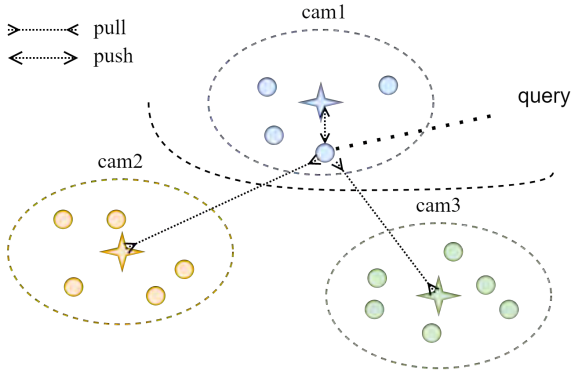

Fig. 3: Camera-Based Contrastive Learning. Where different colors represent different cameras, circular shapes represent the features of the images, and stars represent the camera centroids. Moving the query away from its original camera and closer to other cameras to reduce inter-camera gaps.

where G^+ represents the set of global IDs of the positive centroids of the query, S is the number of elements in G^+, and G^- represents the set of global IDs of the negative centroids of the query.

Similar to intra-camera learning, we also add a hard sample loss.

$$L_{Inter\_Trihard} = \sum_{q \in Q} [\max d(q, p) - \min d(q, n) + m_2]_+, \tag{6}$$

where, p and n respectively represent the positive and negative samples in the batch.

B. Camera-Based Contrastive Learning

The previous MCCN has achieved good results on ICS, but it still has some room for improvement in contrastive learning of inter-camera learning. As can be seen from the previous description, we obtained N_a camera centroids through intra-camera learning, and these camera centroids have different combinations of person and camera labels. Among them, camera centroids with the same pseudo-label will also have large differences, which is because they are the same person but come from different cameras. Inspired by MCRN [34], multiple camera centroids within the same group of people can also be regarded as multi-centroid. Since during intra-camera learning, the camera centroid from the same camera as the query already contains information of the query, there is less information to learn about intra-class diversity. Similar to MCRN [34], we propose a Camera-Based Contrastive Learning (CBCL), where we do not consider this camera centroid as a positive sample.

Due to the characteristics of ICS, we only consider the camera centroids that are not in the same camera as the query as positive samples. This method can also avoid overfitting the model within the same camera and make the query closer to positive samples from other cameras, as shown in Fig.3. Therefore, our inter-camera contrastive learning loss function is similar to ICE [2], but with the addition of negative sample

selection. It can be written as:

$$L_{CBCL} = -\frac{1}{S'} \sum_{i \in G^+, i \neq g} log \frac{exp(\frac{1}{\tau_1} q \cdot M[i])}{Z_{i,g}}, \tag{7}$$

$$\begin{aligned} Z_{i,g} &= exp(\frac{1}{\tau_1} q \cdot M[i]) + exp(\frac{1}{\tau_1} q \cdot M[g]) \\ &+ \sum_{j \in G^-} exp(\frac{1}{\tau_1} q \cdot M[j]), \end{aligned} \tag{8}$$

where g is the global ID of the query and S' is the count of global IDs in G^+ that are not equal to g.

C. Intra-Person Camera Adversarial

Based on the fine-grained characteristic of ReID, we inspired by CAL [12] and propose an adversarial method for learning camera-agnostic features across different camera views of the same person. This method can encourage the model to mine person-invariant features and prevent the interference of negative samples in traditional domain adversarial methods.

We first establish a classifier P with N classes, where each class corresponds to a global ID and we refer to it as the global ID (GI) classifier. We use global ID as the class because in ReID data, each image has two labels, one for the person and one for the camera, and we want the classifier to learn features for both the person and the camera. Therefore, as long as the labels for the person or the camera are different, they are classified into separate classes. In the global ID, the combination of person and camera labels in each class is different. Therefore, training the classifier with global ID as the class enables it to have both person and camera discrimination capabilities at the same time.

During the training phase of the global ID classifier, we apply $L2$ normalization to the model's output feature f, and then detach it before inputting it into the global ID classifier, we denote the output of the classifier as $GI(f)$. The purpose of detaching is to prevent the training of the classifier from affecting the model, as this classifier will classify images of the same person taken from different cameras into different classes (as shown in Fig.4), while our goal is to make the model output similar features for the same person across different cameras. We then use cross-entropy loss to train the GI classifier, with the loss function defined as $L_{GI}(GI(f), G)$, where L_{GI} represents the cross-entropy loss and G is the global ID. The loss can also be written as:

$$L_{GI} = -\frac{1}{n} \sum_{q \in Q} \log \frac{\exp\left(\frac{1}{\tau_2} q \cdot \varphi_g\right)}{\sum_{i=1}^{N_a} \exp\left(\frac{1}{\tau_2} q \cdot \varphi_i\right)}, \tag{9}$$

where q represents the query sample in the batch, g is the global ID of the query sample, φ denotes the parameters of the classifier.

Since the classifier is trained with the global ID as the label during the training phase, the classifier that we obtain can distinguish between different camera views of the same person. However, our goal is to make the features of the same person across different camera views output by the model similar, which means that this classifier should only distinguish between different persons, but not which camera the person is captured by. This creates an adversarial learning scenario between the model and the classifier, where the model tries to produce features that can mislead the classifier, while the classifier tries to accurately distinguish between different camera pictures of the same person. Due to the fact that the number of classes in the classifier is equal to the total number of global IDs, it includes the classes of the same person captured by different cameras. Unlike the training phase, during adversarial training, all classes that share the same pseudo-label as the query, i.e., the classes of the same person captured by different cameras, are considered as the positive class (as shown in Fig.4).

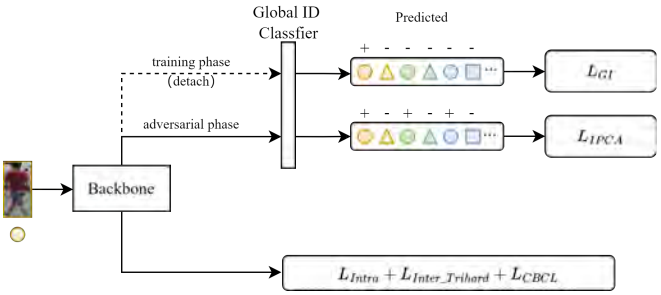

Fig. 4: The process of training a GI classifier and the adversarial relationship between the model and the classifier. The dashed line represents detaching the features before inputting them into the classifier during the training phase. Subsequently, the classifier outputs $q \cdot \varphi$ and uses the global ID as the positive class to train the classifier, enabling it to distinguish images of the same person captured by different cameras. During the adversarial phase, the features are directly inputted into the classifier and get $q \cdot \varphi$. Then all classes that have the same pseudo-label as the query are considered positive examples for training. This allows the model to deceive the classifier and establish an adversarial relationship. Where + indicates positive class and - indicates negative class.

While enhancing the model's recognition performance across different cameras is important, it is equally crucial to prevent the decrease of model's recognition ability under the same camera. To achieve this, we assign higher weights to the global ID class that query is in during the adversarial training, which serves a similar purpose as increasing the same clothing recognition capability in CAL, so we use the same coefficients α as CAL. The corresponding loss function, denoted as L_{IPCA}, is defined as follows:

$$L_{IPCA} = - \sum_{i \in G^+} \alpha_i \log \frac{\exp\left(\frac{1}{\tau_2} q \cdot \varphi_i\right)}{\exp\left(\frac{1}{\tau_2} q \cdot \varphi_i\right) + \sum_{j \in G^-} \exp\left(\frac{1}{\tau_2} q \cdot \varphi_j\right)}, \tag{10}$$

$$\alpha_i = \begin{cases} 1 - \epsilon + \frac{\epsilon}{S} & \text{if } i = g \\ \frac{\epsilon}{S} & \text{if } i \neq g \end{cases}, \tag{11}$$

where, g represents the global ID of the query.

To further address the previous issue, we added margin-based softmax [10] to the global ID class of the query to enhance its recognition ability. The GI classifier tends to classify the query into classes that have the same camera label as the query. This is due to the similarity in camera style among images captured by the same camera. Consequently, classes with the same camera label as the query have a higher probability than classes with different camera labels. Therefore, we introduced a weighting method in face recognition [33] to emphasize the classes of the same camera as the query in the adversarial process, in order to reduce the probabilities of the classifier for these classes. Since $q \cdot \varphi_i$ can also be written as $\cos(\theta_{q,\varphi_i})$, θ represents the angle between the classifier φ_i and the feature vector q, our final formula is:

$$L_{IPCA\_2} = - \sum_{i \in G^+} \alpha_i \log \frac{H_i}{H_i + \sum_{j \in G^-} \beta_j \exp\left(\frac{1}{\tau_2} \cos(\theta_{q,\varphi_j})\right)}, \tag{12}$$

$$H_i = \begin{cases} \exp\left(\frac{1}{\tau_2} \cos(\theta_{q,\varphi_i} + margin)\right) & \text{if } i = g \\ \exp\left(\frac{1}{\tau_2} \cos(\theta_{q,\varphi_i})\right) & \text{if } i \neq g \end{cases}, \tag{13}$$

$$\beta_j = \begin{cases} \exp\left(\frac{1}{\tau_2} t \left(\cos(\theta_{q,\varphi_j}) + 1\right)\right) & \text{if } C_j = C_g \\ 1 & \text{if } C_j \neq C_g \end{cases}, \tag{14}$$

where C_j is the camera label corresponding to the global ID j, and C_g is the camera label corresponding to the query. t is a hyperparameter.

1) Different with CAL: We propose an Intra-Person Camera Adversarial loss inspired by CAL [12]. Our approach differs from CAL in both purpose and formula: (1) CAL aims to extract clothing-irrelevant features, while our goal is to encourage the model to extract camera-irrelevant features. (2) We further improve upon our approach by introducing a weighting method in face recognition [33] to emphasize the classes that have the same camera label as the query.

Finally, integrating the losses in intra-camera and inter-camera, we can obtain the final loss of our model:

$$L = L_{Intra} + L_{Inter\_Trihard} + \frac{1}{n} \sum_{q \in Q} L_{CBCL} + \frac{1}{n} \sum_{q \in Q} L_{IPCA\_2}. \tag{15}$$

IV. Experiment

A. Dataset and Evaluation Metrics

We assessed the effectiveness of our approach on three large-scale datasets, Market1501, DukeMTMC-ReID, and MSMT17, using commonly used evaluation metrics for ReID, Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). To replicate ICS application scenarios, we generated intra-camera labels based on the complete annotations of the datasets and conducted our experiments using the newly generated labels.

TABLE I: Ablation experiments on the method we proposed.

Methods	Market1501		DukeMTMC-ReID		MSMT17	
	mAP	R1	mAP	R1	mAP	R1
MCCN	85.6	93.5	74.6	86.9	44.6	72.4
MCCN+CBCL	86.0	94.0	75.7	87.4	47.7	75.7
MCCN+IPCA	85.9	94.1	75.1	87.1	45.9	74.1
MCCN+IPCA_2	86.4	94.2	75.7	87.2	46.8	75.4
MCCN+CBCL+IPCA	87.2	94.5	76.5	88.1	48.8	77.0
MCCN+CBCL+IPCA_2	87.5	94.8	76.6	88.2	49.2	77.2

B. Implementation Details

We used a ResNet50 [14] pre-trained model on ImageNet [25] as our backbone. We replaced the sub-modules after the 4th layer with a GEM [23] pooling as the same in Cluster Contrast [9], followed by batch normalization [16] and L2 normalization.

Images are resized to 256×128. For training images, we perform random flipping, cropping, and erasing. We use the Adam optimizer to train the ReID model with a weight decay of 5e-4. The initial learning rate is set to 3.5e-4 for the first 10 epochs and a warm-up scheme is used. Then, the learning rate is lowered to 1/10 of its previous value every 20 epochs. We train the model for a total of 50 epochs on Market1501 and DukeMTMC-ReID, and 60 epochs on MSMT17. Similar to CAL [12], our L_{IPCA} is used for training after the 30th epoch. On Market1501, we set the batchsize to 64, with randomly selected $P = 8$ intra-identities and $K = 8$ images for each identity, and train for 408 iterations per epoch. On MSMT17 and DukeMTMC-ReID, the batchsize is set to 72, with $P = 12$ and $K = 6$, and we train for 800 iterations per epoch on MSMT17 and 400 iterations per epoch on DukeMTMC-ReID. Following Cluster Contrast [9] and CAL [12], we set the momentum value m to 0.1, and the temperature coefficient τ_1 and τ_2 to 0.05 and 0.0625, respectively. The $margin$ in Eq.13 is set to 0.7 for the Market1501 dataset, 0.3 for the DukeMTMC-ReID dataset, and 0.5 for the MSMT17 dataset. The t in Eq.14 is set to 0.2 in all datasets. At the same time, set m_1 in Eq.3 and m_2 in Eq.6 to 0.3. Our method is implemented with PyTorch. For the Market1501 dataset, we train the model on a RTX 3060Ti GPU. For the MSMT17 and DukeMTMC-ReID datasets, we train the model on a GTX 1080Ti GPU.

C. Comparison with the State-of-the-arts

In this section, we compare our model with state-of-the-art ReID models and it can be observed that in the Intra-camera supervised model, our model's performance on most evaluation metrics is higher than that of existing optimal models on Market1501, DukeMTMC-ReID, and MSMT17 datasets. Compared to the existing optimal model DCL on the Market1501 dataset, our model's mAP is 1.3% higher, while on the MSMT17 dataset, the mAP/R1 is 4.1%/2.7% higher. Compared to the optimal model CDL on the MSMT17 dataset, our model's mAP/R1 on the Market1501 dataset is 2.9%/0.8% higher, while on the MSMT17 dataset, the mAP/R1 is 1.2%/0.9% higher.

We also compared our method with unsupervised and unsupervised domain adaptation methods, such as unsupervised

TABLE II: Comparison with state-of-the-art methods.

Methods	Reference	Market1501				DukeMTMC-ReID				MSMT17			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Fully Unsupervised Methods													
MMCL [27]	CVPR'2020	80.3	89.4	92.3	45.5	72.4	82.9	85.0	51.4	35.4	44.8	49.8	11.2
SPCL [11]	NeurIPS'2020	88.1	95.1	97.0	73.1	82.9	90.1	92.5	68.8	42.3	55.6	61.2	19.1
Cluster Contrast [9]	ACCV'2021	92.9	97.2	98.0	83.0	-	-	-	-	62.0	71.8	76.7	33.0
ICE [2]	ICCV'2021	93.8	97.6	98.4	82.3	83.3	91.5	94.1	69.9	70.2	80.5	84.4	38.9
PPLR [6]	CVPR'2022	94.3	97.8	98.6	84.4	-	-	-	-	73.3	83.5	86.5	42.2
ISE [37]	CVPR'2022	94.3	98.0	98.8	85.3	-	-	-	-	67.6	77.5	81.0	37.0
Unsupervised Domain Adaptation Methods													
AD-Cluster [36]	CVPR'2020	86.7	94.4	96.5	68.3	72.6	82.5	85.5	54.1	-	-	-	-
SPCL [11]	NeurIPS'2020	90.3	96.2	97.7	76.7	82.9	90.1	92.5	68.8	53.7	65.0	69.8	26.8
GLT [39]	CVPR'2021	92.2	96.5	97.8	79.5	82.0	90.2	92.8	69.2	59.5	70.1	74.2	27.7
IDM [7]	ICCV'2021	93.2	97.5	98.1	82.8	84.6	92.2	94.1	71.9	63.6	75.5	80.2	35.4
AdaDC [17]	TCSVT'2022	92.9	97.5	98.5	83.2	82.3	91.6	94.4	71.4	60.7	73.6	78.7	32.7
IDM++ [8]	arXiv'2022	94.2	97.7	98.5	85.3	84.6	92.2	94.1	73.6	69.5	80.3	84.0	40.5
Fully Supervised Methods													
PCB [26]	ECCV'2018	93.8	-	-	81.6	83.3	-	-	69.2	68.2	-	-	40.4
BoT [19]	TMM'2020	94.4	-	-	86.1	86.4	-	-	76.4	74.1	-	-	50.2
OSNet [40]	ICCV'2019	94.8	-	-	84.9	88.6	-	-	73.5	78.7	-	-	52.9
ABD-Net [3]	ICCV'2019	95.6	-	-	88.3	89.0	-	-	78.6	82.3	90.6	-	60.8
Intra-camera Supervised Methods													
PCSL [21]	TCSVT'2020	87.0	94.8	96.6	69.4	71.7	84.7	88.2	53.5	48.3	62.8	68.6	20.7
ACAN [22]	TCSVT'2021	73.3	87.6	91.8	50.6	67.6	81.2	85.2	45.1	33.0	48.0	54.7	12.6
MATE [41]	IJCV'2021	88.7	-	97.1	71.1	76.9	-	89.6	56.6	46.0	-	65.3	19.1
Precise-ICS [30]	WACV'2021	93.1	97.8	98.6	83.6	83.6	92.6	94.7	72.0	57.7	71.1	76.3	31.3
GCL [32]	Access'2021	93.6	97.5	98.4	85.0	86.4	93.7	**95.5**	75.1	73.4	84.0	87.3	45.6
PIRID [29]	ICASSP'2022	91.0	96.7	97.9	79.6	79.6	88.6	91.4	65.4	60.6	73.8	79.3	34.9
CDL [20]	TMM'2022	94.0	97.8	98.6	84.6	-	-	-	-	76.3	86.2	89.0	48.0
DCL [15]	ICPR'2022	**95.1**	98.0	**98.8**	86.2	-	-	-	-	74.5	84.5	87.6	45.1
ours	-	94.8	**98.1**	**98.8**	**87.5**	**88.2**	**94.2**	**95.5**	**76.6**	**77.2**	**86.6**	**89.3**	**49.2**

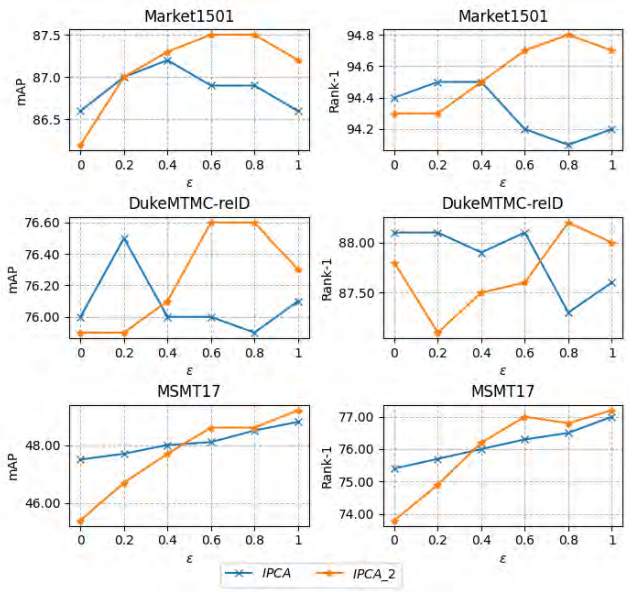

Fig. 5: Analysis of the parameter ϵ in IPCA and IPCA_2.

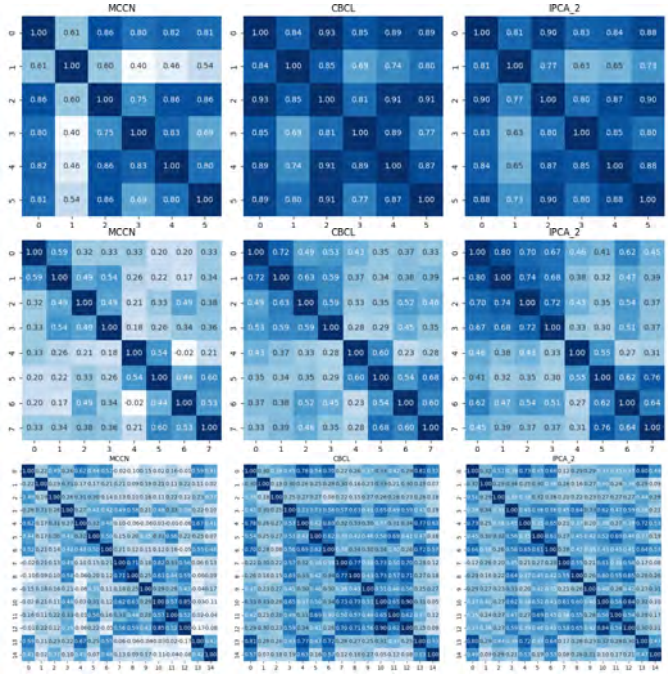

Fig. 6: The pairwise cosine similarity between different cameras in the Market1501, DukeMTMC-ReID, and MSMT17.

methods: MMCL [27], SPCL [11], Cluster Contrast [9], ICE [2], PPLR [6], ISE [37], and unsupervised domain adaptation methods: AD-Cluster [36], SPCL [11], GLT [39], IDM [7], AdaDC [17]. Our method is also superior to them.

Compared to fully supervised methods such as PCB [26], BOT [19], OSNet [40], and ABD-Net [3], our method still has certain competitiveness and outperforms PCB.

D. Ablation Studies

In order to evaluate the effectiveness of each method, we perform ablation experiments on three datasets. As shown in Table I, our baseline model MCCN has achieved high-performance. Furthermore, with the addition of CBCL and IPCA_2, the mAP/R1 improved by 1.9%/1.3% on Market1501, 2.0%/1.3% on DukeMTMC-ReID, and 4.6%/4.8% on MSMT17.

1) Effectiveness of CBCL: We validated the effectiveness of the CBCL method by adding it to both MCCN

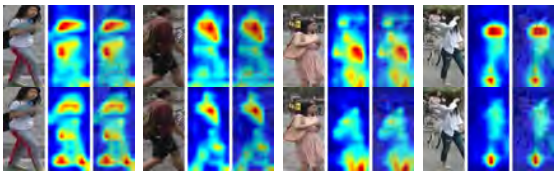

Fig. 7: The first row of the figure shows the activation maps outputted by the MCCN, and the second row shows the activation maps outputted by our final model

and MCCN+IPCA, as shown in Table I, when CBCL was added to MCCN, the mAP/R1 improved by 0.4%/0.5%, 1.1%/0.5%, and 3.1%/3.3% on Market1501, DukeMTMC-ReID, and MSMT17, respectively. In MCCN+IPCA, the addition of CBCL resulted in an improvement of mAP/R1 by 1.3%/0.4%, 1.4%/1.0%, and 2.9%/2.9% on the three datasets. Furthermore, in MCCN+IPCA_2, the addition of CBCL led to an improvement of mAP/R1 by 1.1%/0.6%, 0.9%/1.0%, and 2.4%/1.8% on the respective datasets.

2) Effectiveness of IPCA: As shown in Table I, we can see that IPCA shows improvements in both MCCN and MCCN+CBCL. After adding IPCA to MCCN, mAP/R1 is improved by 0.3%/0.6%, 0.5%/0.2%, and 1.3%/1.7% on the three datasets respectively. And after adding IPCA to MCCN+CBCL, mAP/R1 is improved by 1.2%/0.5%, 0.8%/0.7%, and 1.1%/1.3% on the three datasets respectively. At the same time, it can be seen that IPCA_2 is better than IPCA in all cases. Using IPCA_2 on MCCN, mAP/R1 is improved by 0.5%/0.1%, 0.6%/0.1%, and 0.9%/1.3% more than IPCA on the three datasets respectively. Using IPCA_2 on MCCN+CBCL, mAP/R1 is improved by 0.3%/0.3%, 0.1%/0.1%, and 0.4% 0.2% more than IPCA on the three datasets respectively.

3) The influence of ϵ in IPCA: In this section, we discussed the important parameter ϵ in the IPCA and IPCA_2 methods, which is mainly used to balance the model's cross-camera recognition ability and recognition ability within the same camera. When ϵ is larger, the model focuses more on cross-camera recognition ability. From the graph, it can be seen that IPCA_2 has higher performance than IPCA, and its optimal parameter for ϵ on Market1501, DukeMTMC-ReID, and MSMT17 is 0.8, 0.8, and 1, respectively.

E. Qualitative Analysis

1) Visualization of Cosine Similarity between Cameras: Following Bai et al. [1], we compared the similarity between camera domains in each dataset to further verify the role of CBCL in reducing the gap between camera domains. We used the cosine similarity between the average features of each camera in the dataset as a measure of similarity. From Fig.6, it can be seen that CBCL effectively reduces the distance between multiple camera domains and reduces the average distance (which means increasing their similarity), thus proving the effectiveness of CBCL in reducing the gap between camera domains. Similarly, after incorporating IPCA_2, the cosine similarity between most cameras has improved, and the average similarity has also increased. Therefore, this also proves the effectiveness of the IPCA_2 method.

2) Visualization of Feature Maps: In order to better understand the impact of incorporating two methods on the model, we visualized the activation maps of the model before and after adding the two methods in Fig.7. From the maps, we can see that our method pays more attention to the pedestrians themselves, effectively eliminating the interference of the background environment, and the model's attention is mainly focused on the target pedestrian. For example, as shown in the second and third columns, we can see that the final model is more accurate in recognizing the outline of pedestrians, compared to the MCCN, it can accurately separate pedestrians from complex backgrounds, and after adding the two methods, it can remove the irrelevant background information that the MCCN pays attention to.

V. Conclusion

In this study, we developed a high-performance ICS ReID model called MCCN based on existing ICS models. Based on this, we propose a Camera-Based Contrastive Learning method that further improves the model. During inter-camera learning, this method pushes each sample away from its own camera and pulls them closer to positive samples from other cameras. This helps prevent overfitting within the same camera during the learning phase. We also propose a method that employs adversarial learning between features of the same individual captured by different cameras. This method reduces the domain gap between cameras and avoids the impact of traditional domain adversarial approaches on recognition performance. The experimental results on the three datasets confirm the effectiveness of each of our methods, and the final performance outperforms the existing state-of-the-art methods.

While our method demonstrates significant advantages in our experiments, there are still certain potential limitations. The scalability of our method may be challenged when handling datasets containing a large number of cameras. Future research can focus on addressing these aspects to improve the applicability and robustness of the method.

VI. Acknowledgments

This work is partially supported by National Natural Science Foundation of China (62376231), Tangshan Basic Research Science and Technology Program (23130230E), Sichuan Science and Technology Program (24NSFSC1070,2023YFG0267), Science and technology research and development program of China National Railway Group Co., LTD (K2023T003).

References

[1] Z. Bai, Z. Wang, J. Wang, D. Hu, and E. Ding. Unsupervised multi-source domain adaptation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2021.

[2] H. Chen, B. Lagadec, and F. Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14960–14969, 2021.

[3] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8351–8361, 2019.

[4] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[5] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[6] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7308–7318, 2022.

[7] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L.-Y. Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11864–11874, 2021.

[8] Y. Dai, Y. Sun, J. Liu, Z. Tong, Y. Yang, and L.-Y. Duan. Bridging the source-to-target gap for cross-domain person re-identification with intermediate domains. *arXiv preprint arXiv:2203.01682*, 2022.

[9] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, pages 1142–1160, 2022.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[11] Y. Ge, F. Zhu, D. Chen, R. Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020.

[12] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] S. Hu, X. Zhang, and X. Xie. Decoupled contrastive learning for intra-camera supervised person re-identification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2628–2665. IEEE, 2022.

[16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[17] S. Li, M. Yuan, J. Chen, and Z. Hu. Adadc: Adaptive deep clustering for unsupervised domain adaptation in person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3825–3838, 2021.

[18] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[19] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.

[20] Y.-X. Peng, J. Jiao, X. Feng, and W.-S. Zheng. Consistent discrepancy learning for intra-camera supervised person re-identification. *IEEE Transactions on Multimedia*, 2022.

[21] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao. Progressive cross-camera soft-label learning for semi-supervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2815–2829, 2020.

[22] L. Qi, L. Wang, J. Huo, Y. Shi, X. Geng, and Y. Gao. Adversarial camera alignment network for unsupervised cross-camera person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2921–2936, 2021.

[23] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[24] M. Ren, L. He, X. Liao, W. Liu, Y. Wang, and T. Tan. Learning instance-level spatial-temporal patterns for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14930–14939, 2021.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[26] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[27] D. Wang and S. Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10981–10990, 2020.

[28] G. Wang, J. Lai, P. Huang, and X. Xie. Spatial-temporal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8933–8940, 2019.

[29] L. Wang, W. Zhang, D. Wu, P. Hong, and B. Li. Prototype-based inter-camera learning for person re-identification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4778–4782. IEEE, 2022.

[30] M. Wang, B. Lai, H. Chen, J. Huang, X. Gong, and X.-S. Hua. Towards precise intra-camera supervised person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3229–3238, 2021.

[31] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2764–2772, 2021.

[32] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua. Graph-induced contrastive learning for intra-camera supervised person re-identification. *IEEE Access*, 9:20850–20860, 2021.

[33] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248, 2020.

[34] Y. Wu, T. Huang, H. Yao, C. Zhang, Y. Shao, C. Han, C. Gao, and N. Sang. Multi-centroid representation network for domain adaptive person re-id. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2750–2758, 2022.

[35] S. Xuan and S. Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11926–11935, 2021.

[36] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9021–9030, 2020.

[37] X. Zhang, D. Li, Z. Wang, J. Wang, E. Ding, J. Q. Shi, Z. Zhang, and J. Wang. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7369–7378, 2022.

[38] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 667–676, 2019.

[39] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5310–5319, 2021.

[40] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019.

[41] X. Zhu, X. Zhu, M. Li, P. Morerio, V. Murino, and S. Gong. Intra-camera supervised person re-identification. *International journal of computer vision*, 129:1580–1595, 2021.