PROTOTYPE-BASED INTER-CAMERA LEARNING FOR PERSON RE-IDENTIFICATION

Lin Wang[1 2] *Wanqian Zhang*[1*] *Dayan Wu*[1] *Pingting Hong*[1 2] *Bo Li*[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{wanglin5812,zhangwanqian,wudayan,hongpingting,libo}@iie.ac.cn

ABSTRACT

Person re-identification (ReID) aims at retrieving images of the same person across non-overlapping camera views. The prior works focus on either fully supervised or unsupervised ReID settings, and achieve remarkable performances. In real scenarios, however, the major annotation cost comes from matching identity classes across camera views, thus leading to the Intra-Camera Supervised (ICS) ReID problem. In this work, we propose a Prototype-based Inter-camera ReID (PIRID) method, which tackles the ICS setting through the lens of prototype learning. Specifically, we first introduce the intra-camera learning with non-parametric classifiers to separately generate discriminative features within each camera view. Moreover, the inter-camera prototype learning provides prototypes as the representatives of each class in the common space, making the learned features to be camera-agnostic. Experiments conducted on three benchmarks, i.e., Market-1501, DukeMTMC-ReID, and MSMT17, show the superiority of our method.

Index Terms— person re-identification, intra-camera supervised, prototype learning

1. INTRODUCTION

Person Re-Identification (ReID) tries to return the same pedestrian from a large collection of person images in cross cameras views. ReID has aroused extensive research interest due to its importance in surveillance and public safety. The dominant paradigms mainly focus on fully supervised and unsupervised ReID settings, of which the key difference lies in the access of manual annotations (person identity labels) during the training procedure. Prior works [1, 2, 3, 4, 5, 6] have achieved remarkable performance in both settings respectively.

It is reasonable to believe, however, that for many problems of interest, these two settings are somewhat *'absolute'*. To be specific, in fully supervised setting, a large number of complete annotations are required, which are expensive and time-consuming to obtain. In unsupervised setting, inevitable performance degradation is introduced due to the lack of supervised information. These methods thus greatly reduce the availability and scalability in real-world applications and large-scale deployments.

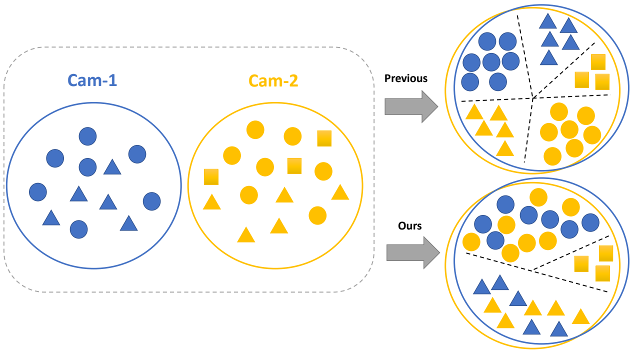


Fig. 1. An illustration of previous methods and our method in ICS ReID task. Training classifiers seperately within single camera will implicitly lead to the reduction of training samples of each inter-camera ID. (best viewed in color).

In this work, we therefore consider instead the less restrictive formulation in which person identity labels are only annotated within each camera, commonly referred to as Intra-Camera Supervised (ICS) ReID task. An early exploration of this setting is introduced by [7], which proposes a more scalable ReID framework with cheaper annotated training data, and pays more attention to exploiting the correspondence between specific identity spaces of cameras. MTML [7] designs a multi branch network which composed of shared feature extraction backbone and multiple classification branches. PCSL [8] adopts the triplet loss within each batch in the intra-camera learning.

Despite the thrilling success achieved by these methods, to date, much of the work in this area has ignored the following issue. As shown in Fig. 1, previous ICS methods focus on generating discriminative features within each camera view. However, we argue that training classifiers seperately within single camera will implicitly lead to the reduction of training samples of each inter-camera ID. To tackle this challenge, in this paper, we propose a Prototype-based Inter-camera ReID (PIRID) method, which tackles the ICS setting through the lens of prototype learning. Specifically, we first introduce the intra-camera learning with non-parametric classifiers to seperately generate discriminative features within each cam-

[*] Corresponding author

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

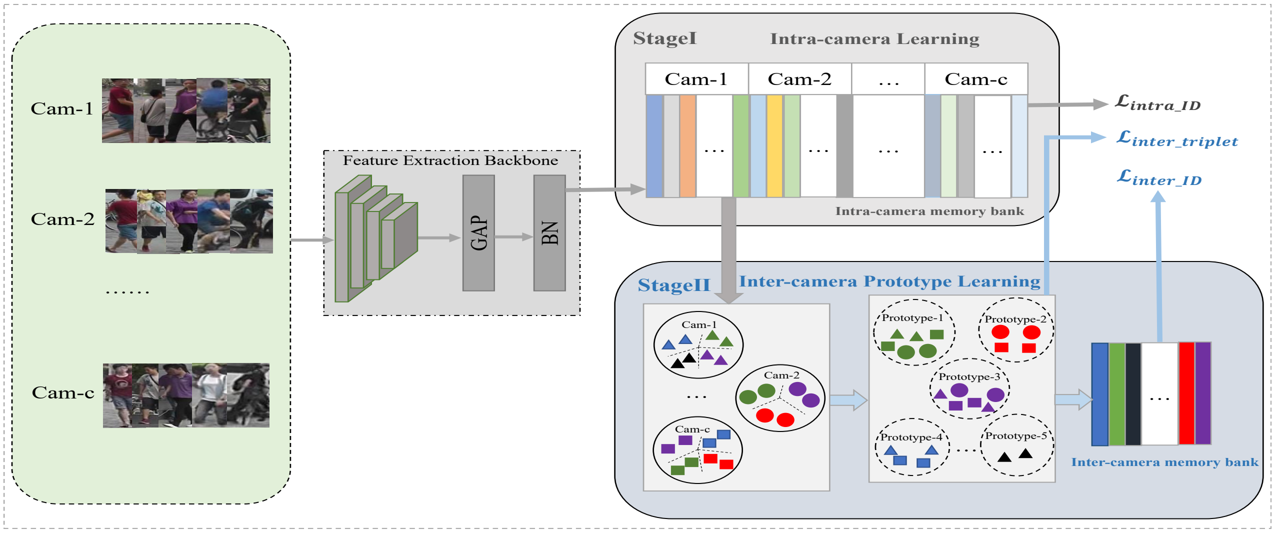


Fig. 2. Schematic diagram of PIRID framework (best viewed in color). Images are fed into ResNet-50 to obtain deep features. The intra-camera learning generates discriminative features within each camera view with non-parametric classifier. Moreover, the inter-camera prototype learning further bridges cross-camera associations for camera-agnostic features.

era view. To address the reduction of training samples within inter-camera IDs, we adopt prototype learning [9, 10] to bridge cross-camera associations and thus generate camera-agnostic features. Prototypes can be viewed as the representatives of each class in the common space, thus it provides valuable clues to discriminate ID information and guide the unsupervised inter-camera learning.

The main contributions of our work are summarized as follows:

- We study the Intra-Camera Supervised (ICS) ReID problem and propose a novel method names Prototype-based Inter-camera ReID (PIRID), which tackles the ICS setting through the lens of prototype learning.
- The intra-camera learning is introduced to generate discriminative features within each camera view, while the inter-camera prototype learning further bridges cross-camera associations, making the learned features to be camera-agnostic.
- Competitive results on three benchmark datasets, i.e., Market-1501, DukeMTMC-ReID, and MSMT17, demonstrate the superiority of PIRID.

2. PROPOSED METHOD

2.1. Problem Formulation

Suppose that there are C cameras in a dataset. We denote the set of the c-th camera by $\mathcal{D}_c = \{(x_i, y_i, c_i)\}_{i=1}^{M_c}$, where image x_i is annotated with an identity label $y_i \in \{1, ..., N_c\}$ and a camera label $c_i \in \{1, ..., C\}$. M_c and N_c are, respectively, the number of total images and IDs in c-th camera view. $N = \sum_{c=1}^{C} N_c$ is the total ID number directly accumulated in all cameras. ICS assumes that identification labels are annotated independently within each camera view and does not provide any identity associations between cameras. Next, we illustrate how to combine the supervised intra-camera learning and the unsupervised inter-camera learning.

2.2. Intra-camera Learning

The overall framework is illustrated in Fig. 2. Given the training set $\mathcal{D} = \bigcup_{c=1}^{C} \mathcal{D}_c$, our goal is to learn a ReID model that can well distinguish the identities within both intra- and inter-camera. We utilize the ResNet-50 [11] pre-trained on ImageNet [12] for feature extraction. Besides, we replace the fully-connected classification layer with the Global Average Pooling (GAP) layer followed by a Batch Normalization (BN) layer. Given an input image x_i, we can obtain the GAP feature $g(x_i)$ and the BN feature $f(x_i)$, of which the dimension is 2048. Inspired by [13] and [14], we adopt a number of non-parametric classifiers with memory bank to effectively guide the intra-camera learning. Each non-parametric classifier optimizes an ID classification loss, pulling the image closer to the centroid of its ID and vice versa. Specifically, we construct a intra-camera memory bank $\mathcal{K} \in \mathcal{R}^{d \times N}$ to store all accumulated IDs, where each column corresponds to each ID. The memory bank can be easily updated as:

$$\mathcal{K}[j] \leftarrow \mu\mathcal{K}[j] + (1 - \mu)f(x_i), \tag{1}$$

where $\mathcal{K}[j]$ is the j-th column of the memory bank. $f(x_i)$ denotes the feature of image x_i that belongs to the j-th ID. Given the dataset $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{N}$, all images within each camera are assigned with a global ID $j = A + y_i$, where $A = \sum_{k=1}^{c_i-1} N_k$ is the total ID number of previous $(c_i - 1)$-th camera view. The memory bank updating rate is $\mu \in [0, 1]$. After each iteration, the updated features in each column can be interpreted as the centroids of the identification classes in the feature space. Thus, the non-parametric softmax function can be defined as:

$$p(j|x_i) = \frac{exp(\mathcal{K}[j]^T f(x_i)/\tau)}{\sum_{k=A+1}^{A+N_{c_i}} exp(\mathcal{K}[k]^T f(x_i)/\tau)}, \tag{2}$$

where τ is a temperature factor.

Different from fully-connected layers, every non-parametric classifier is responsible for the classification task in each specific camera, which greatly reduces the number of parameters to be updated. Thus, the intra-camera ID loss can be defined as:

$$\mathcal{L}_{intra\_ID} = -\sum_{c=1}^{C} (\frac{1}{|\mathcal{D}_c|} \sum_{(x_i,y_i,c_i)\in\mathcal{D}_c} logp(j|x_i)), \tag{3}$$

where $\frac{1}{|\mathcal{D}_c|}$ is a standardized coefficient for balancing the variance of images in different cameras.

2.3. Inter-camera Prototype Learning

The intra-camera learning with non-parametric classifiers can generate discriminative features within each camera view. However, we argue that training classifiers seperately within single camera will implicitly lead to the reduction of training samples of each inter-camera ID. To address this problem, we adopt prototype learning [9, 10] to bridge cross-camera associations. Prototypes can be viewed as the representatives of each class in the common space, thus it provides valuable clues to discriminate ID information. We thereby take prototypes as centroids to provide the guidance for the unsupervised inter-camera learning.

To be specific, we adopt DBSCAN [15] to cluster the prototypes of IDs, and select reliable clusters by ruling out isolated points. All centroids of IDs within each cluster are assigned with the same prototype. In this way, we can obtain inter-camera prototypes of IDs $\mathcal{Q} = \{\mathcal{K}'[i'], \mathcal{S}'[i']\}_{i'=1}^{N'}$, where $\mathcal{K}'[i']$ is the inter-camera memory bank, $\mathcal{S}'[i'] \in \{1, ..., Y\}$ is generated prototypes and Y is the prototype number. Since each image belongs to the corresponding prototype, in a minibatch B,we can adopt the non-parametric softmax loss inspired by fully supervised methods, which can be defined as:

$$\mathcal{L}_{inter\_ID} = -\sum_{b=1}^{B}\sum_{i'=1}^{N'} log \frac{exp(\mathcal{K}'[S'[i']]^T f(x_b)/\tau)}{\sum_{j'=1}^{Y} exp(\mathcal{K}'[j']^T f(x_b)/\tau)}, \tag{4}$$

where j' is the number of images contained in the selected clusters. Therefore, we can obtain discriminative inter-camera classifiers via pulling an instance close to the prototype of its class while pushing away from the prototypes of all other classes.

Despite considering the distance between the sample and the corresponding prototype, ID loss may fail for the hard sample scenario. To that end, we further introduce the batch-hard triplet loss [16] $\mathcal{L}_{inter\_triplet}$, which is applied to the features $g(x)$ output by the GAP layer and can be formulated as:

$$\begin{aligned}\mathcal{L}_{inter\_triplet} = \frac{1}{N'} \sum_{\substack{i,j=1\\ i\neq j}}^{N'} [m_1 + max||g({x_a}^i) - g({x_p}^i)|| \\ - min||g({x_a}^i) - g({x_n}^j)|||].\end{aligned} \tag{5}$$

Thus, the total loss for inter-camera prototype learning is defined as:

$$\mathcal{L}_{inter} = \mathcal{L}_{inter\_ID} + \mathcal{L}_{inter\_triplet}. \tag{6}$$

2.4. Learning Algorithm

The proposed method combines intra- and inter-camera learning together with a two-stage learning procedure. In summary, the overall loss function for model learning is:

$$\mathcal{L} = \mathcal{L}_{intra\_ID} + \mathcal{L}_{inter\_ID} + \mathcal{L}_{inter\_triplet}. \tag{7}$$

To make samples more effective in intra-camera learning, we adopt P-K sampling strategy [16]. In each mini-batch, we choose P IDs and K samples. This sampling strategy performs balanced optimization of all person IDs.

3. EXPERIMENTS

3.1. Datasets and Implementation details

Datasets and evaluation metrics. We evaluate our method on three widely used person ReID datasets under Intra-Camera-Supervised setting, including Market-1501 [17], DukeMTMC-ReID [18], and MSMT17 [19]. Market1501 [17] contains 1,501 persons and 32,668 images from 6 cameras. DukeMTMC-ReID [18] contains 1,404 persons and 36,411 images from 8 cameras. MSMT17 [19] is the largest and most challenging dataset, which contains 4,101 persons with 126,411 images from 15 cameras. For performance evaluation, we adopt the widely used Cumulative Matching Characteristic (CMC) [20] and mean Average Precision (mAP) as metrics.

Implementation details. Images are resized to 256×128 in pixel before the intra- and inter-camera training. We use ADAM as the optimizer. The total epoch number is 50. The initial learning rate is 0.00035 with a warm-up scheme in the first 10 epochs, and is divided by 10 after each 20 epochs. The training batch size is 32. During intra-camera training, we randomly sample 8 persons and 4 images for each intra-camera ID. The memory updating rate μ is set to be 0.05, and the temperature factor τ is 0.01. The triplet loss margin m_1 is set to 0.1. We adopt DBSCAN [15] with a threshold of 0.6 for the inter-camera prototype learning. During training,

Method	Market1501				DukeMTMC-ReID				MSMT17			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
MTML	85.3	-	96.2	65.2	71.7	-	86.9	50.7	44.1	-	63.9	18.6
PCSL	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	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	88.7	-	97.1	71.1	76.9	-	89.6	56.6	46.0	-	65.3	19.1
PIRID	**91.03**	**96.7**	**97.86**	**79.57**	**79.85**	**88.6**	**91.38**	**65.37**	**60.6**	**73.76**	**79.27**	**34.85**

Table 1. Comparison with other Intra-Camera-Supervised ReID methods on three datasets.

Method	Market1501		DukeMTMC-ReID		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
PIRID-1	85.48	69.73	77.11	61.08	50.49	25.56
PIRID-2	90.83	79.18	79.35	64.71	59.72	34.37
PIRID	**91.03**	**79.57**	**79.85**	**65.37**	**60.6**	**34.85**

Table 2. Ablation study on different variants of our method.

we adopt intra-camera loss in the first 5 epochs. While in the remaining epochs, both three losses work together to guide the model learning.

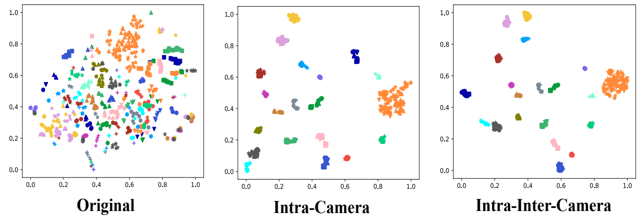
3.2. Comparison with other baselines

We compare our method with some state-of-the-art ICS methods, including MTML [7], PCSL [8], ACAN [21], and MATE [22]. MTML [7] and MATE [22] both adopt the multi-branch structure to learn parametric classifiers in both intra- and inter-camera learning. ACAN [21] develops a multi-camera adversarial learning approach to reduce the cross-camera data distribution discrepancy, while PCSL [8] utilizes a soft-labeling scheme to improve the performance.

Table 1 shows the overall performance results of our method and all the baselines on Market1501, DukeMTMC-ReID, and MSMT17 datasets, respectively. We can find that our method performs best and consistently outperforms state-of-the-art methods. For example, the mAP is 8.47%, 8.77%, and 15.75% relatively higher than the best performances obtained by other methods on Market1501, DukeMTMC-ReID, and MSMT17 respectively. Similarly, the R1 accuracy of our method is 2.33%, 2.95%, and 14.6% higher than the best performances obtained by other methods on three datasets, respectively. In a nutshell, performance comparisons verify the effectiveness of our method.

3.3. Ablation Study

In order to further explore the effectiveness of our method, we comprehensively investigate several variants and show the results in Table 2. Specifically, PIRID-1 is the variant which only adopts intra-camera loss during the whole training procedure. Besides, PIRID-2 is the variant which excludes the inter-camera triplet loss. By analyzing the results of ablation study, we can find the following observations. The results of PIRID-1 and PIRID-2 are worse than the full PIRID method, which shows the superiority of PIRID on tackling the reduction of training samples within each inter-camera prototype in ICS ReID setting. Besides, the result of PIRID-1 is inferior to PIRID-2, which proves the importance of considering the inter-camera prototype learning. We argue that these variants are inferior to our full PIRID method, indicating all the proposed modules contribute to generating camera-agnostic and discriminative features in the ICS ReID task.



Original

Intra-Camera

Intra-Inter-Camera

Fig. 3. The t-SNE visualization results on DukeMTMC-ReID. Different colors and shapes denote different IDs and cameras, respectively (best viewed in color).

3.4. Qualitative Results

To directly demonstrate the efficacy of our method, we further show some qualitative results of features learned on DukeMTMC-ReID. For the sake of simplicity, we randomly choose images of 20 person IDs, and show the t-SNE [23] visualizations in Fig. 3. As can be seen, results of original features (left) are hard to be classified. In contrast, features of intra-camera learning (middle) are well separated and show discriminative structures. Obviously, our method (right) not only separates different IDs well, but also makes images of the same ID more *'clustering'*.

4. CONCLUSION

In this work, we propose a Prototype-based Inter-camera ReID (PIRID) method, which tackles the ICS setting through the lens of prototype learning. We first introduce the intra-camera learning to seperately generate discriminative features within each camera view. Moreover, the inter-camera prototype learning provides prototypes as the representatives of each class in the common space and generates camera-agnostic features. Experiments conducted on three benchmarks show the superiority of our method. In future work, we will explore the ICS ReID task through the lens of data augmentation and identity alignment across different cameras.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grants 62106258 and 62006242.

6. REFERENCES

[1] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European conference on computer vision*, 2018, pp. 480–496.

[2] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8351–8361.

[3] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen, "Densely semantically aligned person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 667–676.

[4] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 994–1003.

[5] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3633–3642.

[6] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 598–607.

[7] X. Zhu, X. Zhu, M. Li, V. Murino, and S. Gong, "Intra-camera supervised person re-identification: A new benchmark," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[8] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[9] Zhixiong Zeng, Shuai Wang, Nan Xu, and Wenji Mao, "Pan: Prototype-based adaptive network for robust cross-modal retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1125–1134.

[10] Yang Liu, Qingchao Chen, and Samuel Albanie, "Adaptive cross-modal prototypes for cross-domain visual-language retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14954–14964.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[13] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[14] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for real-world person re-identification," *arXiv:2006.02631v1*, 2020.

[15] Martin Ester, Hans-Peter Kriegel, J Sander, Xiaowei Xu, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, 1996.

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

[17] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[18] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*, 2016, pp. 17–35.

[19] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[20] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, 2007, number 5, pp. 1–7.

[21] L. Qi, L. Wang, Huo J, Y. Shi, and Y. Gao, "Adversarial camera alignment network for unsupervised cross-camera person re-identification," *arXiv preprint arXiv:1908.00862*, 2019.

[22] X. Zhu, X. Zhu, M. Li, P. Morerio, V. Murino, and S. Gong, "Intra-camera supervised person re-identification," *International Journal of Computer Vision*, 2021.

[23] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, 2008.