CAMERA-AWARE STYLE SEPARATION AND CONTRASTIVE LEARNING FOR UNSUPERVISED PERSON RE-IDENTIFICATION

Xue Li, Tengfei Liang, Yi Jin*, Tao Wang, Yidong Li

School of Computer and Information Technology, Beijing JiaoTong University, Beijing, China
{xue.li, tengfei.liang, yjin, twang, ydli}@bjtu.edu.cn

ABSTRACT

Unsupervised person re-identification (ReID) is a challenging task without data annotation to guide discriminative learning. Existing methods attempt to solve this problem by clustering extracted embeddings to generate pseudo labels. However, most methods ignore the intra-class gap caused by camera style variance, and some methods are relatively complex and indirect although they try to solve the negative impact of the camera style on feature distribution. To solve this problem, we propose a camera-aware style separation and contrastive learning method (CA-UReID), which directly separates camera styles in the feature space with the designed camera-aware attention module. It can explicitly divide the learnable feature into camera-specific and camera-agnostic parts, reducing the influence of different cameras. Moreover, to further narrow the gap across cameras, we design a camera-aware contrastive center loss to learn more discriminative embedding for each identity. Extensive experiments demonstrate the superiority of our method over the state-of-the-art methods on the unsupervised person ReID task.

Index Terms— Person re-identification, unsupervised learning, camera-aware separation, contrastive learning

1. INTRODUCTION

Person re-identification (ReID) aims at retrieving images of the same person across cameras with non-overlapping views, which has important applications in the field of intelligent surveillance and public security. In the past few years, researchers pay more and more attention to this task and make great research progress. However, existing high-performance methods mainly use the supervised learning framework with identities’ annotation and the process of building a dataset is time-consuming and laborious. Besides, the supervised models need to be retrained in a new scenario, which have poor scalability and are not conducive to the practical deployment. Therefore, some researchers recently turn their attention to unsupervised person ReID (UReID) \[1\] \[2\] \[3\] \[4\], which has great potential value to get rid of dependence on manual annotation and become an emerging research hotspot.

Fig. 1. A high-level overview of our proposed method: the CA-UReID method disentangles camera-agnostic feature from the original feature in a explicit way. By jointly optimized with the camera-aware contrastive learning, our method can effectively reduce the intra-class distance and increase the inter-class distance with high matching accuracy.

Existing unsupervised person ReID methods are mainly divided into two categories: unsupervised domain adaptive (UDA) methods and purely unsupervised learning (USL) methods. UDA methods \[1\] \[5\] \[2\] \[6\] first train the model on a supervised source-domain dataset to make the model have a certain discriminant ability for identities. Then the training is carried out in the target domain without annotation to transfer the ability. However, the performance of UDA methods is greatly affected by the quality and size of the annotated source domain datasets, which constrains their scalability. In contrast, the purely unsupervised methods \[3\] \[4\] that requires only the unlabeled target dataset is more flexible. USL methods mainly rely on clustering to generate pseudo labels for supervised learning in the target dataset with the strategy of contrastive learning. However, most USL methods \[7\] \[8\] \[9\] only focus on improving the differences between classes and ignore the intra-class differences caused by camera style variance, including camera property and environment information of cameras, etc. Several methods \[10\] \[11\] \[4\] that consider the influence of cameras are complicated and indirect.
Based on the above observations, we propose a novel purely unsupervised learning method, the camera-aware style separation and contrast learning network (CA-UReID), which directly disentangles camera style at the level of feature map and utilize camera-aware contrastive learning to learn more discriminative feature for each identity (Fig. 1). Specifically, we first design a camera style separation (CSS) module that divides the extracted features into camera-specific and camera-agnostic parts with attention mechanism, so as to alleviate the impact of camera style variance on performance. In the camera-specific branch, the disentangled attention module captures the camera style information under the guidance of the proposed camera separation loss. Then, by using complementary attention maps, the camera-agnostic branch can enforce the network to focus on the persons, which extracts more camera-invariant features with the compact intra-class distribution. After the CSS module, we propose a camera-aware contrastive center (CACC) loss into the optimization objective function during the training. Then we introduce our proposed camera style separation module and camera-aware contrastive center loss, our method can learn more robust features of identities for the final retrieval.

Our main contributions are summarized as follows:

- We propose a novel camera style separation module to explicitly disentangle camera-specific information from feature maps and reduce intra-class variance.
- We design a more robust camera-aware contrastive center loss, which further enhances camera-agnostic features at center level to get discriminative embeddings.
- Extensive experiments on mainstream datasets (Market1501 [12] and DukeMTMC-reID [13]) show superior performance of our proposed CA-UReID method.

2. RELATED WORK

Unsupervised Person Re-identification. Unsupervised person ReID methods can be categorized into the unsupervised domain adaptive (UDA) methods and the purely unsupervised learning (USL) methods. The main difference between them is whether to use labeled source domain datasets. As for the UDA methods, some of them use GAN [14, 15] for style transfer to narrow the gap between the source domain and target domain. Some other methods are designed to generate more reliable pseudo labels: SpCL [2] introduces a cluster reliability criterion to retain reliable clusters and disassemble unreliable clusters to outlier instances. DAAM [11] assigns different weights to samples based on their distance from the class centers. UNRN [6] measured the reliability of pseudo labels by inconsistencies between the teacher network and student network. Compared with the UDA method, the USL method is more flexible and promising without manual annotations. Some approaches explore special clustering methods, such as k-NN [7, 8] and graphs [9], to generate pseudo-labels. MMCL [16] uses a memory-based multi-label classification loss to update the model. CycAs [17] learns pedestrian embedding from original videos to mine temporal information. Most USL methods focus on the distance between classes but often ignore the intra-class variance caused by inconsistent camera styles. In contrast, our USL-based CA-UReID method considers the effects of camera style, which effectively improves the consistency within the class.

Research of Camera Style in UReID. Eliminating the effects of camera style has recently been highlighted by a few unsupervised ReID methods. CIDC [10] method transfers samples to each camera via StarGAN [18]. Some other methods use intra-camera and inter-camera combined training. In the intra-camera training stage, a classifier is set under each camera, and all the classifiers share a backbone network. In the inter-camera training stage, existing methods design different strategies to unify the knowledge learned from different cameras: IICS [11] uses a combination of Instance Normalization (IN) and Batch Normalization (BN) to improve the generalization ability of classifiers. CAP [4] designs camera-based clusters to pull positive clusters and push nearest negative proxies. MetaCam [19] uses meta-learning to learn robust models for cameras. Most of these methods have complex training processes, and their way of eliminating the influence of camera style is indirect. Different from them, our proposed CA-UReID method separates camera style in an explicit way by disentangling camera-specific information in the feature space. Moreover, compared with the existing CAP [4] and ICE [3] methods, our designed camera-aware contrastive loss considers the overall distribution in each camera-aware cluster, which is more robust with better performance.

3. METHOD

In this section, we introduce our camera-aware style separation and contrastive learning (CA-UReID) method in detail. In Sec. 3.1, we first describe the overall architecture of CA-UReID, including the structure of network and the overall objective function during the training. Then we introduce our proposed camera style separation module and camera-aware contrastive center loss in the next two subsections.

3.1. Overall Structure

As shown in Fig. 2 our method follows the memory bank-based framework, constructing an unsupervised learning network. The overall process of CA-UReID mainly includes three stages: identities’ feature extraction, two-branch camera style separation, and self-supervised contrastive learning.
In the feature extraction stage, let $X = \{x_1, x_2, \ldots, x_N\}$ denote the person ReID dataset, and $N$ is the number of samples. Specifically, given an input sample $x_i$, we use ResNet-50 as the backbone to extract the corresponding feature map $F_{x_i} \in \mathbb{R}^{H \times W \times C}$, where $H$, $W$, $C$ denote height, width and channel number respectively. In the stage of two-branch camera style separation, the designed camera separation (Cam-Sep) attention module is used to disentangle the camera-specific feature map $F_{sp,x_i}$ and camera-agnostic feature map $F_{ag,x_i}$ from $F_{x_i}$. The $F_{sp,x_i}$ is fed into the camera-specific branch under the constraints of the proposed camera separation loss $L_{sep}$, and we use the $F_{ag,x_i}$ to generate the camera-agnostic 2048-dim feature embedding $f_i$ with the global average pooling (GAP) and batch normalization (BN) operations.

During the last stage of self-supervised contrastive learning, as shown in the upper part of Fig. 2, our method follows the framework based on memory bank like some previous methods [2]. Before training, the memory bank is initialized with the embeddings extracted from the whole training set by using ImageNet-pretrained ResNet-50. At the beginning of each epoch, the $k$-reciprocal-coded Jaccard distance matrix of all pedestrian features in the bank is used for DBSCAN clustering to obtain pseudo labels $Y = \{y_1, y_2, \ldots, y_N\}$. Base on the pseudo labels, our method can calculate the common InfoNCE loss [20] $L_{base}$ and our proposed camera-aware contrastive center loss $L_{caacc}$. Let $\mathcal{M}(\mathcal{M} \in \mathbb{R}^{d \times N})$ denotes the memory bank, where $d$ is the dimension of the embeddings. At the end of each training iteration, extracted embeddings $f_i$ are used to update the memory bank $\mathcal{M}$ at the instance level with the momentum mechanism:

$$\mathcal{M}[i] \leftarrow \mu \cdot \mathcal{M}[i] + (1 - \mu) \cdot f_i,$$

where $\mathcal{M}[i]$ is the $i$-th item in the memory bank, storing the updated feature of sample $x_i$, and $\mu$ is a momentum coefficient that controls the updating speed.

In the training mode, the overall objective function of this CA-UReID method can be defined as follows:

$$L_{total} = L_{base} + \lambda_1 \cdot L_{sep} + \lambda_2 \cdot L_{caacc},$$

where $\lambda_1$ and $\lambda_2$ are hyper-parameters to balance these different components. In the test mode, the $f_i$ extracted from image is the embedding used for final matching.

3.2. Camera Style Separation Module

The camera style variance, including camera property and environment information, leads to the gap in intra-class features. To solve this problem, we propose the camera style separation (CSS) module to eliminate the camera style directly from the extracted feature maps. Inspired by the way DAAM [1] separates domain-related information in UDA framework, we design a novel strategy to disentangle camera-style information in our USL pipeline. As shown in Fig. 2a, the designed CSS module has two branches: the camera-specific branch and the camera-agnostic branch.

In the camera-specific branch, to separate camera-style features unrelated to pedestrians (e.g. background and camera property, etc.) from feature map, we reframe the learning of...
this branch as a camera classification problem. Note that, although we don’t have ID annotation in USL mode, we can use the cameras annotations, which is easy to obtain in the actual collection of datasets. Specifically, $F_{xi}$ is fed into a Cam-Sep attention module to obtain a weight mask $Attn(F_{xi})$ to focus on camera information. Here, we adopt the HA structure of HA-CNN [21] as the attention module to obtain both spatial and channel attention weights. Through element wise product with the feature map $F_{xi}$, camera-specific feature map $F_{xi}^{sp}$ is calculated by the following formula:

$$F_{xi}^{sp} = Attn(F_{xi}) \odot F_{xi}. \quad (3)$$

The camera-specific embedding $f_{xi}^{sp}$ is generated by GAP and BN operation on $F_{xi}^{sp}$. To guide the attention of this branch, we design the camera separation loss $\mathcal{L}_{sep}$:

$$\mathcal{L}_{sep} = \mathbb{E}[-y_{i}^{cam}\log(\sigma(FC(f_{i}^{sp}))))], \quad (4)$$

where $y_{i}^{cam}$ is the camera label, $\sigma$ is the Softmax function, and FC is the fully connected layer. $f_{xi}^{sp}$ is input into a classifier with the number of categories as the number of cameras to predict probability of belonging to each camera, which can enforce this branch to extract camera-related information.

In the camera-agnostic branch, we carry out element-wise products with the complementary mask of $Attn(F_{xi})$ and the original feature map $F_{xi}$ to obtain the camera-agnostic feature map $F_{xi}^{ag}$, which is defined as follows:

$$F_{xi}^{ag} = (1 - Attn(F_{xi})) \odot F_{xi}. \quad (5)$$

Then the camera-agnostic feature vector $f_{i}$ is obtained from $F_{xi}^{ag}$ and used to calculate $\mathcal{L}_{base}$ and $\mathcal{L}_{cacc}$. By optimizing $\mathcal{L}_{sep}$, the influence of camera styles is reduced and the model can learn more robust pedestrian features.

3.3. Camera-aware Contrastive Center Loss

To further enhance the network to extract camera-invariant embeddings $f_i$, we propose a novel camera-aware contrastive center (CACC) loss $\mathcal{L}_{cacc}$ to narrow the feature distance of different cameras in the same class and increase the feature distance of different classes.

As shown in Fig. 2b, different from the common contrastive learning loss $\mathcal{L}_{base}$ in UReID, the $\mathcal{L}_{cacc}$ considers the camera in the clustering process. Besides, unlike other camera-aware contrastive losses (ICE [3], CAP [4]), our proposed $\mathcal{L}_{cacc}$ constrains the embeddings at the set level by using a camera center instead of a sample as an anchor (Fig. 3), attempting to avoid the impact of noise samples. Specifically, given the camera number $c$ and identity class $k$ for each embedding stored in the memory bank $\mathcal{M}$ and a batch, the camera center $g_{k}^{c}$ and $p_{k}^{c}$ can be respectively formulated as:

$$g_{k}^{c} = \frac{1}{N_{k}^{c}} \sum_{i \in k \cap c} \mathcal{M}[i], \quad p_{k}^{c} = \frac{1}{N_{k}^{c}} \sum_{f_{i} \in k \cap c} f_{i}, \quad (6)$$

where $N_{k}^{c}$ is the number of all embeddings of identity class $k$ with camera number $c$ in $\mathcal{M}$, and $\hat{N}_{k}^{c}$ is the number of embeddings with class $k$ and camera number $c$ in a training batch.

4. EXPERIMENTS

4.1. Datasets and Implementation Details

Datasets. We use the mainstream ReID datasets to evaluate our proposed method: Market1501 [12] contains 32,668 images of 1,501 identities captured by 6 cameras, while DukeMTMC-reID [13] contains 36,411 images of 1,404 identities captured by 8 cameras. As the common practice, Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) are used to measure the performance.

Implementation Details. We use ResNet-50 as backbone with the pre-trained weights on ImageNet. All images are resized to 256 × 128. The data augmentation strategies used in training include random flipping, random cropping, and random erasing. In each epoch, our method uses DBSCAN clustering with k-reciprocal Jaccard distance to generate pseudo-labels. The $\epsilon ps$ in DBSCAN is set to 0.5/0.6 for Market1501/DukeMTMC-reID. Momentum factor $\mu$ is set to 0.2/0.1 on Market1501/DukeMTMC-reID. In $\mathcal{L}_{cacc}$, the temperature factor $\tau_{cacc}$ is set to 0.07, and $J$ is set to 50. The $\lambda_{1}$ and $\lambda_{2}$ in $L_{total}$ is empirically set to 0.4 and 1.0. The batch size is set to 64 with Adam optimizer. The learning rate is set to 0.00035 and multiplied by 0.1 per 20 epoch in training.
4.2. Comparison with the State-of-the-art

In experiments, we compare our method with state-of-the-art unsupervised ReID methods, including the UDA and USL methods. The experimental results are shown in Tab. 1. We list state-of-the-art UDA methods in the top half of Tab. 1. The UDA leverages the annotation information of the source domain to make it easier to achieve better performance on the target dataset. Nevertheless, without any identity annotation, the performance of our method still clearly surpasses them, outperforming the latest OPLG method with +4.5% mAP and +2.6% Rank-1 on Market1501 and +3.1% mAP and +3.0% Rank-1 on DukeMTMC-reID. As shown in the bottom half of Tab. 1, our method also achieves better performance without our proposed CSS branch and CACC loss. Compared with IICS [11], which also considers the influence of camera style, our method has obvious advantages. IICS divides the training process into the intra-camera training stage and the inter-camera training stage, requiring additional AIBN to integrate the knowledge learned under different cameras and reduce the intra-camera variance. Our proposed camera separation structure can separate the feature map, which is more direct than IICS. Compared with CAP [4] and ICE [3] using camera-aware proxies, our method still achieve better performance on both Market1501 and DukeMTMC-reID. It also indicates that our designed center-based cross-camera loss has superiority with less sensitivity to noise samples.

4.3. Ablation Study

In this section, we further discuss and analyze the effectiveness of key components, including the camera style separation module and the camera-aware contrastive center loss.

As shown in Tab. 2 the Baseline (Index-1) is the method without our proposed CSS branch and CACC loss. Compared with the Baseline, the CSS module brings performance improvements on all two metrics (mAP, Rank1) are 2.0%, 0.6% on Market1501 and 0.9%, 0.2% on DukeMTMC-reID. The improvements of $\mathcal{L}_{\text{cacc}}$ (Index-3) over baseline are 4.1%, 2.0% on Market and 3.3%, 2.8% on Duke. The best performance can be obtained by using both CSS and $\mathcal{L}_{\text{cacc}}$ (Index-4), with improvements of 5.8%, 2.8% on Market and 4.6%, 3.0% on Duke. It shows the good complementarity of CSS and $\mathcal{L}_{\text{cacc}}$ and they can jointly optimize the feature learning.

To observe the function of CSS module, we visualize the learned attention maps in the HA structure. As shown in Fig. 4 the camera-specific maps mainly focus on background, which is more related to cameras because of their fixed position. In contrast, the complementary camera-agnostic attention maps focus on the main body of pedestrian, which can narrow the gap between cameras within the class. In experiments, we also compare two different constraints: the camera-aware contrastive sample loss $\mathcal{L}_{\text{casc}}$ and camera-aware contrastive center loss $\mathcal{L}_{\text{cacc}}$, which act on samples and centers, respectively. From Tab. 2 we can notice that the version of center-based loss achieves higher performance with stronger retrieval ability. In general, by using two components together, our method can generate more discriminative embeddings in this person ReID task under USL settings.

5. CONCLUSION

In this paper, we propose a camera-aware style separation and contrastive learning method for purely unsupervised person ReID task, which can effectively separate camera style from the feature map by our proposed camera style separation module and camera-aware contrastive center loss. Experimental results demonstrate that our method shows superior performance against the existing state-of-the-art methods.

---

**Table 1. Comparison with state-of-the-art methods.**

| Method | Venue | Source | Market mAP Rank1 | Source | Duke mAP Rank1 |
|--------|-------|--------|------------------|--------|----------------|
| MMCL [16] | CVPR2020 | Duke | 60.4 84.4 | Market | 51.4 72.4 |
| DAAM [11] | AAAI2020 | Duke | 67.8 86.4 | Market | 63.9 77.6 |
| HGA [22] | AAAI2021 | Duke | 70.3 89.5 | Market | 67.1 80.4 |
| MM [8] | ICLR2020 | Duke | 71.2 87.7 | Market | 65.1 78.0 |
| JVTCA* [23] | CVPR2021 | Duke | 75.4 90.5 | Market | 67.6 81.9 |
| MetaCam [19] | CVPR2021 | Duke | 76.5 90.1 | Market | 65.0 79.5 |
| SpCL [4] | NearPS2020 | Duke | 76.7 90.3 | Market | 68.8 82.9 |
| GCMT [24] | IJCAI2021 | Duke | 77.1 90.6 | Market | 67.8 81.1 |
| UNRN [16] | AAAI2021 | Duke | 78.1 91.9 | Market | 69.1 82.0 |
| GLT [25] | CVPR2021 | Duke | 79.5 92.2 | Market | 69.2 82.0 |
| OPLG [26] | ICCV2021 | Duke | 80.0 91.5 | Market | 70.1 82.2 |

**Table 2. Ablation studies of the designed components.**

| Index | Methods | Market1501 mAP Rank1 | Duke-reID mAP Rank1 |
|-------|---------|-----------------------|---------------------|
| 1     | Baseline | 78.7 91.3 | 68.6 82.2 |
| 2     | Baseline + CSS | 80.7 91.9 | 69.5 82.4 |
| 3     | Baseline + $\mathcal{L}_{\text{cacc}}$ | 82.8 93.3 | 71.9 85.0 |
| 4     | Baseline + CSS + $\mathcal{L}_{\text{cacc}}$ | 84.5 94.1 | 73.2 85.2 |
| 5     | $\mathcal{L}_{\text{cacc}}$ (with sample) | 83.5 93.4 | 72.0 84.0 |
| 6     | $\mathcal{L}_{\text{cacc}}$ (with center) | 84.5 94.1 | 73.2 85.2 |

**Fig. 4. Visualization of attention map in the CSS module.**

As shown in Tab. 2 the Baseline (Index-1) is the method without our proposed CSS branch and CACC loss. Compared with the Baseline, the CSS module brings performance improvements on all two metrics (mAP, Rank1) are 2.0%, 0.6% on Market1501 and 0.9%, 0.2% on DukeMTMC-reID. The improvements of $\mathcal{L}_{\text{cacc}}$ (Index-3) over baseline are 4.1%, 2.0% on Market and 3.3%, 2.8% on Duke. The best performance can be obtained by using both CSS and $\mathcal{L}_{\text{cacc}}$ (Index-4), with improvements of 5.8%, 2.8% on Market and 4.6%, 3.0% on Duke. It shows the good complementarity of CSS and $\mathcal{L}_{\text{cacc}}$ and they can jointly optimize the feature learning.

To observe the function of CSS module, we visualize the learned attention maps in the HA structure. As shown in Fig. 4 the camera-specific maps mainly focus on background, which is more related to cameras because of their fixed position. In contrast, the complementary camera-agnostic attention maps focus on the main body of pedestrian, which can narrow the gap between cameras within the class. In experiments, we also compare two different constraints: the camera-aware contrastive sample loss $\mathcal{L}_{\text{casc}}$ and camera-aware contrastive center loss $\mathcal{L}_{\text{cacc}}$, which act on samples and centers, respectively. From Tab. 2 we can notice that the version of center-based loss achieves higher performance with stronger retrieval ability. In general, by using two components together, our method can generate more discriminative embeddings in this person ReID task under USL settings.
6. REFERENCES

[1] Yangru Huang, Peixi Peng, Yi Jin, Yidong Li, and Junliang Xing, “Domain adaptive attention learning for unsupervised person re-identification,” in AAAI, 2020, pp. 11069–11076.

[2] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li, “Self-paced contrastive learning with hybrid memory for domain adaptive object re-id,” in NeurIPS, 2020.

[3] Hao Chen, Benoit Ladadec, and François Brémond, “ICE: inter-instance contrastive encoding for unsupervised person re-identification,” CoRR, vol. abs/2103.16364, 2021.

[4] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaoqin Gong, and Xian-Sheng Hua, “Camera-aware proxies for unsupervised person re-identification,” in AAAI, 2021, pp. 2764–2772.

[5] Yixiao Ge, Dapeng Chen, and Hongsheng Li, “Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification,” in ICLR, 2020.

[6] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha, “Exploiting sample uncertainty for domain adaptive person re-identification,” in AAAI, 2021, pp. 3538–3546.

[7] Minxian Li, Xiaticn Zhu, and Shaoqang Gong, “Unsupervised person re-identification by deep learning tracklet association,” in ECCV, 2018, pp. 772–788.

[8] Yanbei Chen, Xiaticn Zhu, and Shaoqang Gong, “Deep association learning for unsupervised video person re-identification,” in BMVC, 2018, p. 48.

[9] Jinlin Wu, Hao Liu, Yang Yang, Zhen Lei, Shengcai Liao, and Stan Z. Li, “Unsupervised graph association for person re-identification,” in ICCV, 2019, pp. 8320–8329.

[10] Jiajie Tian, Qihao Tang, Rui Li, Zhu Teng, Baopeng Zhang, and Jianping Fan, “A camera identity-guided distribution consistency method for unsupervised multi-target domain person re-identification,” ACM Trans. Intell. Syst. Technol., vol. 12, no. 4, pp. 38:1–38:18, 2021.

[11] Shiyu Xuan and Shiliang Zhang, “Intra-inter camera similarity for unsupervised person re-identification,” in CVPR, 2021, pp. 11926–11935.

[12] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, “Scalable person re-identification: A benchmark,” in ICCV, 2015, pp. 1116–1124.

[13] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in ECCV Workshops, 2016, pp. 17–35.

[14] Weijian Deng, Liang Zheng, Qiixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in CVPR, 2018, pp. 994–1003.

[15] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, “Person transfer GAN to bridge domain gap for person re-identification,” in CVPR, 2018, pp. 79–88.

[16] Dongkai Wang and Shiliang Zhang, “Unsupervised person re-identification via multi-label classification,” in CVPR, 2020, pp. 10978–10987.

[17] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang, “Cy-cas: Self-supervised cycle association for learning re-identifiable descriptions,” in ECCV, 2020, pp. 72–88.

[18] Yunjie Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in CVPR, 2018, pp. 8789–8797.

[19] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzengtai Cai, Yaqin Lin, Shaozi Li, and Nicu Sebe, “Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification,” in CVPR, 2021, pp. 4855–4864.

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding,” CoRR, vol. abs/1807.03748, 2018.

[21] Wei Li, Xiatian Zhu, and Shaoqang Gong, “Harmonious attention network for person re-identification,” in CVPR, 2018, pp. 2285–2294.

[22] Minying Zhang, Kai Liu, Yidong Li, Shihui Guo, Hongtao Duan, Yimin Long, and Yi Jin, “Unsupervised domain adaptation for person re-identification via heterogeneous graph alignment,” in AAAI, 2021, pp. 3360–3368.

[23] Hao Chen, Yaohui Wang, Benoit Ladadec, Antitza Dantcheva, and François Brémond, “Joint generative and contrastive learning for unsupervised person re-identification,” in CVPR, 2021, pp. 2004–2013.

[24] Xiaobin Liu and Shiliang Zhang, “Graph consistency based mean-teaching for unsupervised domain adaptive person re-identification,” in IJCAI, 2021, pp. 874–880.

[25] Kecheng Zheng, Wu Liu, Lingxia He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha, “Group-aware label transfer for domain adaptive person re-identification,” in CVPR, 2021, pp. 5310–5319.

[26] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Dongliang Qi, and Dapeng Chen, “Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification,” in ICCV, 2021, pp. 8371–8381.

[27] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang, “A bottom-up clustering approach to unsupervised person re-identification,” in AAAI, 2019, pp. 8738–8745.

[28] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian, “Unsupervised person re-identification via soft-ened similarity learning,” in CVPR, 2020, pp. 3387–3396.

[29] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo, “Hierarchical clustering with hard-batch triplet loss for person re-identification,” in CVPR, 2020, pp. 13654–13662.

[30] Yiming Wu, Xintian Wu, Xi Li, and Jian Tian, “MGH: metadata guided hypergraph modeling for unsupervised person re-identification,” in ACM Multimedia, 2021, pp. 1571–1580.