Second-order Camera-aware Color Transformation for Cross-domain Person Re-identification

Wangmeng Xiang[1][0000−0002−4373−2610] Hongwei Yong[1][0000−0001−7603−6497]
Jianqiang Huang[2][0000−0001−5735−2910] Xian-Sheng Hua[2][0000−0002−8232−5049]
and Lei Zhang[1,2][0000−0002−2078−4215]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{cswxiang,cshyong,cslzhang}@comp.polyu.edu.hk
[2] Artificial Intelligence Center, Alibaba DAMO Academy, Hangzhou, China
{jianqiang.hjq,xiansheng.hxs}@alibaba-inc.com

Abstract. In recent years, supervised person re-identification (person ReID) has achieved great performance on public datasets, however, cross-domain person ReID remains a challenging task. The performance of ReID model trained on the labeled dataset (source) is often inferior on the new unlabeled dataset (target), due to large variation in color, resolution, scenes of different datasets. Therefore, unsupervised person ReID has gained a lot of attention due to its potential to solve the domain adaptation problem. Many methods focus on minimizing the distribution discrepancy in the feature domain but neglecting the differences among input distributions. This motivates us to handle the variation between input distributions of source and target datasets directly. We propose a Second-order Camera-aware Color Transformation (SCCT) that can operate on image level and align the second-order statistics of all the views of both source and target domain data with original ImageNet data statistics. This new input normalization method, as shown in our experiments, is much more efficient than simply using ImageNet statistics. We test our method on Market1501, DukeMTMC, and MSMT17 and achieve leading performance in unsupervised person ReID.

1 Introduction

Person re-identification (person ReID) is an important computer vision task, which aims to identify the same person from a set of images captured under different cameras [1]. The task of person ReID is very challenging due to the large variation in camera viewpoint, lighting, resolution, and human pose etc. In recent years, supervised person ReID has achieved great performance under the single domain dataset [2–7, 4, 8–17]. State-of-the-art methods have achieved over 95% top1 accuracy and nearly 90% in mAP. Existing researches in supervised single domain person ReID methods can be roughly grouped into three categories: 1) transferring and improving powerful CNN architectures to person ReID [2–7], where off-the-shelf feature extractors are used as parts of the

network architecture; 2) designing more effective metrics [10–15]; 3) combining priori into network architecture for fine-grained feature learning [4, 8, 9, 16, 17].

Despite supervised single domain person ReID has achieved great accuracy, the performance of the model would drop dramatically when the model is applied to an unseen new dataset. In other words, the performance of ReID model trained on the labeled source dataset (source) is often inferior on the new unlabeled target dataset, which is due to the data-bias between these two datasets (or two domains). As it is expensive to label the target datasets, many researchers treat this task as an unsupervised domain adaptation (UDA) problem. Unfortunately, many existing UDA solutions can not be simply applied to unsupervised person ReID due to the differences in problem settings. Generally, UDA setting requires the categories of source and target domain to be the same, or at least to have overlap. However, in person ReID, the identities in source and target datasets are totally different. In recent years, approaches that aim specifically to improve the performance of unsupervised domain adaptation of person ReID have been proposed [18–21]. We categorize these methods into two different settings: direct transfer and progressive learning. In the direct transfer setting, most of methods [19, 20] minimize the discrepancy of two domains by applying carefully designed loss functions. While in progressive learning [21], pseudo labels are generated for training, and the model is trained in an iterative manner.

Although many methods focus on decreasing the domain discrepancy in the feature level, there are very few methods working on the image level. Several existing methods use generative adversarial networks [22] (GAN) to generate new data that are similar to the target domain for training. For example, PT-GAN [18] transfers the appearance of the labeled source dataset to the unlabeled target dataset using generative adversarial networks (GAN). ECN [19] makes fine-grained camera style transfer by utilizing the camera id information in the target dataset. Camera sensor variation is pointed out in [23], and they use GAN to generate domain-specific images for every view of the camera. Our proposed method also works on image level, unlike previous works that use additional GAN for data generation, we aim to minimize the discrepancy of datasets by matching the camera-wise input distributions to the second-order ImageNet [24] data statistics. We found that there are actually two "domain shift" steps in the typical unsupervised person ReID. 1) ImageNet to the source: most methods in person ReID would use a backbone network pre-trained on ImageNet, the input distribution of ImageNet is usually different from the source dataset, which means the first domain shift is from ImageNet to source dataset. 2) source to target: the second step is domain shift from source dataset to target dataset. To solve the domain shift problem mentioned above, we apply the camera-aware color transformation, which matches every camera view individually to the ImageNet statistics. This fine-grained color transformation boosts the performance of ReID model in the cross-domain setting by a large margin. In addition, we also apply a color equalization data augmentation to increase the adaptive ability of our trained model. As far as we know, it is the first time second-order

color transformation and color equalization data augmentation are applied for unsupervised person ReID. We summarize the contributions as follows:

- We propose to use the color transformation as a pre-processing step to compensate for the color changes in unsupervised person ReID in both source domain and target domain, which is a fine-grained camera-aware color transformation that can handle camera color shift and statistics changes of different cameras. It is easy to implement, fast and simple, and there is no need for tedious parameter tuning.
- We propose to use color equalization augmentation in cross-domain person ReID. This augmentation could ease the differences in the input distribution of different datasets.
- We conduct extensive experiments and ablation studies on several popular person ReID benchmarks including Market1501 [25], DukeMTMC-ReID [26], MSMT17 [18] to demonstrate the effectiveness of proposed color transformation and data augmentation solution and achieve leading performance in these datasets.

2 Related works

2.1 Supervised person ReID

In recent years, state-of-the-art supervised person ReID methods have achieved over 95% top1 accuracy in large-scale public datasets. Researchers have been working on novel network structures [2–4, 7], combining other human body prior in the training process [9, 27, 28, 8], more efficient loss functions [10, 29–31, 6, 32, 33, 11, 2] etc. For example, Sun *et al.* [3] used CaffeNet and ResNet as backbone networks. Chen *et al.* [27] developed a multi-scale network architecture with a saliency-based feature fusion. Zhou *et al.* [28] built a part-based CNN to extract discriminative and stable features for body appearance. Shi *et al.* [11] trained their network using triplet loss with hard positive pairs mining. Although these methods perform well on the single domain dataset, when directly test the model trained from the source dataset on the new unlabeled target dataset, the performance of the model would drop dramatically. This performance gap has led many researchers to cross-domain person ReID or unsupervised person ReID.

2.2 Unsupervised domain adaptation

Unsupervised domain adaptation (UDA) [34–36] aims to transfer knowledge from a labeled source dataset to an unlabeled target dataset. Many UDA methods focus on the feature domain and try to decrease the discrepancy of source and target feature distributions. For example, Gretton *et al.* [34] minimize the difference between the means of features from two domains. Sun *et al.* [35] learn a linear transformation that aligns the mean and covariance of feature distribution between source and target domain. Ganin *et al.* [36] propose a gradient reversal layer and integrate it into a deep neural network for minimizing the classification

loss while maximizing domain confusion loss. Chen *et al.* [37] propose a high-order moment matching between two domains in feature space. Some existing works try to generate pseudo-labels on the target set and utilize this information for training. For example, Sener *et al.* [38] infer the labels of unsupervised target data points in a k-NN graph and jointly train a unified deep learning framework in an end-to-end fashion. Saito *et al.* [39] propose to assign pseudo-labels to unlabeled target samples based on the predictions from two classifiers trained on source samples and one network is trained by the samples to obtain target discriminative representations. However, these methods assume the class labels are the same across domains, which is not true in person ReID. In addition, they were not designed to address the camera shifts in ReID problem. Therefore they can not achieve good performance in unsupervised person ReID and some of them can not be applied to this problem.

2.3 Unsupervised person ReID

Recently, works specifically focus on unsupervised person ReID have been proposed to tackle the scalability problem. Several methods utilize GAN to generate new data that looks similar to target domain for training. For example, Wei *et al.* [18] transfer the appearance of labeled source dataset to the unlabeled target dataset using cycle GAN [40]. Zhong *et al.* [19] make fine-grained camera style transfer by utilizing the camera id information in the target dataset. Some works use additional attribute information for cross-domain knowledge transfer. For instance, Wang *et al.* [41] propose to learn an attribute-semantic and identity discriminative feature representation space for the target domain. They utilize attribute information of person to bridge the source and target domain. Cross-camera scenes variation in person ReID is significant in many ways including image resolution, color, and viewpoints changes. These variants lead to a huge discrepancy in image statistics of images captured by different cameras. Existing methods handle cross-camera scene variation by camera-to-camera alignment at image level [19, 20] or feature level [42]. Zhong *et al.* [20] try to use GAN to generate different camera style images of the target dataset. Wu *et al.* [43] consider the domain shift among different cameras and propose to keep cross camera-aware similarity consistency and intra-camera similarity preservation by minimizing two consistency loss. UPR [44] adjusts images' hue, saturation, lightness, and contrast to enhance the adaptation ability of models by data augmentation. Different from previous works, we propose a second-order color transform method that focuses on the distribution of input images. Besides, color equalization augmentation is used to make model less sensitive to the data distribution changes.

3 Proposed method

As we mentioned in Section 1, we noticed that the DNNs' backbone model used for person ReID are usually pre-trained on ImageNet to obtain better performance. However, the input statistics of source and target ReID datasets (e.g.

Market1501, DukeMTMC) are different from ImageNet. Therefore, matching the color statistics of the source and target ReID dataset with ImageNet would reduce the domain discrepancy. Another observation is that the image distribution of camera views are inconsistent due to the camera sensor variation. As we can see from Fig. 1, the color of the clothes of the same person looks different under different camera views. The color mean statistic of six cameras in Market1501 in Fig. 2 shows that there are clear differences in input color distribution of different cameras. This observation motivates us to make color statistics of all the cameras be the same. More specifically, we use a linear transformation to match the first-order statistics (mean) and second-order statistics (covariance) of input images with ImageNet statistics.

Fig. 1. Samples of color changes of different camera views of Market1501. The column 1-6 and 7-12 are sampled from camera view 1 to 6 respectively. As we can see the camera view 6 (column 6,12) is very different from other camera views.

3.1 Color Statistics Calculation

For color matching of different camera views, we need to obtain the statistics of datasets. Due to the limitation of computation resources, we cannot load the whole dataset into memory to make the full computation, especially for the large scale dataset (e.g. ImageNet). Therefore, we adopt an incremental computation method. Suppose $\mathbf{X}_k \in \mathtt{R}^{m \times n \times 3}$ the k-th input image of training dataset, $m \times n$ is the size of input images and 3 is the number of channels (i.e. R,G and B), and $\mathbf{x}_{ijk} \in \mathtt{R}^3$ denotes the color vector of a pixel in image $\mathbf{X}_k$. We use the following incremental rules to obtain the color mean and covariance of input images:

$$\begin{aligned} S^k[\mathbf{x}] &= S^{k-1}[\mathbf{x}] + \sum_{i,j} \mathbf{x}_{ijk}, \\ S^k[\mathbf{x}\mathbf{x}^T] &= S^{k-1}[\mathbf{x}\mathbf{x}^T] + \sum_{i,j} \mathbf{x}_{ijk}\mathbf{x}_{ijk}^T, \\ N^k &= N^{k-1} + mn \end{aligned} \tag{1}$$

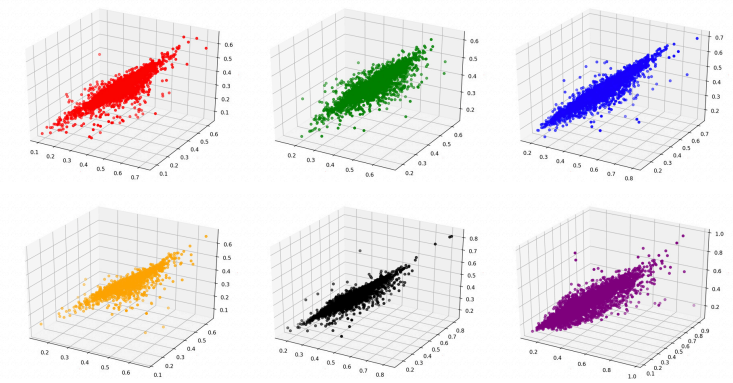

Fig. 2. Color mean statistic of different camera views on Market1501. It is clear that the scale and shape of color mean statistic distribution are different among various camera views.

where $S^k[\mathbf{x}]$ and $S^k[\mathbf{x}\mathbf{x}^T]$ is the incremental statistics of the first k images, respectively, N^k is the number of pixels of the first k images, and $S^0[\mathbf{x}]$ and $S^0[\mathbf{x}\mathbf{x}^T]$ is a zero vector and matrix, respectively, and $N^0 = 0$. After we obtain the final statistics $S[\mathbf{x}]$ and $S[\mathbf{x}\mathbf{x}^T]$ of all input images, the mean and covariance matrix are:

$$\boldsymbol{\mu} = \frac{1}{N} S[\mathbf{x}], \quad \boldsymbol{\Sigma} = \frac{1}{N} S[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{2}$$

We can use these incremental rules to get mean and covariance of input images without occupying too much memory. Because we only need to save $S^k[\mathbf{x}]$, $S^k[\mathbf{x}\mathbf{x}^T]$, N^k and current image $\mathbf{X}_k$ at every step.

3.2 Mean and Covariance Matching

Given the statistics μ and $\boldsymbol{\Sigma}$ for input images of training dataset and the target statistics μ_0 and $\boldsymbol{\Sigma}_0$, we need to find a linear transformation to make the first and second order statistics of transformed pixel the same as target statistics. The pixel $\mathbf{x}$ is transformed as: $\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A}$ is a 3×3 matrix and $\mathbf{b}$ is a 3-vector. We need to find proper $\mathbf{A}$ and $\mathbf{b}$ to satisfy $E[\mathbf{x}'] = \boldsymbol{\mu}_0$ and $E[\mathbf{x}'\mathbf{x}'^T] - \boldsymbol{\mu}_0^2 = \boldsymbol{\Sigma}_0$, then we can obtain the following conditions:

$$\begin{aligned} \mathbf{A}\mu + \mathbf{b} &= \boldsymbol{\mu}_0 \\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T &= \boldsymbol{\Sigma}_0 \end{aligned} \tag{3}$$

The solution for $\mathbf{A}$ and $\mathbf{b}$ are not unique under this conditions. There remain a family of solutions. Suppose the eigenvalue decomposition for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$ is $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ and $\boldsymbol{\Sigma}_0 = \mathbf{U}_0\boldsymbol{\Lambda}_0\mathbf{U}_0^T$, Then we have

$$\begin{aligned} \mathbf{A} &= \mathbf{U}_0\boldsymbol{\Lambda}_0^{\frac{1}{2}}\mathbf{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T \\ \mathbf{b} &= \boldsymbol{\mu}_0 - \mathbf{A}\boldsymbol{\mu} \end{aligned} \tag{4}$$

where $\mathbf{Q}$ is any orthogonal matrix (i.e. $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$). Therefore, we can choose a proper orthogonal matrix $\mathbf{Q}$ to get a specific solution. The most two common choices for $\mathbf{Q}$ is $\mathbf{I}$ and $\mathbf{U}_0^T\mathbf{U}$, and their corresponding solution for $\mathbf{A}$ are $\mathbf{A} = \mathbf{U}_0\mathbf{\Lambda}_0^{\frac{1}{2}}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$ and $\mathbf{A} = \mathbf{U}_0\mathbf{\Lambda}_0^{\frac{1}{2}}\mathbf{U}_0^T\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$, respectively. In our experiments, we use the second choice for $\mathbf{Q}$.

3.3 Transformation for Different Cameras

In order to reduce the difference of color distribution among different cameras, we use a linear transformation for the images from one camera to make color statistics the same as images from ImageNet. Suppose we have obtained the mean and covariance of the images from different cameras according to Eq. (1) and (2), and we use $\{\boldsymbol{\mu}_c\}_{c=1}^C$ and $\{\mathbf{\Sigma}_c\}_{c=1}^C$ to denote the color mean and covariance matrix of images from C cameras, and use $\boldsymbol{\mu}_I$ and $\mathbf{\Sigma}_I$ to denote the color mean and covariance matrix of image from ImageNet. Then we adopt the Eq. (4) to get the linear transformation parameters $\{\mathbf{A}\}_{c=1}^C$ and $\{\mathbf{b}\}_{c=1}^C$ for images from different cameras:

$$\begin{aligned} \mathbf{A}_c &= \mathbf{U}_I\mathbf{\Lambda}_I^{\frac{1}{2}}\mathbf{U}_I^T\mathbf{U}_c\mathbf{\Lambda}_c^{-\frac{1}{2}}\mathbf{U}_c^T \\ \mathbf{b}_c &= \boldsymbol{\mu}_I - \mathbf{A}_c\boldsymbol{\mu}_c, \quad c = 1, 2, ..., C, \end{aligned} \tag{5}$$

where $\mathbf{\Sigma}_c = \mathbf{U}_c\mathbf{\Lambda}_c\mathbf{U}_c^T$ and $\mathbf{\Sigma}_I = \mathbf{U}_I\mathbf{\Lambda}_I\mathbf{U}_I^T$. Therefore, for a pixel $\mathbf{x}$ from the c-th camera, the input pixel for the backbone DNNs is $\mathbf{x}' = \mathbf{A}_c\mathbf{x} + \mathbf{b}_c$. After making the color statistics of images from all cameras the same as that of images from ImageNet, we also need to use the same normalization parameters (mean and variance of each channel) as that used in training ImageNet for input images. This normalization method can be viewed as simple linear transformation so it can be fused with color transformation. We use the diagonal matrix $\mathbf{A}_0 = \text{Diag}\{\frac{1}{\sigma_r}, \frac{1}{\sigma_g}, \frac{1}{\sigma_b}\}$ and $\mathbf{b}_0 = [\mu_r, \mu_g, \mu_b]^T$ to denote the normalization transformation parameters. Then, the fused linear transformation for c-th camera changes to:

$$\begin{aligned} \mathbf{x}' &= \mathbf{A}'_c\mathbf{x} + \mathbf{b}'_c, \\ \mathbf{A}'_c &= \mathbf{A}_0\mathbf{A}_c, \\ \mathbf{b}'_c &= \mathbf{A}_0\mathbf{b}_c + \mathbf{b}_0. \end{aligned} \tag{6}$$

We only need to compute and save the parameters $\mathbf{A}'_c$ and $\mathbf{b}'_c$ for each cameras, and it can work as a pre-processing step in network training.

3.4 Progressive unsupervised learning

Progressive unsupervised learning is an effective paradigm used in [21, 45] to boost the performance of unsupervised person ReID. We first train a base network on the labeled source dataset, then we follow the training strategy in [21] and conduct the progressive training on target training set in an iterative manner. The overall framework can be seen in Fig. 3.

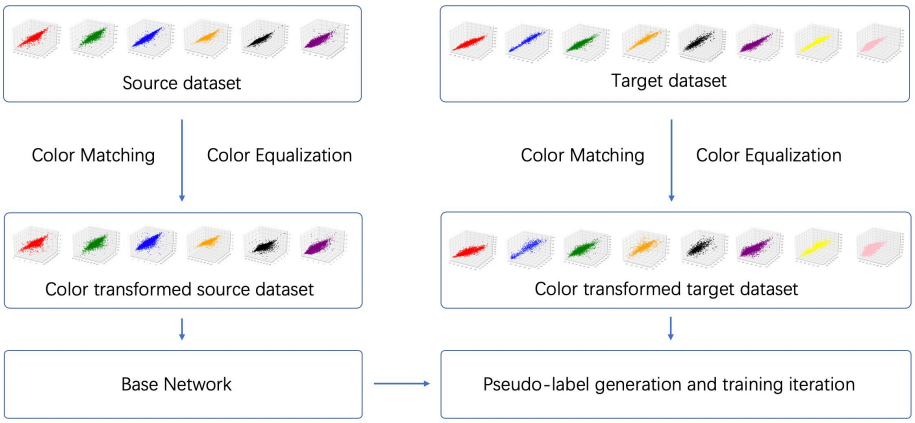

Fig. 3. Pipeline of our training framework. We use Market1501 (6 camera views) as source dataset, DukeMTMC (8 camera views) as target dataset in this figure. After training base network using transformed source data, pseudo-labels are generated from transformed target data with base network. Then progressive learning is applied to train the model in an iterative manner.

Base network training. We use ResNet50 as the backbone network for a fair comparison, as it is used by most of the state-of-the-art methods. We follow the training strategy and network structure in [45] to fine-tune on the ImageNet pre-trained model. We discard the last fully connected (FC) layer and add two FC layer. The output of the first FC layer is 2,048-dim, followed by batch normalization [46], ReLU. The output of the second FC layer is the identity number in the labeled training set. As mentioned in the previous section, we add a preprocessing step to make our model invariant to color statistic changes.

Progressive learning. Inspired by previous ReID work [47, 21], we adopt a clustering-training iterative strategy, and both global and local features are considered in the iterative training process. We denote the unlabeled image on the target training set to be I^i, after feeding the input image to the base network, the output feature could then be denoted X^i, which is a $H \times W \times C$ feature map. Besides global feature map X^i, we further split the X^i into upper body and lower body feature maps $X^i_{up}, X^i_{low} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. The feature maps are then average pooled to feature vector f^i_g, f^i_{up} and f^i_{low}. To keep the model consistency with the pre-trained model, we add a 2048-dim FC layer after f_g to get f_e.

After generating features of all target training set, we then apply DBSCAN [48] to cluster and generate pseudo-label y^i_g, y^i_{up} and y^i_{low} for each sample, respectively. Re-ranking [49] is also applied to generate a more reliable distance matrix. We use these pseudo-labels as supervised information to train the model on the target training set. We apply hard triplet loss function to fine-tune the model, and the formula can be represented as follows:

$$\mathcal{L}_t = \sum_{i \in P \times K} [\underbrace{\max \|f_a^i - f_p^i\|_2^2}_{\text{hardest positive}} - \overbrace{\left\|f_a^i - f_n^i\right\|_2^2}^{\text{hardest negative}} + m]_+ \tag{7}$$

where P is the identity number, K is the number of samples per identity, f_a^i represents anchor feature vector, f_p^i stands for positive feature vector, f_n^i stands for negative feature vector, m is the margin. This formula aims to make largest positive distance smaller than smallest negative distance by margin m. We apply hard triplet loss on both global and local features, the whole loss function can be represented as:

$$L = L_t\left(f_g, y_g\right) + L_t\left(f_{up}, y_{up}\right) + L_t\left(f_{low}, y_{low}\right) + L_t\left(f_e, y_g\right) \tag{8}$$

We also apply identity dictionary strategy proposed in [21] and randomly sample an identity in each cluster group as the identity agent. The pseudo-label of other samples would then be decided by the feature distance to the identity agent. This leads to new pseudo label set $y_{n-g}, y_{n-up}, y_{n-low}$. We apply the same hard triplet loss in Eq. 8 on these new labels. The overall loss function becomes:

$$\begin{aligned} L = & L_t\left(f_g, y_g\right) + L_t\left(f_{up}, y_{up}\right) + L_t\left(f_{low}, y_{low}\right) + L_t\left(f_e, y_g\right) \\ & + L_t\left(f_g, y_{n-g}\right) + L_t\left(f_{up}, y_{n-up}\right) + L_t\left(f_{low}, y_{n-low}\right) + L_t\left(f_e, y_{n-g}\right) \end{aligned} \tag{9}$$

The clustering and training process is kept several times until iterations.

Fig. 4. Samples of color equalization augmentation of different views of Market1501. The first row are originals, the second row are the images after color equalization.

Color equalization augmentation. Color equalization is a simple yet efficient technique for image visual effect enhancement. It equalizes the image histogram and improves the contrast by effectively spreading out the most frequent intensity

values, i.e., stretching out the intensity range of the image. In person ReID, the contrast of images is inconsistent among different camera views. Therefore, using equalization augmentation would ease the inconsistency and improve the generalization ability. The visual effect of color equalization can be seen from Fig. 4. As we can see, the images after color equalization are lighter and feel more constant among different views. Our implementation of color equalization augmentation is based on the Python Image Library(PIL) [3], which applies the library function $ImageOps.equalize()$. In all our experiments, we set the color equalization augmentation probability to 0.5.

4 Experiments

4.1 Implementation details

Baseline training. As described in Section 3.4, we first train a baseline model on the source dataset with the color transformation layer. We resize all the input images to 256×128. For data augmentation, we employ random cropping, flipping, random erasing [50], and our proposed color equalization. For hard triplet mining, in each mini-batch, we randomly selected P = 4 identities and sampled K = 8 images for each identity from the training set, so that the mini-batch size is 32. And in our experiment, we set the margin parameter to 0.5. During training, we use the Adam [51] with weight decay 0.0005 to optimize the parameters for 150 epochs. The initial learning rate is set to 3×10^{-4} and decays to 3×10^{-5} after 100 epochs.

Unsupervised training. For unsupervised training, we follow the same data augmentation strategy and triplet loss setting. And we decrease the initial learning rate from 3×10^{-4} to 6×10^{-5} and change the training epoch from 150 to 70. Besides, the whole framework is trained for 30 iterations.

4.2 Ablation Study

Effectiveness of camera-aware color transformation. We investigate the influences of different normalization and color transformation strategy. The "Baseline" method applies the mean and standard deviation of ImageNet, which is used by most of the methods. As listed in Table 1, using domain-specific mean and deviation "MD" outperforms using ImageNet statistics by around $3-4\%$ in mAP/top1 when transferred from DukeMTMC to Market1501. While camera-aware first-order normalization C_{first} improves 1.9% mAP and 1.8% top1 accuracy over "MD" when tested from D→M. This shows the effectiveness of camera-aware normalization, even though it is a simple first-order method, it can still largely boost the performance. Our proposed C_{second} achieves the best performance and outperforms other methods by a large margin. We use C to represent C_{second} for convenience in the experiments below.

[3] http://www.pythonware.com/products/pil/

Table 1. Comparison of using different color transformation methods. "Baseline" denotes using ImageNet image statistics for normalization. "MD" stands for using mean and std for normalization. "C_{first}" stands for using each camera view's mean and standard deviation of source and target datasets for input normalization. "C_{second}" stands for using the proposed second-order camera-aware color transformation.

Method	DukeMTMC → Market1501				Market1501 → DukeMTMC			
	mAP	Top1	Top5	Top10	mAP	Top1	Top5	Top10
Baseline	20.3	46.5	65.4	72.4	15.9	29.0	45.2	52.3
Baseline + MD	23.1	50.1	67.0	73.8	16.8	29.9	45.3	50.9
Baseline + C_{first}	25.0	51.9	68.2	74.6	19.8	35.5	51.3	57.5
Baseline + C_{second}	30.6	58.2	74.0	80.1	21.9	38.5	55.1	62.3

In Table 2 we can see that with camera-aware color transformation, we improve the performance by 10.3% and 15.4% in mAP and rank-1 accuracy comparing to baseline when the model is transferred from DukeMTMC to Market1501. Similarly, when the model is trained on Market1501 and tested on DukeMTMC, the performance gain becomes 6.0% and 9.5% in mAP and rank-1 accuracy, respectively. Moreover, color transform keeps boosting the performance when combined with progressive learning. It can further improve the performance by 5.9% in mAP and 4.6% in rank-1 accuracy when transfer from DukeMTMC to Market1501, which show its strong adaptive ability.

Effectiveness of equalization augmentation. We conduct an ablation study to prove the effectiveness of equalization data augmentation in Table 2. This augmentation could relieve the lighting differences between different camera views. From Table 2, we can see that with the color equalization augmentation the performance is improved by 8.6% and 11.9% in mAP and top1 accuracy comparing to "Baseline" when the model is transferred from DukeMTMC to Market1501. When the model is trained on Market1501 and tested on DukeMTMC, the performance gain becomes 8.5% and 13.9% in mAP and top1 accuracy, respectively. Moreover, when we combine color equalization and camera-aware color transformation, the performance would further be improved by around 2% in top1 accuracy and $1-2\%$ in mAP.

Effectiveness of progressive unsupervised learning. We perform several ablation studies to prove the effectiveness of progressive unsupervised learning (PUL) as listed in Table 2. Specifically, comparing "Baseline + C + E" with "Baseline + C + E + P", we improve the performance by 35.5% and 28.4% in mAP and rank-1 accuracy when the model is transferred from DukeMTMC to Market1501. Similarly, when the model is trained on Market1501 and tested on DukeMTMC, the performance gain becomes 35.0% and 32.3% in rank-1 accuracy and mAP, respectively. Although the progressive unsupervised learning and direct transfer have a huge performance gap, our proposed method can constantly

Table 2. Comparison of various methods on the target domains. When tested on DukeMTMC-reID, Market-1501 is used as the source and vice versa. "Baseline" means directly applying the source-trained model on the target domain. "Baseline+x" means using "x" method upon baseline model. "E" means trained with color equalization augmentation. "C" denotes camera-aware color transformation. "P" stands for progressive learning methods.

Method	DukeMTMC → Market1501				Market1501 → DukeMTMC			
	mAP	Top1	Top5	Top10	mAP	Top1	Top5	Top10
Baseline	20.3	46.5	65.4	72.4	15.9	29.0	45.2	52.3
Baseline + E	28.9	58.4	74.6	80.1	24.4	42.9	58.8	65.2
Baseline + C	30.6	58.2	74.0	80.1	21.9	38.5	55.1	62.3
Baseline + C + E	32.4	61.9	77.8	83.6	25.4	44.7	60.7	67.1
Baseline + P	62.0	82.0	92.7	94.9	54.9	71.8	82.9	86.0
Baseline + E + P	65.5	84.6	93.8	95.6	57.5	75.1	84.6	88.0
Baseline + C + E + P	67.9	86.6	94.5	96.9	60.4	76.0	85.0	88.9

improve the performance under these two settings. With progressive learning, camera-aware color transformation and color equalization augmentation achieve the best overall performance.

4.3 Comparison with State-of-the-arts

In this section, we compare the proposed Second-order Camera-aware Color Transformation (SCCT) with state-of-the-art unsupervised learning methods on Market1501, DukeMTMC in Table 3. SCCT outperforms existing approaches with a clear advantage. In particular, our model outperforms the best published method SSG [21] by almost 9.7% on mAP when testing on Market1501 and DukeMTMC-reID dataset.

Results on Market1501 On Market-1501, we compare our results with state-of-the-art methods including Bag-of-Words (BoW) [25], local maximal occurrence (LOMO) [52], UMDL [53], PUL [54] and CAMEL [55], PTGAN [18], SPGAN [56], TJ-AIDL [41], ARN [57], UDAP [45], MAR [58], PDA-Net [59], PAST [60], SBSGAN [61], CASCL [43] and SSG [21]. Methods that trained on target training set with clustering and pseudo-label (UDAP, PAST, SSG) always obtain higher results than other methods. Therefore, we show the performance of SCCT in two different settings. When tested under direct transfer setting, our SCCT-DIRECT outperforms many complicated state-of-the-art direct transfer methods including TJ_AIDL, SBSGAN, and HHL. TJ_AIDL use additional attribute label information from source data. SBSGAN use additional JPPNet to obtain foreground masks. While SCCT-DIRECT only apply linear camera-aware color transformation and color equalization augmentation. We believe if combined with SCCT, these methods would further boost performance. When

comparing with progressive learning methods, SCCT-PUL achieves rank-1 accuracy = 86.6% and mAP = 67.9%, which outperforms unsupervised version SSG in [21] by 6.6% and 9.6%.

Results on DukeMTMC A similar improvement can also be observed when we test it on the DukeMTMC dataset. Although the camera view discrepancy in DukeMTMC is not as large as Market1501, SCCT can still significantly improve the model performance over the Baseline. Specifically, we achieve mAP = 60.4%, top1 accuracy = 76.0%, top5 accuracy = 85.0% and top10 accuracy = 88.9% by unsupervised learning. Compared with the best unsupervised method, our result is 7.0%/3.0%/4.6%/5.7% higher on mAP and top1/top5/top10 accuracy. Therefore the superiority of our camera-aware color transformation methods with color equalization augmentation can be concluded.

Table 3. Comparison of proposed approach with state-of-the-arts unsupervised domain adaptive person re-ID methods on Market1501 and DukeMTMC dataset.

Method	DukeMTMC → Market1501				Market1501 → DukeMTMC			
	mAP	Top1	Top5	Top10	mAP	Top1	Top5	Top10
LOMO [52]	8.0	27.2	41.6	49.1	4.8	12.3	21.3	26.6
BOW [25]	14.8	35.8	52.4	60.3	8.3	17.1	28.8	34.9
UMDL [53]	12.4	34.5	52.6	59.6	7.3	18.5	31.4	37.4
PTGAN [18]	-	38.6	-	66.1	-	27.4	-	50.7
PUL [54]	20.5	45.5	60.7	66.7	16.4	30.0	43.4	48.5
SPGAN [56]	22.8	51.5	70.1	76.8	26.2	46.4	62.3	68.0
CAMEL [55]	26.3	54.5	-	-	-	-	-	-
SPGAN+LMP [56]	26.7	57.7	75.8	82.4	26.2	46.4	62.3	68.0
TJ_AIDL [41]	26.5	58.2	74.8	81.1	23.0	44.3	59.6	65.0
SBSGAN [61]	27.3	58.5	-	-	30.8	53.5	-	-
HHL [20]	31.4	62.2		84.0	27.2	46.9	61.0	66.7
CASCL [43]	35.5	65.4	80.6	37.8	86.2	59.3	73.2	77.8
ARN [57]	39.4	70.3	80.4	86.3	33.4	60.2	73.9	79.5
MAR [58]	40.0	67.7	81.9	-	48.0	67.1	79.8	-
ECN [19]	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
PDA-Net [59]	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
UDAP [45]	53.7	75.8	89.5	93.2	49.0	68.4	80.1	83.5
PAST [60]	54.6	78.4	-	-	54.3	72.4	-	-
SSG [21]	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
SCCT-DIRECT	32.4	61.9	77.8	83.6	25.4	44.7	60.7	67.1
SCCT-PUL	67.9	86.6	94.5	96.9	60.4	76.0	85.0	88.9

Results on MSMT17 We further validate the effectiveness of our proposed Second-order Camera-aware Color Transformation (SCCT) and color equalization augmentation on MSMT17 [18] dataset as listed in Table 4. Using color

Table 4. Experiments on MSMT17 dataset. "Baseline" denotes using ImageNet image statistics for normalization. "C" stands for using the proposed second-order camera-aware color transformation. "E" means applying color equalization augmentation.

Method	DukeMTMC → MSMT17				Market1501 → MSMT17			
	mAP	Top1	Top5	Top10	mAP	Top1	Top5	Top10
Baseline	5.6	17.0	27.0	32.2	2.8	8.7	15.6	19.7
PTGAN [18]	3.3	11.8	-	27.4	2.9	10.2	-	24.4
ECN [19]	10.2	30.2	41.5	46.8	8.5	25.3	36.3	42.1
TAUDL [62]	12.5	28.4	-	-	-	-	-	-
UTAL [63]	13.1	31.4	-	-	-	-	-	-
Baseline + C	7.1	21.7	32.2	37.8	4.1	12.6	21.1	25.5
Baseline + E	12.1	33.8	48.0	54.0	7.2	21.8	33.8	39.4
SCCT-DIRECT	**13.2**	**37.8**	**51.2**	**57.0**	8.3	23.6	36.0	42.1

transformation can boost the performance by 1.5% on mAP and 4.7% on top1 accuracy comparing to the baseline model when transferring from DukeMTMC to MSMT17. Similarly, it improves 1.3% on mAP and 3.9% on top1 accuracy when transferring from Market1501 to MSMT17. Color equalization boosts the performance even larger on these datasets. On DukeMTMC to MSMT17, it improves the baseline method by 6.9% in mAP and 15.8% on top1 accuracy. On Market1501 to MSMT17, it surpasses baseline by 4.4% on mAP and 13.1% on top1. When color transformation and color equalization are combined, it achieves 13.2%/37.8% on DukeMTMC to MSMT17 and 8.3%/23.6 on Market1501 to MSMT17, which is very significant as it outperforms or achieves similar performances with many state-of-the-art methods (e.g. ECN [19], TAUDL [62], UTAL [63]) with simple image-level color transformation and augmentation.

5 Conclusions

In this work, we proposed camera-aware second-order color transformation for person ReID, which can reduce the discrepancy of source and target data caused by the input distribution, constantly improve performance on direct transfer setting and progressive learning settings. It is a novel input normalization method, which is often neglected by previous unsupervised person ReID methods. It is simple to implement and can be easily combined with many existing state-of-the-art methods. We also investigate the color equalization data augmentation under the unsupervised person ReID setting, which is very effective and can boost the generalization ability of the ReID model. Extensive experimental results demonstrate that the performance of our approach outperforms the state-of-the-arts by a clear margin.

Acknowledgements:

This research is supported by the China NSFC grant (no. 61672446).

References

1. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR. (2006)
2. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR. (2017)
3. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV. (2017)
4. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: ICCV. (2017)
5. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: ECCV. (2016)
6. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: ECCV. (2016)
7. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244 (2016)
8. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR. (2017)
9. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR. (2017)
10. Yi, D., Lei, Z., Li, S.Z.: Deep metric learning for practical person re-identification. arXiv preprint arXiv:1407.4979 (2014)
11. Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., Li, S.Z.: Embedding deep metric for person re-identication A study against large variations. In: ECCV. (2016)
12. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. (2015)
13. Jose, C., Fleuret, F.: Scalable metric learning via weighted approximate rank component analysis. In: ECCV. (2016)
14. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR. (2016)
15. Liao, S., Li, S.Z.: Efficient psd constrained asymmetric metric learning for person re-identification. In: ICCV. (2015)
16. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. In: CVPR. (2017)
17. Zhang, Y., Li, X., Zhao, L., Zhang, Z.: Semantics-aware deep correspondence structure learning for robust person re-identification. In: IJCAI. (2016)
18. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 79–88
19. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identication. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
20. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: ECCV. (2018)
21. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 6112–6121

22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
23. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR. (2018)
24. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
25. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: CVPR. (2015)
26. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: CVPR. (2017)
27. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: CVPR. (2017)
28. Zhou, S., Wang, J., Shi, R., Hou, Q., Gong, Y., Zheng, N.: Large margin learning in set to set similarity comparison for person re-identification. IEEE Transactions on Multimedia (2017)
29. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR. (2015)
30. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR. (2014)
31. Wu, L., Shen, C., Hengel, A.v.d.: Personnet: Person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255 (2016)
32. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
33. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR. (2016)
34. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13** (2012) 723–773
35. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Thirtieth AAAI Conference on Artificial Intelligence. (2016)
36. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. (2015) 1180–1189
37. Chen, C., Fu, Z., Chen, Z., Jin, S., Cheng, Z., Jin, X., Hua, X.S.: Homm: Higher-order moment matching for unsupervised domain adaptation. order **1** (2020) 20
38. Sener, O., Song, H.O., Saxena, A., Savarese, S.: Learning transferrable representations for unsupervised domain adaptation. In: Advances in Neural Information Processing Systems. (2016) 2110–2118
39. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 2988–2997
40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232
41. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2275–2284
42. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020)

43. Wu, A., Zheng, W.S., Lai, J.H.: Unsupervised person re-identification by camera-aware similarity consistency learning. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 6922–6931
44. Lan, X., Zhu, X., Gong, S.: Universal person re-identification (2019)
45. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognition (2020) 107173
46. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. (2015)
47. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8295–8302
48. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Volume 96. (1996) 226–231
49. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1318–1327
50. Zhun Zhong, Liang Zheng, G.K.S.L.Y.Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
51. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
52. Shengcai Liao, Yang Hu, X.Z., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. (2015)
53. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1306–1315
54. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14** (2018) 1–18
55. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE international conference on computer vision. (2017) 994–1002
56. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: 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) 994–1003
57. Li, Y.J., Yang, F.E., Liu, Y.C., Yeh, Y.Y., Du, X., Frank Wang, Y.C.: Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2018) 172–178
58. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2148–2157
59. Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.C.F.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7919–7929
60. Zhang, X., Cao, J., Shen, C., You, M.: Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 8222–8231

61. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Sbsgan: Suppression of inter-domain background shift for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 9527–9536
62. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. In: Proceedings of the European conference on computer vision (ECCV). (2018) 737–753
63. Li, M., Zhu, X., Gong, S.: Unsupervised tracklet person re-identification. IEEE transactions on pattern analysis and machine intelligence (2019)