Local Soft Attention Joint Training and Dual Cross-neighbor Label Smoothing for Unsupervised Person Re-identification

Qing Han[1], Longfei Li[1], Weidong Min[1,2,3]*, Qi Wang[1], Qingpeng Zeng[1], Shimiao Cui[1], Jiongjin Chen[1]
[1]School of Mathematics and Computer Science, Nanchang University, Nanchang, China
[2]Institute of Metaverse, Nanchang University, Nanchang, China
[3]Jiangxi Key Laboratory of Smart City, Nanchang, China
hanqing@ncu.edu.cn, lilongfei@email.ncu.edu.cn, minweidong@ncu.edu.cn, wangqi@ncu.edu.cn
zengqingpeng@ncu.edu.cn, 406100210099@email.ncu.edu.cn, 416100210351@email.ncu.edu.cn

Abstract

Existing unsupervised person re-identification approaches fail to fully capture the fine-grained features of local regions, resulting in that some persons with similar appearances and different identities have been given the same label after clustering. Then, the identity-independent information contained in different local regions will lead to different levels of local noise. To address the above challenges, Local Soft Attention Joint Training and Dual Cross-neighbor Label Smoothing are proposed in this paper. Firstly, the whole joint training is divided into global and local parts, and then the soft attention mechanism in the local branch is proposed to accurately capture the subtle differences in local regions, which improves the ability of the Re-ID model to identify local significant parts of the person. Secondly, Dual Cross-neighbor Label Smoothing (DCLS) is designed to progressively mitigate label noise in different local regions. The DCLS realizes the semantic alignment between the global and local regions of the person by the global and locals similarity metric, and then further establishes the proximity association between local regions by the cross information of neighbor regions, which achieves the label smoothing of the global and locals throughout the training process. Extensive experiments show that the proposed method outperforms the existing work under unsupervised settings on several standard person re-identification datasets.

Keywords: ***Person re-identification, unsupervised learning, local soft attention joint training, dual cross-neighbor label smoothing***

*Corresponding author

1. Introduction

Person re-identification (Re-ID) aims to match query person images with person images of the same identity across non-overlapping camera views. It and vehicle Re-ID play a vital role in areas, such as intelligent video surveillance, intelligent security, and intelligent transport [15, 46, 47]. In recent years, unsupervised person Re-ID has attracted much attention because it does not require massive manually labeled data. Existing unsupervised person Re-ID approaches can be divided into unsupervised domain adaptation (UDA) [1, 27, 28, 39] and unsupervised learning (USL) [9, 26, 31, 49] Re-ID; the former focuses on learning from existing labeled data and transferring the knowledge learned to unlabeled data [11]. Compared with UDA, USL Re-ID is more practical because it abandons the dependence on existing labeled data and focuses on the mining of unlabeled data information.

Existing unsupervised person Re-ID methods [52, 55] mainly adopt clustering to generate labels for unlabeled data and then train the model in a supervised manner, but this method depends on good clustering results. Some existing approaches [12, 32, 54] optimize clustering results by combining global and local training strategies. Despite their effectiveness, they more or less ignore two vital factors during this process. (1) The inadequate capture of the local fine-grained features interferes with the clustering result. Clustering results depend on the metric ranking of the global features extracted by the model. However, some approaches ignore the capture of the fine-grained feature in local regions, thus affecting the entire model to learn discriminant feature embedding from data. Some persons with similar appearances and different identities have been given the same label after clustering, directly affecting the training and final results of the model. (2) The identity-independent information contained in different local regions lead to different levels of local noise. Given the changes in lighting, background environment, and human posture, even the

features of the same person's data are also changed. The noise in different local regions interferes with the extraction of global features, leading to a large gap between features within the class, which introduces noise into the label of the global feature after clustering. The errors of the noisy label accumulate in the training process and hinder the representation learning of the global person data, thus affecting the accuracy of the model.

On the basis of the above discussion, this study attempts to address the above two challenges to achieve robust unsupervised person Re-ID. Hence, a method combining local soft attention joint training and dual cross-neighbor label smoothing (DCLS) is proposed to solve the unsupervised person Re-ID issues. The contributions of this study are summarized as follows:

1. We adopt the strategy of global and local joint training. The soft attention mechanism in the local branch is proposed to capture the subtle differences in local regions accurately, thereby improving the ability of the Re-ID model to identify local regions.

2. DCLS is designed to mitigate label noise in different local regions progressively. DCLS realizes the semantic alignment between the global and local regions of the person by the global and local similarity metrics and then further establishes the proximity association between local regions by the cross information of neighbor regions, thus achieving the label smoothing of the global and locals throughout the training process.

3. The proposed method is evaluated on several standard person Re-ID datasets. Extensive experiments show that the proposed method outperforms the existing work under unsupervised settings and demonstrates the effectiveness of the proposed method.

The rest of this paper is organized as follows: In Section 2, related work is discussed. Section 3 is divided into four subsections. Section 3.1 presents an overview of the proposed framework. Section 3.2 describes the local soft attention mechanism. Sections 3.3 and 3.4 describe the DCLS and loss function, respectively. The experiments are shown in Section 4. The conclusion is indicated in Section 5.

2. Related work

Unsupervised Re-ID can be divided into two categories: UDA and USL Re-ID. UDA tasks focus on learning from existing labeled data, and some UDA methods narrow the data distribution difference between source and target domains by feature alignment [20, 29] and image style transfer [5, 7, 45, 50, 66].

Although UDA and USL Re-ID are trained on different data conditions, their mainstream methods both use clustering-based learning strategies. Currently, clustering-based methods [13, 14] achieve state-of-the-art results due to improvements in label quality. To make the clustering results align with the real-world distribution, Lin et al. [30] proposed a bottom-up clustering (BUC) approach to optimize jointly a CNN model and the relationship among the individual samples. To deal with data imbalance, DBC [8] learns the data distribution and then uses pairwise sample relationships to achieve improved cluster balance in the clustering process. Zeng et al. [56] proposed a method that combines hierarchical clustering with hard-batch triplet loss (HCT); the method maximizes the similarity among samples in the target dataset by hierarchical clustering and reduces the influence of hard examples by HCT to generate high-quality labels. Lin et al. [32] proposed softened similarity learning (SSL) to soften labels by computing similarities that introduce camera labels and fine-grained details. Ge et al. [13] proposed the mutual mean-teaching method, which provides more reliable soft labels by learning from each other through two networks. SpCL [14] proposed a reliability criterion for measuring cluster independence and compactness to guarantee clustering quality.

In recent studies, supervised methods [44, 60] of joint local feature training have achieved state-of-the-art performance. Sun et al. [40] proposed a part-based convolutional baseline network to extract local features. Zheng et al. [60] proposed a pyramid model, which incorporates local and global features that can match at different scales and then search for the correct image of the same identity even when the image pairs are not aligned. Wang et al. [44] designed the multiple granularity network and uniformly partitioned the images into several stripes by varying the number of parts in different local branches to obtain local feature representations with multiple granularities. Given the remarkable achievements of these methods in supervised settings, some works [12, 17] have attempted to apply joint local feature learning to unsupervised person Re-ID. Fu et al. [12] proposed a self-similar grouping method to cluster global and local features separately and supervise model training jointly. Subsequently, some works exploited local features to achieve impressive performance on unsupervised person Re-ID tasks.

3. Method

3.1. Overview of the proposed method

Figure 1 shows an overview of the proposed local soft attention joint training and DCLS method. Specifically, (1) the network structure is adopted to extract global features and the corresponding local features; (2) the global features are clustered by the DBSCAN [10] algorithm, and the label generated by clustering are shared by local features; (3) the initial smoothing of the local label is dynamically guided by global and local similarity measures, and then the further smoothing of the local label is guided by the cross information between neighbor locals; the smoothing of the

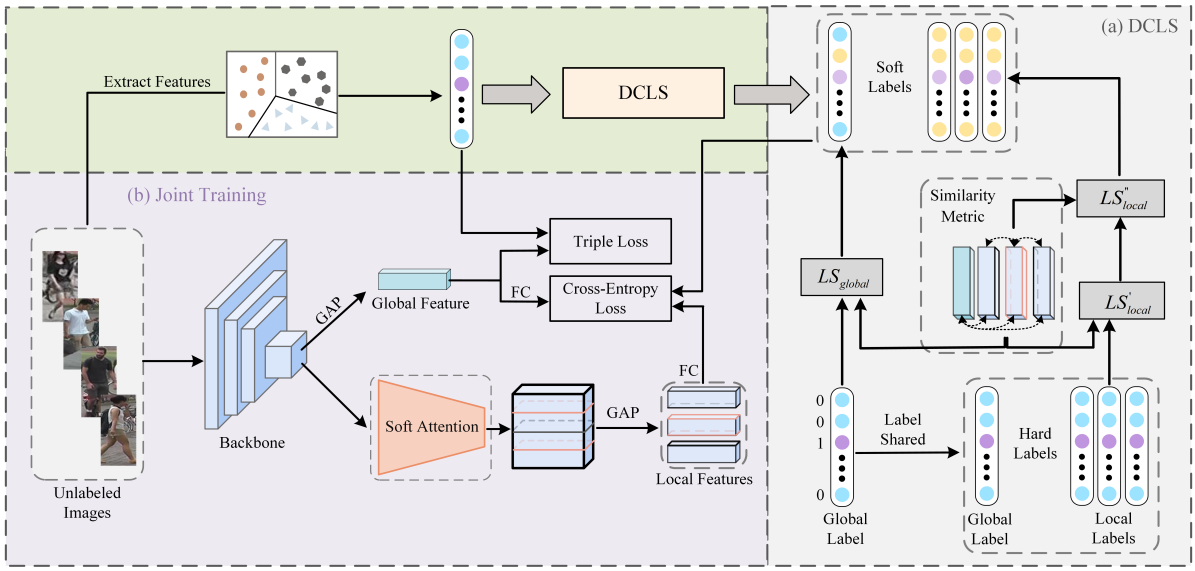

Figure 1. Overview of our proposed method. The overall process is divided into two parts. The first part is DCLS, which assigns the clustered hard label to the local, and then the global and local label smoothing by similarity metric. The second part is the global and local joint training, and the soft attention mechanism is proposed in local branches. The soft labels processed by DCLS are used as the supervision signals of each branch to train the model jointly.

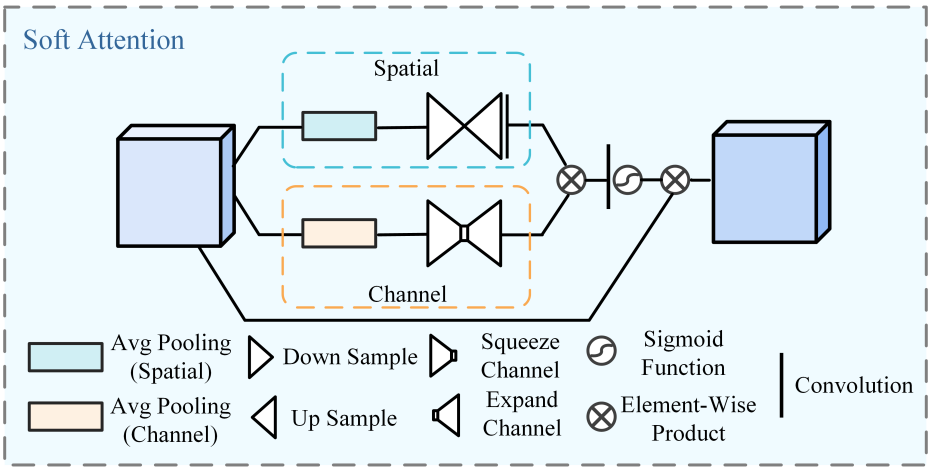

Figure 2. Soft attention mechanism.

global label is guided by the learning of local regions; (4) the global and local labels are simultaneously used as supervision signals to train the model jointly.

3.2. Local soft attention joint training

The framework of the proposed method is trained in a global and local joint training way and follows the settings in existing methods [14, 30, 56]. The ResNet-50 [16] model is used as the backbone to extract global features, which participate in clustering and subsequent label smoothing [41]. For local feature extraction, on the basis of the global feature map, the soft attention mechanism in the local branch is proposed to capture the subtle differences in local regions accurately, thereby improving the ability of the Re-ID model to identify the key local parts of the person. Figure 2 shows the soft attention mechanism, which combines spatial attention and channel attention. Spatial attention uses the spatial relationship of features to generate the spatial attention map, which is complementary to the channel attention [51]. The feature map displays the regions with prominent feature information by the global average pooling operation on the channel dimension [23]. Then, the high-level feature information of the feature graph is captured by downsampling and restored to its size by upsam-

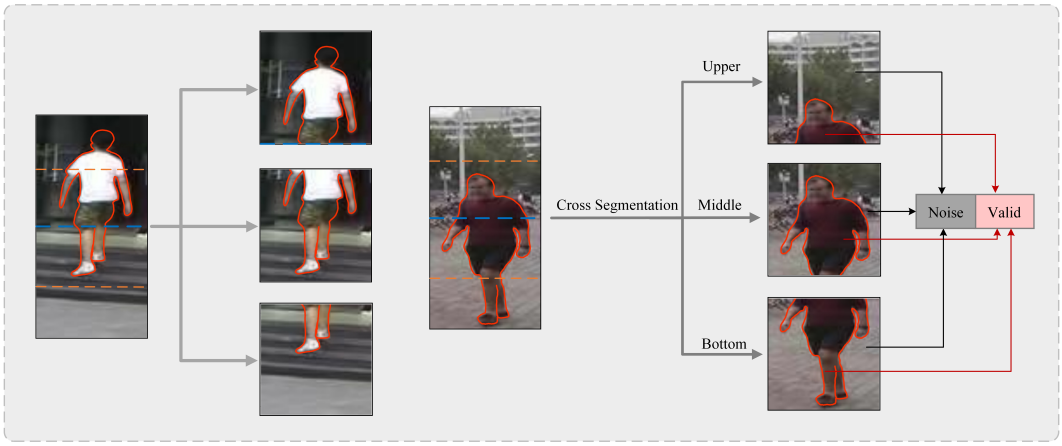

Figure 3. Cases with noise in different local regions. The inner part of the red line represents the valid region, and the outer part of the red line represents the noise region.

pling. Finally, the feature map by the convolution layer to obtain the spatial attention map $F_s \in \mathbb{R}^{1\times H\times W}$ to integrate with the channel attention map. For channel attention, the feature graph first aggregates the spatial feature information by global average pooling on the height and width dimensions to generate the spatial context descriptor $G^c(F)$, which denotes average-pooled features, where F indicates input features. Then, the relationship between channels is fully captured by the excitation operation of squeezing and expanding channels to obtain the channel attention map $F_c \in \mathbb{R}^{C\times 1\times 1}$. In summary, channel attention is calculated as Eq. (1):

$$F_c = W_1^c\left(W_0^c\left(G^c(F)\right)\right) \tag{1}$$

where $W_1^c \in \mathbb{R}^{C/r\times C}$ and $W_0^c \in \mathbb{R}^{C\times C/r}$ are the parameters for squeezing and expanding channels, channel weights are generated by W_1^c and W_0^c, and $r = 16$ is the reduction ratio. The ReLU activation function is followed by W_1^c and W_0^c. Then, spatial attention and channel attention are combined and calculated as Eq. (2):

$$F = F_s \otimes F_c \tag{2}$$

where $F \in \mathbb{R}^{C\times H\times W}$ represents the weight map, and $\otimes$ represent the element-wise product. Then, based on the complementary relationship between spatial and channel attention, the weight map combines spatial and channel attention by the convolution layer, and the combined weight map is normalized by the sigmoid activation function to obtain the final weight map of soft attention, which is fused with the input feature map.

Under the effect of the soft attention mechanism, the global feature map is divided into multiple local feature maps, and neighbor feature maps are ensured to have overlapping regions. This split method enables the local cross information to play a guiding role in the subsequent work of DCLS and can improve the performance of the model. The local feature map obtains local feature by global average pooling, which participates in the subsequent label smoothing work. The global and locals simultaneously participate in the training of the model.

3.3. DCLS

Although the soft attention mechanism can capture the discriminative features of local regions, the local regions of some personal data contain too little discriminative information, and substantial information noise becomes unrelated to identity, which is inevitable. Figure 3 shows cases with noise in different local regions. The identity-independent information contained in different local regions leads to different levels of local noise. This noise interferes with the extraction of global features, leading to a large gap between features within the class, which introduces noise into the label of the global feature after clustering. To solve this issue, DCLS is designed to mitigate label noise in different local regions progressively, thereby achieving the label smoothing of the global and locals throughout the training process and strengthening the representation learning of global person data. The label smoothing [33] is calculated as Eq. (3):

$$y' = y\left(1-\beta\right) + \frac{\beta}{K} \tag{3}$$

where $\beta \in (0, 1)$ represents the smoothing factor of the label, and K represents the number of image categories. Then, the label smoothing factor of each local is dynamically determined by the similarity metric. Specifically, the square Euclidean distance algorithm is introduced to measure the feature similarity, the measurement is calculated as

Eq. (4):

$$d(x,y) = \left\| \sum_{i=1}^{D} (x_i - y_i)^2 \right\|^2 \tag{4}$$

where $||\cdot||$ represents the Euclidean distance, D represents the feature dimension, and (x_i, y_i) represents the coordinates of feature x and feature y in the feature space. To obtain the label smoothing factor of each local, the distance is normalized as Eq. (5):

$$\beta_{x,y} = e^{-d(x,y)} \in (0,1) \tag{5}$$

the value of $\beta_{x,y}$ is negatively related to the distance between features. DCLS realizes the semantic alignment between the global and local regions of the person by the global and local similarity metrics to obtain the smoothing factor $\beta_{i,g}$ of each local label. The factors are inserted into Eq. (3), and the local label initial smoothing is calculated as Eq. (6):

$$y_i^{'} = y\left(1 - \beta_{i,g}\right) + \frac{\beta_{i,g}}{K} (i = 1,2,3) \tag{6}$$

where the value of i corresponds to local features, and g represents the global feature. As the training continues to advance, the representation of global and local features improves, their distance in the feature space gradually increases, and the label smoothing factor $\beta_{i,g}$ of each local gradually decreases. Given the prominent feature representation, the local label retains great credibility, that is, the smoothness of the label is low; Eq. (6) can well show this change. If a local feature is close to the global feature in the feature space, its similarity with the global feature is relatively high, and the corresponding label reliability is relatively high. However, the formula does not consider this detail, and only the global and local similarity metrics guide the local label smoothing neglecting the local context, which may not be conducive to model training. To solve this issue, the DCLS further establishes the proximity association between local regions by the cross information of neighbor regions, thereby achieving the further smoothing of local labels. Specifically, the further smoothing of local labels is guided by the local and neighbor local similarity metrics and is calculated as Eq. (7):

$$y_i^{''} = \begin{cases} y_i^{'}\left(1 - \beta_{i,j}\right) + y_j^{'}\beta_{i,j} (i = 1,3; j = 2) \\ y_i^{'}\left(1 - W_l\right) + \sum_{n=1}^{1,3} \dfrac{y_n^{'} W_l \beta_{i,n}}{\sum_{j=1}^{1,3} \beta_{i,j}} (i = 2) \end{cases} \tag{7}$$

where the values of i, j and n correspond to local features, and $W_l \in [0,1]$ is a weight super parameter. The formula can well adjust the label smoothness of each local, that is, local features with high similarity to the global feature have high label reliability, and vice versa. Then, the learning of local regions is combined to guide global label smoothing and is calculated as Eq. (8):

$$y_g^{'} = y\left(1 - W_g\right) + \sum_{i=1}^{3} \frac{W_g \beta_{i,g} \operatorname{Softmax}\left(\hat{y}_i\right)}{\sum_{j=1}^{3} \beta_{j,g}} \tag{8}$$

where the values of i and j correspond to local features, and $W_g \in [0,1]$ is a weight super parameter. $\operatorname{Softmax}(\cdot)$ is a softmax function used to normalize the input element value, and $\hat{y}_i$ is the output value of the model, representing the probability classification of the corresponding local feature.

3.4. Loss function

The cross-entropy and triple loss commonly used in person Re-ID tasks are used as the loss function of the model in the training stage. The formula of cross entropy is calculated as Eq. (9):

$$L = -y \log(\hat{y}) \tag{9}$$

where $\hat{y}$ indicates the predicted probability distribution. Then, the local label after Eq. (7) smoothing is inserted into Eq. (9) to obtain the cross-entropy loss of the local features and is calculated as Eq. (10):

$$L_i = -\left[y\left(1 - R_i\right) + \frac{R_i}{K} \right] \log\left[\operatorname{Softmax}\left(\hat{y}_i\right)\right] \tag{10}$$

$$R_i = \begin{cases} \beta_{i,g} - \beta_{i,g}\beta_{i,j} + \beta_{j,g}\beta_{i,j} (i = 1,3; j = 2) \\ \beta_{i,g}\left(1 - W_l\right) + \sum_{n=1}^{1,3} \dfrac{\beta_{n,g} W_l \beta_{i,n}}{\sum_{j=1}^{1,3} \beta_{i,j}} (i = 2) \end{cases}$$

similarly, the global label after Eq. (8) smoothing is inserted into Eq. (9) to obtain the cross-entropy loss of the global feature and is calculated as Eq. (11):

$$L_g = -\left[y\left(1 - W_g\right) + \sum_{i=1}^{3} \frac{W_g \beta_{i,g} \operatorname{Softmax}\left(\hat{y}_i\right)}{\sum_{j=1}^{3} \beta_{j,g}} \right] \times \log\left[\operatorname{Softmax}\left(\hat{y}_g\right)\right] \tag{11}$$

where $\hat{y}_g$ represents the classification probability value of the global feature. The triple loss only applies to global feature.

4. Experiments

4.1. Datasets and Evaluation protocols

The proposed method is evaluated on three standard person re-ID datasets, namely, Market-1501 [63], DukeMTMC-reID [37] and MSMT17 [50]. Market-1501 contains 32,688 images of 1,501 person identities captured by six cameras; these images are divided into 12,936 training images of 751 person identities and 19,732 test images of 750 person identities, with 3,368 test images used as queries. DukeMTMC-reID contains 34,183 images of

Methods	Market-1501					DukeMTMC-reID				
	Source	mAP	Rank1	Rank5	Rank10	Source	mAP	Rank1	Rank5	Rank10
Unsupervised Learning (USL)										
SSL [32]	None	37.8	71.7	83.8	87.4	None	28.6	52.5	63.5	68.9
BUC [30]	None	38.3	66.2	79.6	84.5	None	27.5	47.4	62.6	68.4
DBC [8]	None	41.3	69.2	83.0	87.8	None	30.0	51.5	64.6	70.1
MMCL [43]	None	45.5	80.3	89.4	92.3	None	40.2	65.2	75.9	80.0
JVTC [24]	None	47.5	79.5	89.2	91.9	None	50.7	74.6	82.9	85.3
MPRD [21]	None	51.1	83.0	91.3	93.6	None	43.7	67.4	78.7	81.8
HCT [56]	None	56.4	80.0	91.6	95.2	None	50.7	69.6	83.4	87.4
GCL [4]	None	66.8	87.3	93.5	95.5	None	62.8	82.9	87.1	88.5
SpCL [14]	None	72.5	87.8	94.7	96.3	None	61.4	77.1	86.7	89.6
HHCL [19]	None	78.7	90.4	96.4	97.5	None	67.1	80.7	89.7	92.1
ICE [3]	None	82.4	92.7	97.6	98.4	None	66.2	80.3	88.9	90.9
Trans-SSP [35]	None	87.3	94.2	**98.0**	**98.5**	None	71.4	82.7	89.9	92.5
Ours(*w/o* RK)	None	81.4	92.4	97.3	98.2	None	67.6	81.1	**90.4**	92.7
Ours	None	**91.2**	**94.3**	96.6	97.1	None	**80.7**	**85.0**	90.1	**92.8**
Unsupervised Domain Adaptation (UDA)										
ECN [25]	Duke	43.0	75.1	87.6	91.6	Market	40.4	63.3	75.8	80.4
JVTC [24]	Duke	67.2	86.8	95.2	97.1	Market	66.5	80.4	89.9	92.2
AD-Cluster [53]	Duke	68.3	86.7	94.4	96.5	Market	54.1	72.6	82.5	85.5
HGA [58]	Duke	70.3	89.5	93.6	95.5	Market	67.1	80.4	88.7	90.3
MMT [13]	Duke	71.2	87.7	94.9	96.9	Market	65.1	78.0	88.8	92.5
Mixup [34]	Duke	71.5	88.1	94.4	96.2	Market	65.2	79.5	88.3	91.4
NRMT [59]	Duke	71.7	87.8	94.6	96.5	Market	62.2	77.8	86.9	89.5
DCML [2]	Duke	72.6	87.9	95.0	96.7	Market	63.3	79.1	87.2	89.4
GCL [4]	Duke	75.4	90.5	96.2	97.1	Market	67.6	81.9	88.9	90.6
MEB-Net [57]	Duke	76.0	89.9	96.0	97.5	Market	66.1	79.6	88.3	92.2
UNRN [61]	Duke	78.1	91.9	96.1	97.8	Market	69.1	82.0	**90.7**	**93.5**
GLT [62]	Duke	79.5	92.2	96.5	97.8	Market	69.2	82.0	90.2	92.8
Ours(*w/o* RK)	Duke	81.4	92.2	**97.4**	**98.2**	Market	68.2	81.0	90.4	93.1
Ours	Duke	**91.2**	**93.8**	96.5	97.4	Market	**80.7**	**84.9**	90.6	93.4

Table 1. Comparison with the state-of-the-art unsupervised Re-ID methods on Market1501 and DukeMTMC-reID datasets.

1,404 person identities captured by eight cameras; these images are divided into 16,522 training images of 702 person identities and 17,661 test images of 702 person identities, with 2,228 test images used as queries. MSMT17 is the largest person Re-ID dataset, which contains 126,411 images of 4,101 person identities captured by 15 cameras; these images are divided into 32,621 training images of 1,041 person identities and 93,820 test images of 3,060 person identities, with 11,659 test images used as queries. Mean average precision (mAP) and cumulative match characteristic (CMC) Rank-1/5/10 accuracy are used to evaluate the performance of the proposed method on three standard datasets. In the experiments, the re-ranking [64](RK) technique is adopted.

4.2. Implementation details

Following the settings in existing methods [8, 14, 30, 32, 56], the ResNet-50 [16] model pretrained on ImageNet [6] as the backbone. During training, random flipping, random cropping, and random erasing [65] are used for data augmentation, using the Adam [22] optimizer with weight decay of 5×10^{-4}. We set the mini-batch size as 32, consisting of randomly selected eight pseudo-classes and four instance images of each class. The initial learning rate is set to 3.5×10^{-4} and is reduced by a factor of 10 at 40 and 70 epochs, for a total of 80 epochs. In each epoch, the model is trained for 400 iterations. The DBSCAN [10] method clusters the training dataset before each epoch. During testing, only global features are used for evaluation.

4.3. Comparison with state-of-the-art

To verify the performance of the proposed method further, several state-of-the-art unsupervised person Re-ID methods are compared on the Market-1501, DukeMTMC-reID, and MSMT17 datasets. Table 1 shows the results of the proposed method and state-of-the-art methods on

Methods	MSMT17				
	Source	mAP	R1	R5	R10
Unsupervised Learning (USL)					
AE [9]	None	8.5	26.6	37.0	41.7
MMCL [43]	None	11.2	35.4	44.8	49.8
MPRD [21]	None	14.6	37.7	51.3	57.1
JVTC [24]	None	17.3	43.1	53.8	59.4
SpCL [14]	None	19.1	42.3	55.6	61.2
GCL [4]	None	21.3	45.7	58.6	64.5
Ours(*w/o* RK)	None	27.6	52.8	66.0	**71.5**
Ours	None	**39.0**	**59.2**	**67.4**	**71.5**
Unsupervised Domain Adaptation (UDA)					
ECN [25]	Market	8.5	25.3	36.3	42.1
AE [9]	Market	9.2	25.5	37.3	42.6
NRMT [59]	Market	19.8	43.7	56.5	62.2
Mixup [34]	Market	20.4	43.7	56.1	61.9
DG-Net++ [67]	Market	22.1	48.4	60.9	66.1
MMT [13]	Market	22.9	49.2	63.1	68.8
JVTC [24]	Market	25.1	48.6	65.3	68.2
UNRN [61]	Market	25.3	52.4	64.7	69.7
GCL [4]	Market	27.0	51.1	63.9	69.9
Ours(*w/o* RK)	Market	28.3	53.7	67.0	**72.5**
Ours	Market	**39.9**	**59.6**	**68.3**	72.3
ECN [25]	Duke	10.2	30.2	41.5	46.8
AE [9]	Duke	11.7	32.3	44.4	50.1
NRMT [59]	Duke	20.6	45.2	57.8	63.3
DG-Net++ [67]	Duke	22.1	48.8	60.9	65.9
MMT [13]	Duke	23.3	50.1	63.9	69.8
Mixup [34]	Duke	24.3	51.7	64.0	68.9
UNRN [61]	Duke	26.2	54.9	67.3	70.6
JVTC [24]	Duke	27.5	52.9	70.5	**75.9**
GCL [4]	Duke	29.7	54.4	68.2	74.2
Ours(*w/o* RK)	Duke	30.4	57.4	70.0	75.2
Ours	Duke	**42.7**	**63.2**	**71.3**	75.3

Table 2. Comparison with the state-of-the-art unsupervised Re-ID methods on MSMT17 dataset.

Methods	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
w/ DCLS	79.6	91.3	66.3	79.4
+ECA [48]	78.9	90.8	66.0	79.8
+CBAM [51]	79.8	91.5	65.5	79.2
+Soft Attention	**81.4**	**92.4**	**67.6**	**81.1**

Table 3. Ablation study on different attention mechanisms.

Market-1501 and DukeMTMC-reID. Table 2 shows the results of the proposed method and state-of-the-art methods on MSMT17.

Under the USL setting, we evaluate the performance of the proposed method. On Market-1501, Compared with HHCL [19], our method(*w/o* RK) improves Rank-1 accuracy by 2.0% and mAP by 2.7%. On DukeMTMC-reID, Compared with ICE [3] and HHCL [19], our method(*w/o* RK) improves Rank-1 accuracy by 0.8% and 0.4% and mAP by 1.4% and 0.5%, respectively, which is slightly lower than GCL [4] in Rank-1 accuracy but 4.8% higher mAP than it. On MSMT17, our method(*w/o* RK) achieves 52.8% Rank-1 accuracy and 27.6% mAP, which are respectively 7.1% and 6.3% higher than GCL [4]. After implementing reranking [64](RK), our method outperforms all baselines by a large margin on Market-1501 and DukeMTMC-reID. Compared to the currently best published method TransReID-SSL [35], our method surpasses it by 3.9% on Market-1501 and 9.3% on DukeMTMC-reID in mAP.

Under the UDA setting, we also evaluate the performance of the proposed method. Compared to the current best method GLT [62], our method(*w/o* RK) improves mAP by 1.9% on Market-1501, which is slightly below it on DukeMTMC-reID but also surpasses other methods. On MSMT17, our method(*w/o* RK) surpasses other methods by a large margin, regardless of whether the source do-

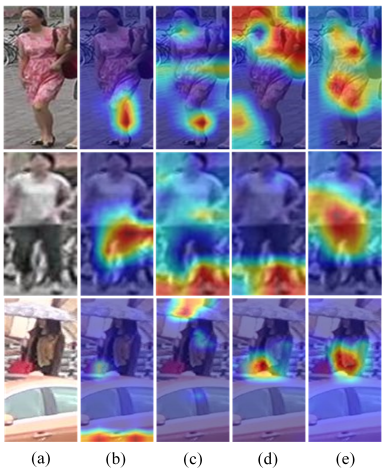

Figure 4. Visualized of the global features under the influence of different attention mechanism modules based on Grad-CAM [38]: (a) Original images; (b) without attention; (c) with ECA [48]; (d) with CBAM [51]; (e) with Soft Attention(Ours).

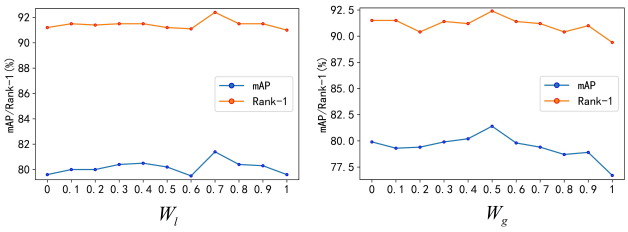

Figure 5. Sensitivity Analysis of Hyper-parameter W_l and W_g on Market-1501 dataset.

Methods	Market		Duke	
	mAP	R1	mAP	R1
w/ Soft Attention	78.4	90.6	63.4	76.7
$+LS_l^{'}$	79.3	91.2	65.1	78.6
$+LS_l^{''}$	79.9	91.4	66.1	78.8
$+LS_g$	79.6	90.8	65.5	79.4
$+LS_l^{'} + LS_g$	80.1	91.5	66.5	79.2
$+LS_l^{''} + LS_g$	**81.4**	**92.4**	**67.6**	**81.1**
$+LS_l^{''} + LS_g$(non-cross)	79.6	91.3	64.9	79.5

Table 4. Ablation study on different label smoothing methods.

N_l	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
0	73.4	88.0	62.6	77.6
2	79.6	91.2	67.1	79.7
3	**81.4**	**92.4**	**67.6**	**81.1**

Table 5. Evaluation results of different number of local features.

main dataset is Market-1501 or DukeMTMC-reID. After implementing reranking [64](RK), the performance of our method by large margins.

4.4. Ablation study

Effectiveness of soft attention mechanism. To verify the effectiveness of the soft attention mechanism, two attention mechanisms, i.e., CBAM [51] and ECA [48], are compared. CBAM integrates channel attention and spatial attention and shows that the best results can be achieved by using the serialization method of channel attention first and then spatial attention. ECA is an improved channel attention mechanism based on the SE [18] mechanism. It achieves light weight and high efficiency by avoiding dimensionality reduction and cross-channel interaction. The experimental results of different attention mechanisms under the same settings are shown in Table 3. As shown in the table, soft attention has considerable performance advantages. Soft attention aims to improve the discriminative properties of the whole model for fine-grained information so that the global features that are noticed to be more valid information play a good role in the clustering and evaluation stages. In this regard, to better illustrate the explanation, the visualization of global features based on different attention mechanisms is shown in Figure 4. The figure shows that compared with the other two attention mechanisms, soft attention can extract more fine-grained local information, thus improving the ability of the Re-ID model to identify the important local parts of the person.

Analysis of hyper-parameters. The effect of the number of local features N_l and hyper-parameters W_l, W_g on performance is explored. Under the same settings, the experimental results of their different values are shown in Table 5 and Figure 5. When W_l is set to 0, the middle local label only uses the information between it and the global feature for smoothing and is not guided by the cross information between it and the neighbor local regions, resulting in inaccurate label optimization, thus limiting the performance. When W_l is set to 1, the middle local label is completely guided by the cross information between it and the neighbor local regions for smoothing, but the neighbor local regions contain information that the middle local feature does not have, leading to false guidance. When W_g is set to 0, the global label is not smoothed, limiting the performance. When W_g is set to 1, the global label smoothing is completely guided by combining the learning of local regions, showing a considerable performance drop. Therefore, according to the above experimental results, we set N_l to 3, W_l to 0.7, and W_g to 0.5.

Metrics	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
Euclidean	**81.4**	**92.4**	67.6	**81.1**
Cosine	80.1	91.7	**68.6**	80.9

Table 6. Evaluation results of different distance metrics.

Effectiveness of DCLS. To verify the effectiveness of DCLS, Table 4 reports the experimental results of ablation for various label smoothing methods. In the table, "LS_g" denotes the global label smoothing, "$LS_l^{'}$" denotes the local label initial smoothing, and "$LS_l^{''}$" denotes the local label further smoothing. As shown in the table, each label smoothing method can improve the performance to a certain extent; especially when local and global label smoothing are conducted simultaneously, improved labels can be obtained, thus considerably improving the performance. On this basis, compared with the guidance using non-cross local features, the way of using cross-information guidance improves mAP by 1.8% and 2.7%, verifying the vital of cross information between local regions. To intuitively analyze the DCLS, Figure 6 visualize the embeddings of the global feature of our method with and without DCLS on Market-1501 dataset. The comparison in the figure shows that in the absence of DCLS, the feature points of different identities are embedded too close to each other. On the contrary, they are well separated in the embedding space under the action of DCLS. This shows that DCLS can effectively strengthening the representation learning of global person data.

In DCLS, the degree of global and local label smoothing depends on the similarity metric between features. In the main paper, the squared euclidean distance is used to metric the similarity of features. The cosine distance is another common metric used in person Re-ID and is calculated as

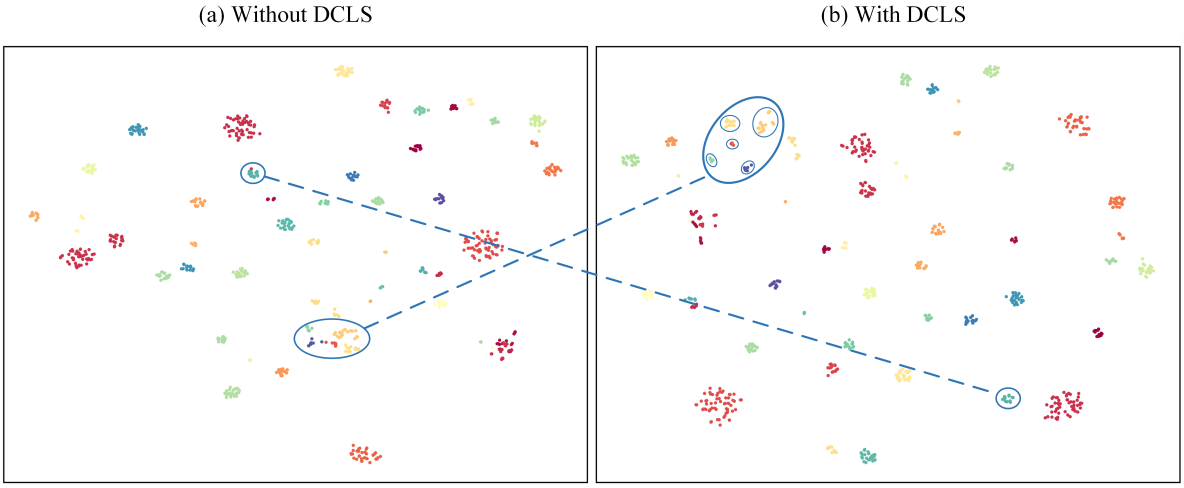

Figure 6. T-SNE [42] visualization of the learned global feature embeddings on the Market-1501 training dataset (Random 50 identities). Points of the same color represent images of the same identity.

Backbones	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
ResNet50	81.4	92.4	67.6	81.1
IBN-ResNet50	**84.0**	**92.8**	**70.4**	**82.1**

Table 7. Evaluation results of different backbones.

Eq. (12):

$$C(x,y) = 1 - \frac{\sum_{i=1}^{D} x_i y_i}{\sqrt{\sum_{j=1}^{D} x_j^2}\sqrt{\sum_{k=1}^{D} y_k^2}} \tag{12}$$

where D represents the feature dimension, and (x_i, y_i) represents the coordinates of feature x and feature y in the feature space. To better understand the sensitivity of DCLS to measurement methods, Table 6 shows the performance of the proposed method under different measurement methods. On the whole, the squared euclidean distance is more popular with DCLS, but no matter which measurement method, the proposed method shows a competitive performance.

Instance-batch normalization(IBN) [36] has been proven more effective than batch normalization(BN) in both supervised and UDA Re-ID tasks. Table 7 compares the performance of our method under the ResNet50 and IBN-ResNet50 backbones. With IBN-ResNet50, the performance of our method can be further improved.

Label smoothing [41] is a common technique within various domains, and to further verify the effectiveness of DCLS, Table 8 shows the performance of DCLS and constant label smoothing (CLS). CLS means using a constant smoothing factor (default is 0.1) in Eq. (3). From the table, DCLS shows a better performance than CLS. CLS only blindly adjusts the label distribution without considering the correlation between features, and the improvement is limited. DCLS reliably smoothes global and local labels by the similarity measure between features and achieves better performance.

Methods	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
w/ Soft Attention	78.4	90.6	63.4	76.7
$+CLS_l$	79.8	90.7	64.9	78.7
$+LS_l^{''}$	**79.9**	**91.4**	**66.1**	**78.8**
$+CLS_g$	79.4	90.7	64.1	78.5
$+LS_g$	**79.6**	**90.8**	**65.5**	**79.4**
$+CLS_l + CLS_g$	80.3	91.3	66.6	79.3
$+LS_l^{''} + CLS_g$	80.8	91.6	66.2	79.4
$+LS_g + CLS_l$	80.5	91.5	67.1	80.4
$+LS_l^{''} + LS_g$	**81.4**	**92.4**	**67.6**	**81.1**

Table 8. Comparison of DCLS and constant label smoothing (CLS) on Market1501 and DukeMTMC-reID datasets.

To present the experimental results clearly, Figure 7 shows the top 10 visualization results of the different modules' rank-list on the Market-1501 dataset. Images in the green boxes denote correct matches, whereas the red box represents the wrong match. The above three examples show that when the soft attention mechanism and DCLS are used together, the correct matching hit rate is high, especially the first hit, which is crucial to the person Re-ID task.

5. Conclusion

In this work, we present a local soft attention joint training and dual cross-neighbor label smoothing (DCLS) ap-

Figure 7. Top 10 visualization results of the different modules' rank-list on Market-1501 dataset.

proach for unsupervised person Re-ID. First, the entire joint training is divided into global and local parts, and then the soft attention mechanism in the local branch is proposed to capture the subtle differences in local regions accurately, thereby improving the ability of the Re-ID model to identify the important local parts of the person. Second, DCLS is designed to mitigate label noise in different local regions progressively. Under the unsupervised setting of the standard person datasets, extensive ablation experiments prove that the proposed method can effectively capture the local discriminative information of the person. Moreover, it can alleviate the person identity-independent information noise contained in different local regions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62076117 and 62166026, the Jiangxi Key Laboratory of Smart City under Grant No. 20192BCD40002 and the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011.

References

[1] Y. Bai, C. Wang, Y. Lou, J. Liu, and L. Duan. Hierarchical connectivity-centered clustering for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, 30:6715–6729, 2021. 1

[2] G. Chen, Y. Lu, J. Lu, and J. Zhou. Deep credible metric learning for unsupervised domain adaptation person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 643–659, 2020. 6

[3] H. Chen, B. Lagadec, and F. Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision, pages 14960–14969, 2021. 6, 7

[4] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2004–2013, 2021. 6, 7

[5] Y. Chen, X. Zhu, and S. Gong. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 232–242, 2019. 2

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 6

[7] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. 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*, pages 994–1003, 2018. 2

[8] G. Ding, S. Khan, Z. Tang, J. Zhang, and F. Porikli. Towards better validity: Dispersion based clustering for unsupervised person re-identification. *arXiv preprint arXiv:1906.01308*, 2019. 2, 6

[9] Y. Ding, H. Fan, M. Xu, and Y. Yang. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1):1–19, 2020. 1, 7

[10] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 2, 6

[11] H. Fan, L. Zheng, C. Yan, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018. 1

[12] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, 2019. 1, 2

[13] Y. Ge, D. Chen, and H. Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 2, 6, 7

[14] Y. Ge, F. Zhu, D. Chen, and R. Zhao. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020. 2, 3, 6, 7

[15] Q. Han, H. Liu, W. Min, T. Huang, D. Lin, and Q. Wang. 3d skeleton and two streams approach to person re-identification using optimized region matching. *ACM Transactions on Multimidia Computing Communications and Applications*, 2022. 1

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[17] T. He, L. Shen, Y. Guo, G. Ding, and Z. Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 879–887, 2022. 2

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 8

[19] Z. Hu, C. Zhu, and G. He. Hard-sample guided hybrid contrast learning for unsupervised person re-identification. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 91–95. IEEE, 2021. 6, 7

[20] Y. Huang, P. Peng, Y. Jin, Y. Li, and J. Xing. Domain adaptive attention learning for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11069–11076, 2020. 2

[21] H. Ji, L. Wang, S. Zhou, W. Tang, N. Zheng, and G. Hua. Meta pairwise relationship distillation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3661–3670, 2021. 6, 7

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[23] N. Komodakis and S. Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3

[24] J. Li and S. Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 483–499, 2020. 6, 7

[25] J. Li and S. Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *European Conference on Computer Vision*, pages 483–499, 2020. 6, 7

[26] M. Li, X. Zhu, and S. Gong. Unsupervised tracklet person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1770–1782, 2020. 1

[27] Y. Li, H. Yao, and C. Xu. Test: Triplet ensemble student-teacher model for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 30:7952–7963, 2021. 1

[28] Y. Li, H. Yao, and C. Xu. Intra-domain consistency enhancement for unsupervised person re-identification. *IEEE Transactions on Multimedia*, 24:415–425, 2022. 1

[29] S. Lin, H. Li, C. Li, and A. C. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*, 2018. 2

[30] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8738–8745, 2019. 2, 3, 6

[31] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang. Unsupervised person re-identification via cross-camera similarity exploration. *IEEE Transactions on Image Processing*, 29:5481–5490, 2020. 1

[32] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF conference on computer vi-*

sion and pattern recognition, pages 3390–3399, 2020. 1, 2, 6

[33] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458, 2020. 4

[34] C. Luo, C. Song, and Z. Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 224–241, 2020. 6, 7

[35] H. Luo, P. Wang, Y. Xu, F. Ding, Y. Zhou, F. Wang, H. Li, and R. Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021. 6, 7

[36] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 9

[37] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35, 2016. 5

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[39] J. Sun, Y. Li, H. Chen, Y. Peng, and J. Zhu. Unsupervised cross domain person re-identification by multi-loss optimization learning. *IEEE Transactions on Image Processing*, 30:2935–2946, 2021. 1

[40] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 2

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3, 9

[42] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 9

[43] D. Wang and S. Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10981–10990, 2020. 6, 7

[44] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 2

[45] Q. Wang, W. Min, Q. Han, Q. Liu, C. Zha, H. Zhao, and Z. Wei. Inter-domain adaptation label for data augmentation in vehicle re-identification. *IEEE Transactions on Multimedia*, 24:1031–1041, 2021. 2

[46] Q. Wang, W. Min, Q. Han, Z. Yang, X. Xiong, M. Zhu, and H. Zhao. Viewpoint adaptation learning with cross-view distance metric for robust vehicle re-identification. *Information Sciences*, 564:71–84, 2021. 1

[47] Q. Wang, W. Min, D. He, S. Zou, T. Huang, Y. Zhang, and R. Liu. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. *Science China Information Sciences*, 63(11):1–12, 2020. 1

[48] Q. Wang, B. Wu, P. Zhu, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7, 8

[49] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh. Learning sparse and identity-preserved hidden attributes for person re-identification. *IEEE Transactions on Image Processing*, 29:2013–2025, 2019. 1

[50] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2, 5

[51] S. Woo, J. Park, J. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3, 7, 8

[52] J. Wu, S. Liao, X. Wang, Y. Yang, and S. Z. Li. Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 886–891, 2019. 1

[53] F. Yang, Z. Zhong, Z. Luo, Y. Cai, Y. Lin, S. Li, and N. Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4855–4864, 2021. 6

[54] Q. Yang, H. Yu, A. Wu, and W. Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3633–3642, 2019. 1

[55] H. Yu, A. Wu, and W. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 994–1002, 2017. 1

[56] K. Zeng, M. Ning, Y. Wang, and Y. Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13657–13665, 2020. 2, 3, 6

[57] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 594–611, 2020. 6

[58] M. Zhang, K. Liu, Y. Li, S. Guo, H. Duan, Y. Long, and Y. Jin. Unsupervised domain adaptation for person re-identification via heterogeneous graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3360–3368, 2021. 6

[59] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, and L. Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 526–544, 2020. 6, 7

[60] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8522, 2019. 2

[61] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha. Exploiting sample uncertainty for domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3538–3546, 2021. 6, 7

[62] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5310–5319, 2021. 6, 7

[63] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 5

[64] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 6, 7, 8

[65] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 6

[66] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[67] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 87–104, 2020. 7