Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8859088, 9 pages
https://doi.org/10.1155/2021/8859088

Research Article

Deep Camera-Aware Metric Learning for Person Reidentification

Wei Liu,[1] Ping Liang,[1,2] Lei Liu,[1] Zhiqiang Hao,[3] and Xin Xu[1]

[1]*School of Computer Science and Technology, Wuhan University of Science and Technology, 430065 Wuhan, China*
[2]*Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, 430065 Wuhan, China*
[3]*Key Laboratory of Metallurgical Equipment and Control Technology (Wuhan University of Science and Technology), Ministry of Education, 430065 Wuhan, China*

Correspondence should be addressed to Ping Liang; lpnjh@wust.edu.cn

Received 17 September 2020; Revised 17 December 2020; Accepted 24 December 2020; Published 6 January 2021

Academic Editor: Zhili Zhou

Person reidentification (re-id) suffers from a challenging issue due to the significant inconsistency of the camera network, including position, view, and brands. In this paper, we propose a deep camera-aware metric learning (DCAML) model, where images on the identity-level spaces are further projected into different camera-level subspaces, which can explore the inherent relationship between identity and camera. Furthermore, we exploit dynamic training strategy to jointly multiple metrics for identity-camera relationship learning and thus consumedly elevating the retrieval accuracy. Extensive experiments on the three public datasets demonstrated that our method performs competitive results compared to the state-of-the-art person re-id methods.

1. Introduction

Person reidentification (re-id) attracts increasing research interests due to its significance in video surveillance. Although a noticeable improvement has been obtained in recent years, existing person re-id approaches still suffer from some challenging issues: (1) dramatic variations of visual appearance, (2) inconsistency in camera network, and (3) confusion between two similar pedestrians.

To address these problems, classical approaches generally focus on discriminative embedding learning or searching for effective similarity measurement. For example, semantic feature learning is studied via multistage ROI pooling in the work [1]. In the work [2], relations among individual body parts are explored through a GCP network. Additionally, metric learning aims to map semantically similar persons from some manifold onto metrically close person points in another space. In the work [3], an enhanced triplet loss is proposed to learn a distance metric between two pedestrian images. However, these methods are unable to discover the inherent relationship between identity and camera. Although Das et al. [4, 5] proposed a camera network reidentification approach which exploits the camera label information, however, the information was only exploited in their matching part and not utilized in the training stage. Lin et al. [6] exploited intracamera and intercamera consistent-aware information both in the training and testing stages. However, they ignore the inherent relationship between the camera and the pedestrian's features.

Fortunately, we find that the learned features of one person contain no camera-level information. As shown in Figure 1, the images of some pedestrians captured from several cameras are visualized in the same space via TSNE, which performs a disorderly distribution of camera-level information. Different cameras have different geographical locations, viewpoints, and brands. Thus, the specific camera may provide camera-level discriminative information for personal identities, which is usually ignored by existing methods. Therefore, pedestrian discrimination might benefit from joint information of camera and identity.

In this paper, we propose a novel metric approach called the deep camera-aware metric learning (DCAML) model, where person features are projected into a unified identity space, and each identity space is modelled according to camera-level distribution. In this circumstance, the essential relationship between identity and camera can be discovered.

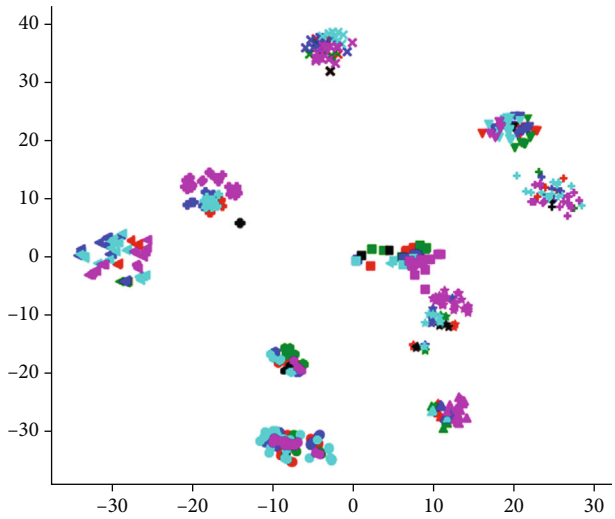

Figure 1: Visualization of representations of several pedestrians. (1) Different shapes denote different identities. (2) Different colors denote different cameras.

Meanwhile, multiple metrics are exploited to formulate the learning of the identity space and camera subspaces. Instead of treating them as separate progress, a dynamic training strategy is developed to integrate them into one optimization objective. In addition, we further consider the hard samples for both intracamera and cross-cameras to improve training quality. Extensive experiments conducted on public datasets show the effectiveness of our method compared with the state-of-the-art approaches.

In summary, the main contributions of this paper include the following:

(1) We propose a deep camera-aware metric learning (DCAML) model to discover the relationship between camera and identity, where camera-level and identity-level information jointly contribute to the retrieval accuracy

(2) We develop a dynamic training strategy to integrate multiple metrics as a unified optimization objective

(3) We introduce online hard example mining into DCAML to further improve the model performance

2. Related Work

Person re-id is often viewed as a subproblem of image retrieval [7, 8]. Recently, with the use of deep learning in person re-id, the performance of person re-id methods has improved to an unprecedented level. Mainstream deep learning methods are divided into supervised and unsupervised learning. The supervised learning approach is adopted in this paper. There are two main classes of methods in the field of person re-id, feature learning and metric learning, which will be introduced in this section.

2.1. Feature Learning for Re-Id. Recent developments in person re-id adopt some form of localized representation learning to achieve improved performance on challenging datasets. For example, Zhao et al. [9] decomposed pedestrian images into different parts and extracted representations of parts and then aggregated as the overall representation. Li et al. [10] proposed to localize parts and learn part features through spatial transformer networks, then combining local and global features for classification. Su et al. [11] exploited pose information as a supervisory signal to learn normalized human part features. Meanwhile, the attention mechanism has been used in person re-id to tackle the localization problem. For example, Liu et al. [12] proposed a HydraPlus-Net network to extract low and semantic-level features for discriminative representations. Li et al. [13] proposed to simultaneously learn region-level and pixel-level attentive features for a multigranular representation. Li et al. [14] trained a predefined attention model for each specific body part and then aggregated them employing a temporal attention model. Additionally, to describe pedestrians with detailed information, a patch-based model [15, 16] slices person images into horizontal grids for better representations. To leverage human parts, aided by pose estimator, pose-based models [9, 13] extract pose maps to obtain part-level features. Another way is to compute attention maps for discriminative regions.

However, these methods do not consider the essential relationship between identity and camera, which may waste the annotations of the camera index. In contrast to them, we consider person representation from the perspective of identity and camera-level distributions.

2.2. Metric Learning for Re-Id. Inspired by the great success of deep learning in computer vision tasks [17–20], many types of research integrate the feature and metric learning jointly in a unified deep framework, where the learning is under the supervision of the distance metric loss. For example, Ding et al. [21] presented a scalable distance-driven framework to introduce triplet loss into person re-id. Based on triplet loss, Hermans et al. [22] designed a variant of the triplet loss for end-to-end person re-id. Besides, compared to the triplet loss, Chen et al. [23] proposed a quadruplet loss to make outputs with a larger interclass variation and a smaller intraclass. Inspired by the hard sample mining method, Xiao et al. [24] proposed a new metric learning loss called margin sample mining loss using hard sample mining. However, all of the above methods do not take advantage of the intrinsic connection between the pedestrian picture and the camera to design the loss.

3. The Proposed Method

3.1. Problem Formulation and Overview. Given a probe image, the objective of the person re-id is to obtain a matched list of images from a gallery across different cameras. Define an image $I_i = (x_i, y_i, c_i)$ where c_i is the camera label, y_i is the identity label, and x_i is the feature extracted by a re-id model.

Figure 2 shows the proposed backbone for feature extraction. We employ the pretrained ResNet50 model as the basic extractor where the last layer is removed and two extra fully

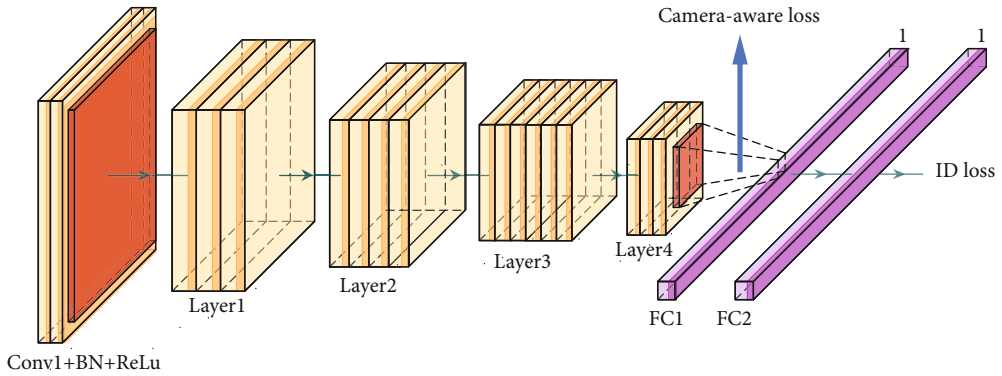

Figure 2: Illustration of the proposed backbone. The last layer is removed, and two extra fully connected (FC) layers are appended. The first FC layer reduces the embedding dimension to 2048, followed by batch normalization. The second FC layer reduces the embedding dimension to the class number as final outputs.

connected (FC) layers are appended. The first FC layer reduces the embedding dimension to 2048, followed by batch normalization. The second FC layer reduces the dimension of the feature tensor to the class number as final outputs. Furthermore, our optimization objective includes ID loss L_{id} and camera-aware lossL_{ca}. ID loss is to model the identity-level distribution, while camera-aware loss (CA loss) is to model the camera-level distribution. These two losses are integrated into a unified framework via a dynamic training strategy. Figure 3 illustrates the overall framework for ID loss and CA loss. During modelling the camera-level distribution, features under the same camera of loss are to model the identity-level distribution, while camera-aware loss is to model the camera-level distribution. These two losses are integrated into a unified framework via a dynamic training strategy. Figure 3 illustrates the overall framework for ID loss and CA loss.

During modelling the camera-level distribution, features under the same camera of the same identity are pulled closer, while features from different cameras of the same identity are pushed away in an appropriate distance. To this end, a mini-batch consists of a series of quadruplets that are denoted as $X_i = \langle x_1^{c1}, x_2^{c1}, x_1^{c2}, x_2^{c2} \rangle$. We will introduce the quadruplet form in Section 3.3.

3.2. Camera-Aware Person Re-Id. Metric learning is widely studied for person re-id. The goal is to explore an effective mapping function $f_\theta(x)$: $R^S \to R^T$ that can map semantically similar person points from the manifold in R^S into metrically close person points in R^T. The θ is the parameter in mapping function f_θ and can be represented ranging from a linear transform to a nonlinear transform of convolutional neural network.

We define Distance(x, y): $R^T \times R^T \to R$ as a distance metric function in the embedding space. For convenience, we use the simple form $D_{i,j} = \text{Distance}(f_\theta(x_i), f_\theta(x_j))$ while ignoring the parameter θ.

Ding et al. [21] investigate the distance relation between intraclass and interclass points, which aims to decrease the intraclass variation while increasing the interclass variation. They formulated it as a metric learning function named "triplet loss" to optimize f_θ:

$$L_{\text{triplet}}(\theta) = \left(L_{\text{pull}}(\theta) + L_{\text{push}}(\theta) + \alpha\right)_+, \tag{1}$$

where α is a tradeoff parameter between positive (intraclass) and negative (interclass) pairs. For an explicit definition, pulling person points of the same identity is defined as:

$$L_{\text{pull}}(\theta) = \sum_{\substack{a,p \\ y_a = y_p}} \left[D_{a,p}\right]_+. \tag{2}$$

while pushing the person points of different identities is defined as:

$$L_{\text{push}}(\theta) = \sum_{\substack{a,n \\ y_a \neq y_n}} \left[D_{a,n}\right]_+. \tag{3}$$

The whole optimization objective can be written as:

$$L_{\text{triplet}}(\theta) = \sum_{\substack{a.p.n \\ y_a = y_p \neq y_n}} \left[D_{a,p} - D_{a,n} + \text{margin}\right]_+. \tag{4}$$

We observe that triplet loss can effectively set identity margins for different identities. Inspired by the form of "triplet loss," we design a novel loss function called "camera-aware loss" (CA Loss). In detail, the motivation of the camera-aware re-id is to construct the bridge between identity and camera. Thus, a similar form of triplet loss can be used to learn appropriate margins for the modelling of camera-level distribution. The camera-aware loss can be written as:

$$L_{\text{CA}}(\theta) = \sum_{\substack{ac.pc.nc \\ y_{ac} = y_{pc} \neq y_{nc}}} \left[D_{ac,pc} - D_{ac,nc} + \text{margin}\right]_+, \tag{5}$$

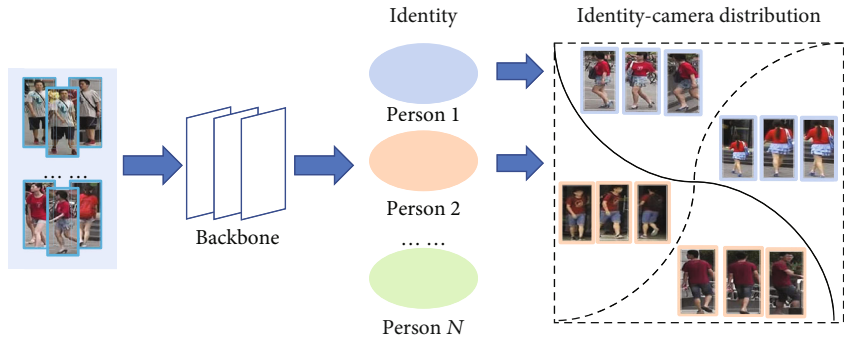

Figure 3: Illustration of the framework. The active line is the split line between image pairs of different identities, while the broken line is the split line between image pairs of different cameras.

where the a means the probe image. The p means the image which constitutes a positive pair with the probe image. The n means the image which constitutes a negative pair with the probe image. The c means the camera label of the image.

To alleviate the ambiguous representation caused by cross-camera variations, the underlying relation between camera and identity should be well mined. Given an anchor point x_c under camera c, the projection of a point under camera c is closer to the anchor's projection than those under another camera p, where $p \neq c$, by at least a margin for camera-level distribution. Via optimizing it on the whole dataset for long enough, all possible cross-camera variation pairs $(x_{ac}, x_{pc})\&(x_{ac}, x_{nc})$ will be searched. In this case, camera-level distribution can be well modelled and cross-camera information can be well learned during training.

3.3. Batch Hard Pair Selection. A main difficulty of the CA loss is that as the dataset or camera number gets larger, the number of cross-camera pairs also grows quickly, rendering a difficult training process. Due to redundancy, the information of most cross-camera pairs is uninformative and trivial. Therefore, for the fast convergence of the model, it is crucial to select hard pairs that are most similar. Intuitively, we consider that the intracamera image of the same identity should be closer and more similar, where the outliers may be the hard one. Meanwhile, the cross-camera pairs of the same identity should be slightly dissimilar, and the most similar cross-camera pairs may be the hard ones.

In mathematics, given a person image x_i^{ac} under camera c, we hope to select x_i^{pc} under camera c that satisfies:

$$\arg\max_{x_i^{pc}} D\left(f_\theta(x_i^{ac}), f_\theta\left(x_i^{pc}\right)\right), \tag{6}$$

and select x_i^{nm} under camera m where $m \neq c$ that satisfies:

$$\arg\min_{x_i^{nm}} D(f_\theta(x_i^{ac}), f_\theta(x_i^{nm})). \tag{7}$$

However, it is time-consuming to calculate the values of argmax and argmin under the whole training set. Besides, it may lead to a worse training process as the hardest images are usually noisy such as wrongly labelled or wrongly detected. To address this problem, we focus on the online hard example mining within a minibatch for calculations of argmax and argmin.

To achieve this goal, we define the intracamera pair as the positive pair and the cross-camera pair as the negative pair. For the effective representation of the CA loss, it needs to be ensured that positive and negative pairs of one identity are present in each minibatch together. Therefore, instead of random sampling for a minibatch, we construct a quadruplet form for each identity within the minibatch:

$$\text{Quadruplet} = \left\langle x_1^A, x_2^A, x_1^B, x_2^B \right\rangle, \tag{8}$$

where $x_{1,2}^A$ are of person x under camera A, $x_{1,2}^B$ are of person x under camera B.

To form a minibatch, P identities are randomly sampled, thus resulting in a minibatch $4 \times P$ images. In each quadruplet, positive pairs are:

$$\left\langle x_1^A, x_2^A \right\rangle, \left\langle x_1^B, x_2^B \right\rangle, \tag{9}$$

while negative pairs are:

$$\left\langle x_1^A, x_1^B \right\rangle, \left\langle x_1^A, x_2^B \right\rangle, \left\langle x_2^A, x_1^B \right\rangle, \left\langle x_2^A, x_2^B \right\rangle, \tag{10}$$

where x_i^c is the i-th image of person x under camera c. As the training progresses, we notice that negative pairs grow larger than positive pairs, due to the limited images of one identity under one camera. Thus, hard positive pairs are not necessary. Therefore, we can mine the hardest negative pairs within the minibatch when computing the camera-aware loss, and we call it batch hard negative pair. The loss is written as:

$$L_{\text{BHNP}}(\theta) = \sum_{a=1}^{P} \sum_{c=1}^{2} \sum_{i=1}^{2} \cdot \left[D(f_\theta(x_i^c), f_\theta(x_{3-i}^c)) - \min_{k=1,2} D\left(f_\theta(x_i^c), f_\theta\left(x_k^{3-c}\right)\right) + \text{margin} \right]_+, \tag{11}$$

where P is the number of identities in a minibatch, and x_i^c corresponds to the$i$-th image of the person x under camera c.

3.4. Multiloss Dynamic Training. Person re-id is aimed at identifying pedestrians across nonoverlapping cameras. The core task is the identification, so identification loss is the key component of optimization objectives. Camera-aware loss is used to alleviate the influence of the cross-camera variations via a bridge between identity and camera. These two optimization objectives can benefit from each other with explicit potential connections and can achieve a better generalization performance.

(1) *Identification loss*: as a point-wise classification loss, identification loss adopts the cross-entropy form for identity prediction, which is defined as:

$$L_{\mathrm{id}} = -\sum_{i}^{K} S\left(\left(W_i\right)^T x_i\right), \tag{12}$$

where i denotes the true identity of the input image x_i. K is the number of persons. S is the softmax function, and W_i is the weight matrix of the last FC layer for i-th identity.

(2) *Dynamic weighing*: how to integrate identification loss and camera-aware loss is a crucial problem due to the different effects. Most multitask researches weigh the different tasks by balancing parameters and formulate some tasks as the regularization items in the loss function. However, in our framework, (1) it is difficult to choose an appropriate parameter to fairly weigh two tasks. And (2) inappropriate parameter setting may produce negative effects on person re-id.

On the one hand, for the most identities, the intracamera variations are slighter than cross-camera variations. Thus, camera-aware loss only provides a small loss value for updating parameters, for which camera-aware loss contributes little to the learning if its weight is too small. On the other hand, in the early learning procedure, the re-id model needs to treat identification as the main task; otherwise, the camera-aware loss may influence the learning of the discriminative information. Besides, from the essence of the person re-id, these two tasks are conflicting when the weight of the camera-aware loss is too large, leading to an intraclass variance. Therefore, with the progress of the training, a camera-aware loss should play a progressive role.

In this work, we propose a progressive balance strategy to ensure the best combination of two losses. For identification loss, it requires no extra weights as the main part. For camera-aware loss, we define a gradient ascent method to estimate the weight of its growth. Suppose O_{id} be the orders of magnitude for ID loss at the initial time. To approximate this order for a balance, we change the weight α of camera-aware loss in a linear way to avoid loss oscillation. We can calculate the gradient grad_{α} of weight α according to:

$$\mathrm{grad}_{\alpha} = \frac{O_{\mathrm{ID}}}{N_{\mathrm{batch}} \times N_{\mathrm{epoch}}}, \tag{13}$$

where $N_{\mathrm{batch}} \times N_{\mathrm{epoch}}$ is the number of the minibatch × epoch. Based on the minibatch index k, the weight can be written as:

$$\alpha = \mathrm{grad}_{\alpha} \times k = k \times \frac{O_{\mathrm{ID}}}{N_{\mathrm{batch}} \times N_{\mathrm{epoch}}}. \tag{14}$$

In the case of weight α increasing, the ratio of ID loss and CA loss is decreasing gradually. Obviously, at the initial time, α is equal to zero with no effects on identification task learning. When the learning is progressing, the larger α ensures that the re-id model can learn from the different viewpoints of one person. Finally, the overall objective can be rewritten as:

$$\begin{aligned} L(\theta) &= L_{\mathrm{id}}(\theta) + L_{\mathrm{BHNP}}(\theta) \\ &= \sum_{i=1}^{N} \left[L_{\mathrm{id}}(\theta, x_i) + \alpha L_{\mathrm{BHNP}}\left(\theta, x_i^{c1}, x_i^{*c1}, x_i^{c2}, x_i^{*c2}\right)\right], \end{aligned} \tag{15}$$

where the x_i^{*c} corresponds to the $3 - i$-th image of the person x under camera c.

4. Experiments

4.1. Experimental Settings

(1) *Datasets*: three public person re-id datasets are available for evaluation, including Market1501 [25], DukeMTMC re-ID [26, 27], and MSMT17 [28].

The Market1501 dataset is collected at the campus under 6 cameras. It includes 1,501 person identities with 19,732 testing images and 12,936 training images. We follow the standard evaluation protocol to ensure fair comparisons, which is defined as follows: (1) the fixed 750 identities are used as the training set to train the re-id model, and (2) 3,368 probe images are matched with the fixed gallery including 751 identities.

The MSMT17 dataset includes 15 cameras. 4,101 identities are captured with 126,441 labelled person boxes. It has 1,041 training identities and 3,060 testing identities. Besides, the person boxes are cropped from the video by the Faster RCNN detector. We adopt the standard evaluation protocol proposed in [26], which is defined as (1) the dataset is randomly split into two parts and (2) the training set and testing set are split according to the ratio of 1 : 3.

The DukeMTMC-reID dataset is a subset of the Duke Dataset. It has 16,522 images of 702 identities for training, while 2,228 probe images and 17,661 gallery images of the other 702 identities for testing. We follow the protocol proposed in [26], defined as follows: (1) 702 identities are randomly selected as the training set and the remaining 702 identities are as the testing set. (2) In the testing set, one image of each identity is randomly selected as the query under each camera, and the remaining images are in the gallery.

(2) *Parameters*: our framework is implemented with Pytorch. The dimension of the embedding for matching is set as 512-dim, and the batch size is 32 for all

Table 1: Comparison results (%) on the Market1501 dataset where the bold font denotes the best method. These methods explore extracting global features for a person re-id. Our proposed model achieves the best results compared with these methods.

Method	Reference	Rank-1	mAP
Ours	—	**92.7**	**78.7**
DML	CVPR [29]	87.7	68.8
Triplet	arXiv [22]	84.9	68.8
Transfer	arXiv [30]	83.7	65.5
PAN	TCSVT [31]	82.8	63.4
SVDNet	ICCV [32]	82.3	62.1
CADL	CVPR [6]	80.9	50.6

Table 2: Comparison results (%) on Market1501 dataset where the bold font and underline denotes the best method; bold font denotes the 2nd best method, and underline denotes the 3rd best method. These methods explore part-based features for a person re-id. Our proposed model achieves the 4th best results on all the 2 evaluation metrics.

Method	Reference	Rank-1	mAP
Ours	—	92.7	78.7
MGN	ACMMM [33]	**95.7**	**88.2**
HPM	AAAI [16]	**94.2**	**82.7**
PCB+RPP	ECCV [15]	93.8	81.6
PCB	ECCV [15]	92.3	77.4
GLAD	ACMMM [34]	89.9	73.9
Partloss	TIP [35]	88.2	59.3
PDC	ICCV [11]	84.4	63.4
Multiloss	IJCAI [36]	83.9	64.4
PAR	ICCV [9]	81.0	63.4
Hydra	ICCV [12]	76.9	—
MRegion	AVSS [37]	66.4	41.2

datasets while the dropout rate is set as 0.5. Random cropping and resize are exploited for data argumentation, and the images are resized to 256×128. SGD is adopted as the 0.9. The initial learning rate is set as 0.05. Moreover, the margin is set as 0.1.

(3) *Evaluation metrics*: The rank-1 and mAP (mean average precision) are used to visually show the model performance.

4.2. Comparison with the State-of-the-Arts. In this section, we compared the proposed framework with more than 30 state-of-the-art methods, which are proposed in recent years, on the three datasets. For the comparison on each dataset, we provide detailed results as follows.

(1) *Market1501*: for this dataset, we compared two kinds of approaches including local feature and global feature approaches. It is illustrated in Table 1 that our method can perform the best accuracy scores on both rank and mAP compared with the global feature methods. Our method exceeds DML 9.9% on metric mAP and 5% on rank-1.

Table 3: Comparison results (%) on DukeMTMC-reID dataset where the bold font and underline denotes the best method; bold font denotes the 2nd best method, and underline denotes the 3rd best method. Our proposed model achieves the 3rd best result on rank-1 and 4th best result on mAP.

Method	Reference	Rank-1	mAP
Ours	—	83.6	67.8
HPM	AAAI [16]	**86.6**	**74.3**
PA	ECCV [38]	**84.4**	**69.3**
PCB	ECCV [15]	83.3	69.2
DuATM	CVPR [39]	81.8	64.6
MLFN	CVPR [40]	81.2	62.8
HA-CNN	CVPR [13]	80.5	63.8
ATWL	CVPR [41]	79.8	63.4
PSE	CVPR [42]	79.8	62.0
DPFL	ICCVW [43]	79.2	60.6
CamStyle	CVPR [44]	78.3	57.6
AACN	CVPR [45]	76.8	59.3
dMpRL	TIP [46]	76.8	58.6
SVDNet	ICCV [32]	76.7	56.8
PAN	TCSVT [31]	71.6	51.5

The results indicate that local feature-based methods perform better than the only global feature-based methods in general. Compared with local feature-based methods in Table 2, our method can obtain 4-*th* best performance only using pre-trained ResNet50. Although MGN and PCB+RPP and HPM perform slightly better than ours, MGN and HPM explore both global and local information for person re-id, while PCB+RPP takes multiple parts for person re-id. Both MGN and PCB+RPP for person re-id. have a more complex architecture with multiple branches. Besides, our method can outperform PCB without RPP 0.4% on rank and 1.3% on mAP.

(2) DukeMTMC-reID: As illustrated in Table 3, we can conclude that our method can exceed the most SOTA methods on this dataset on both metrics mAP and rank-1. The HPM method outperforms our method by 3.0% on rank-1 and 6.5% mAP but using both global and local information, while our method only uses global information. The part-aligned method outperforms our method by a small gap with 0.7% on rank-1 and 1.5% mAP but using part-aligned information, while our method only needs to extract a simple global feature without other operations. MLFN is a method to extract multi-level features that perform worse than ours on rank-1 (81.2% vs. 83.6%). Besides, our model exceeds PCB 0.3% on rank-1. In addition, our method also performs better than some attention based methods such as HA-CNN.

Table 4: Comparison results (%) on MSMT17 dataset at 2 evaluation metrics: mAP and rank-1.

Method	Reference	Rank-1	mAP
Ours	—	50.5	33.1
GoogleNet	CVPR [28]	47.6	23.0
PDC	CVPR [28]	58.0	29.7
GLAD	CVPR [28]	61.4	34.0

Table 5: Ablation study results (%) on three datasets. If without dynamic weighting, the ratio of ID loss and CA loss is 1 : 1. ID is identification loss while CA is a camera-aware loss.

Dataset	Model	Rank-1	mAP
Market1501	ResNet50+ID	87.8	71.6
	ResNet50+ID+CA	91.6	77.9
	ResNet50+ID+CA+DW	92.7	78.7
DukeMTMC-reID	ResNet50+ID	80.2	62.2
	ResNet50+ID+CA	81.3	65.0
	ResNet50+ID+CA+DW	83.6	67.8
MSMT17	ResNet50+ID	45.8	29.9
	ResNet50+ID+CA	47.5	31.4
	ResNet50+ID+CA+DW	50.5	33.1

(3) *MSMT17*: MSMT17 dataset is a recently released dataset that contains complex illumination, scenes, and background. Due to limited works for MSMT17, we merely compared the works presented in [28] that releases the MSMT17 dataset. As illustrated in Table 4, our method can elevate the identification accuracy without extra information. In detail, compared with the state-of-the-art methods, our method exceeds the GoogleNet compared with the state-of-the-art methods; our method exceeds the GoogleNet 2.9% on rank-1 and 10.1% on mAP and exceeds PDC 0.4% on mAP while approximate rank-1.

4.3. Ablation Study. To further discuss every component in our framework, we conducted a series of comprehensive ablation studies for the different submodules. The performance results at 2 metrics (mAP and rank-1) are shown in Table 5. Each result is obtained with only one submodule changed, and the rest submodules are the same as the original.

We only used the fine-tuned ResNet50 to extract the feature for a person re-id. And then we added batch hard negative pair on ResNet50 to test the performance. From Tables 1–5, we can conclude (1) BHNP sampling method is more useful than the random sampling method, which indicates that the cross-camera quadruplets are effective. Besides, via BHNP, overfitting is effectively avoided during training. (2) With BHNP and CM loss, the performance can further exceed that of the ResNet50, which indicates that CA loss is effective. The margins among camera embedding spaces can successfully reduce the confusing

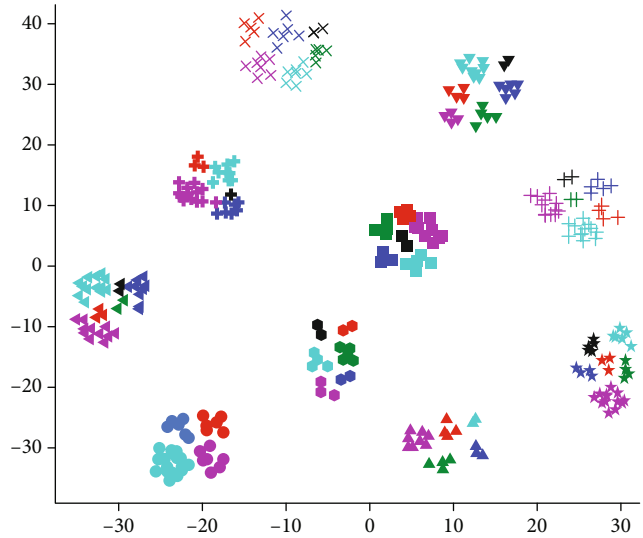

Figure 4: Visualization of representations of several pedestrians via DCAML. (1) Different shapes denote different identities. (2) Different colors denote different cameras.

wrong matching pairs. And (3) dynamic weighting is important for the CA loss to product performance improvement. It indicated that the dynamic weighting strategy can improve the function of the CA loss. Without dynamic weighting, the CA loss may provide a negative influence on the performance.

4.4. Visualization Results. To directly indicate the effectiveness of our method, we visualized the same images shown in Figure 1, by TSNE and PCA. As shown in Figures 1 and 4), different shapes denote different identities. (2) Different colors denote different cameras. The different identities can be well classified by our method, while there exists obvious camera-level information in the same identity category. In fact, camera-level information also can provide help for identification because the visual representations captured from cameras contain the characteristics of the cameras, such as camera style, viewpoints, and scale.

5. Conclusion

In this paper, we propose a deep camera-aware metric learning (DCAML) model for person reidentification, where images on the identity-level spaces are further projected into different camera-level subspaces. We explore the inherent relationship between identity and camera. Furthermore, multiple loss functions are utilized to supervise the learning of the identity-level spaces and camera-level subspaces. In addition, we also consider joint multiple metrics for identity-camera relationship learning via a designed dynamic training strategy. Extensive experiments on the three public datasets demonstrated that our method performs competitive results compared to the state-of-the-art person re-id methods.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of China (U1803262, 61602349, and 61440016).

References

[1] H. Zhao, M. Tian, S. Sun et al., "Spindle net: person re-identification with human body region guided feature decomposition and fusion," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1077–1085, Honolulu, HI, USA, 2017.

[2] H. Park and B. Ham, "Relation network for person re-identification," 2019, https://arxiv.org/abs/1911.09318.

[3] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1335–1344, Las Vegas, NV, USA, 2016.

[4] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury, "Network consistent data association," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1859–1871, 2016.

[5] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *Computer Vision – ECCV 2014. ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8690 of Lecture Notes in Computer Science, pp. 330–345, Springer, Cham, 2014.

[6] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5771–5780, Honolulu, HI, USA, 2017.

[7] Z. Zhou, Q. J. Wu, Y. Yang, and X. Sun, "Region-level visual consistency verification for large-scale partial-duplicate image search," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–25, 2020.

[8] Z. Zhou, Y. Mu, and Q. J. Wu, "Coverless image steganography using partial-duplicate image retrieval," *Soft Computing*, vol. 23, no. 13, pp. 4927–4938, 2019.

[9] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3219–3228, Venice, Italy, 2017.

[10] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 384–393, Honolulu, HI, USA, 2017.

[11] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3960–3969, Venice, Italy, 2017.

[12] X. Liu, H. Zhao, M. Tian et al., "Hydraplus-net: attentive deep features for pedestrian analysis," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 350–359, Venice, Italy, 2017.

[13] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Salt Lake City, UT, USA, 2018.

[14] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 369–378, Salt Lake City, UT, USA, 2018.

[15] 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 *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11208 of Lecture Notes in Computer Science, pp. 480–496, Springer, Cham, 2018.

[16] Y. Fu, Y. Wei, Y. Zhou et al., "Horizontal pyramid matching for person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8295–8302, 2019.

[17] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3d action recognition," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 519–529, 2016.

[18] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[19] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5292–5303, 2018.

[20] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4242–4251, Salt Lake City, UT, USA, 2018.

[21] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, https://arxiv.org/abs/1703.07737.

[23] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 403–412, Honolulu, HI, USA, 2017.

[24] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: a deep learning based method for person re-identification," 2017, https://arxiv.org/abs/1710.00478.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: a benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, Santiago, Chile, 2015.

[26] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3754–3762, Venice, Italy, 2017.

[27] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Computer Vision – ECCV 2016 Workshops. ECCV 2016*, G. Hua and H. Jégou, Eds., vol. 9914 of Lecture Notes in Computer Science, pp. 17–35, Springer, Cham, 2016.

[28] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88, Salt Lake City, UT, USA, 2018.

[29] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*pp. 4320–4328, Salt Lake City, Utah, USA, 2018.

[30] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," 2016, https://arxiv.org/abs/1611.05244.

[31] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.

[32] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3800–3808, Venice, Italy, 2017.

[33] 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*pp. 274–282, New York, USA, 2018.

[34] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*pp. 420–428, Mountain View, California, USA, 2017.

[35] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019.

[36] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," 2017, https://arxiv.org/abs/1705.04724.

[37] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, Lecce, Italy, 2017.

[38] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11218 of Lecture Notes in Computer Science, pp. 402–419, Springer, Cham, 2018.

[39] J. Si, H. Zhang, C.-G. Li et al., "Dual attention matching network for context-aware feature sequence based person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5363–5372, Salt Lake City, UT, USA, 2018.

[40] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2109–2118, Salt Lake City, UT, USA, 2018.

[41] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6036–6046, Salt Lake City, UT, USA, 2018.

[42] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 420–429, Salt Lake City, UT, USA, 2018.

[43] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2590–2600, Venice, Italy, 2017.

[44] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, Salt Lake City, UT, USA, 2018.

[45] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2119–2128, Salt Lake City, UT, USA, 2018.

[46] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1391–1403, 2019.