Spectral Feature Transformation for Person Re-Identification

Chuanchen Luo[1,3] Yuntao Chen[1,3] Naiyan Wang[2] Zhaoxiang Zhang[1,3,4]
[1] University of Chinese Academy of Sciences [2] TuSimple
[3] Center for Research on Intelligent Perception and Computing, CASIA
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS
[1]{luochuanchen2017, chenyuntao2016, zhaoxiang.zhang}@ia.ac.cn
[2]winsty@gmail.com

Abstract

With the surge of deep learning techniques, the field of person re-identification has witnessed rapid progress in recent years. Deep learning based methods focus on learning a discriminative feature space where data points are clustered compactly according to their corresponding identities. Most existing methods process data points individually or only involves a fraction of samples while building a similarity structure. They ignore dense informative connections among samples more or less. The lack of holistic observation eventually leads to inferior performance. To relieve the issue, we propose to formulate the whole data batch as a similarity graph. Inspired by spectral clustering, a novel module termed Spectral Feature Transformation is developed to facilitate the optimization of group-wise similarities. It adds no burden to the inference and can be applied to various scenarios. As a natural extension, we further derive a lightweight re-ranking method named Local Blurring Re-ranking which makes the underlying clustering structure around the probe set more compact. Empirical studies on four public benchmarks show the superiority of the proposed method. Code is available at https://github.com/LuckyDC/SFT_REID.

1. Introduction

Person re-identification (ReID) is an indispensable component in surveillance video analysis. Given the probe, person ReID aims at identifying images of the same person across multiple non-overlapping camera views. Thanks to the emergence of deep learning techniques and large scale datasets [62, 64, 21, 55], the field of person identification evolves rapidly. Though having achieved much progress, it remains challenging due to drastic pose variation, occlusion, and background cluttering.

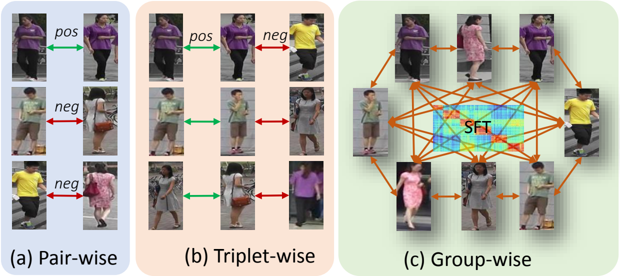


Figure 1. The illustration of the similarity structure built by different methods. Contrastive and triplet loss focus on pair-wise(a) and triplet-wise(b) relationship, respectively. While our method capture dense connections between samples by formulating the data as a graph(c).

Deep learning based ReID methods focus on exploiting the powerful capability of neural networks to learn discriminative feature. When projected to the obtained feature space, data points are expected to gather into clusters according to their labels. In order to intensify intra-class compactness and inter-class separability of the feature, many efforts have been made in recent years. Besides designing tailor-made neural architectures, a large variety of loss functions have also been proposed. The two most prevalent types of loss functions in ReID are classification loss(*e.g.* softmax cross entropy loss) [63, 10, 44, 43] and metric learning based loss(*e.g.* triplet loss and contrastive loss) [5, 15, 60]. Classification loss has promising convergence but is vulnerable to overfitting. It processes samples individually and only builds connections implicitly through the classifier. Metric learning based loss explicitly optimizes the distances between samples. While the similarity structure it builds only involves a pair/triplet of data points and ignores other informative samples. This leads to a large proportion of trivial pairs/triplets which could overwhelm the training process and eventually makes the model suffer from slow convergence. To relieve the issue, many methods [40, 29, 27] incorporate more samples while building the similarity structure. Whereas, they are still limited by the number of samples considered which may impair the performance.

Motivated by aforementioned observations, we propose to capture more informative structure by taking all instances in the batch into account. Specifically, the whole data batch is regarded as a similarity graph in our method. The insight is illustrated intuitively in Figure 1. To encourage the learning of discriminative feature, we borrow the inspiration from spectral clustering which operates on the similarity graph of the input data. Given the input data, spectral clustering partitions them into groups. It is expected that samples from different groups have very low similarities and those within the same group have high similarities. Under the setup of supervised learning, the ground-truth partitions (*i.e.* identity label) are also given. In addition, group-wise similarity derived from the feature can be learned in the deep learning scheme. Thus, the objective becomes to optimize *group-wise* similarities such that the given partitions are optimal in the case. Whereas, it is non-trivial to optimize group-wise similarities directly. Alternatively, we perform a feature transformation with the guidance of the derived transition probabilities. Then, the supervision is imposed on the transformed feature. To push the performance ahead, we further combine the model with an auxiliary classification branch. The whole process is fully differentiable and only brings marginal computational cost. Despite its simplicity, the proposed method improves the performance significantly over strong baselines.

Furthermore, we adapt the online feature transformation to the offline post-processing stage. In the properly learned embedding space, there underlies a clustering structure in the local neighborhood of each data point. The proposed local blurring re-ranking acts as a pre-clustering process. It makes ambient clustering structure more compact which could diminish ambiguity in retrieval.

In summary, this paper has following contributions:

- To efficiently capture more informative structure, we form the data in one batch into a similarity graph. Inspired by spectral clustering, a novel feature transformation is proposed which facilitates the optimization of group-wise similarities on the graph. It introduces no extra cost to the inference and can be readily adapted to other tasks which require embeddings.
- A lightweight re-ranking method is naturally derived. It makes the underlying clustering structure more compact in the neighborhood of the probe set.
- Extensive experiments validate the effectiveness of our method. Competitive performances are achieved on all four public benchmarks.

2. Related Works

Person re-identification has witnessed rapid progress lately with the power of deep neural networks. Recent efforts on deep learning based person ReID can be roughly categorized into two directions. One is to customize the network architecture for person ReID. Besides common techniques in CNN such as multi-scale feature aggregation [30] or attention modules [22, 47], tailor-made architectures [44, 42, 33, 53, 49, 11] for person ReID are also devised. Sun *et al.* [44] split the feature map into several horizontal parts and imposed supervision on them directly. Suh *et al.* [42] employed a sub-network to learn body part feature and fused it with appearance feature via a bilinear-pooling layer. These methods explicitly consider the structure of human body to alleviate the impact of occlusion or inaccurate detections, thus improve the performance.

The other direction concentrates on developing discriminative loss functions. There are two dominant streams in this direction. One is to introduce the classical metric learning into deep learning, such as contrastive loss [12] and triplet loss [34]. The performances of these methods are highly dependent on the similarity structure built in training. Several works make improvement by incorporating more informative samples [40, 29, 27]. Another stream improves on classification loss. Center loss [57] regularizes the distance between data points and their corresponding class center. Large-margin softmax [25] and its variants [24, 50, 48] enforce various types of margin on the vanilla softmax cross entropy loss. They all have demonstrated effectiveness in face recognition and person ReID.

Spectral clustering is a conventional algorithm for data clustering. It was pioneered by Donath *et al.* [8] and became popular in the pattern recognition community since some landmark works [38, 28, 26, 46]. It is based on the spectral graph theory and converts the data clustering problem into the graph partition problem. In contrast to K-Means, spectral clustering makes no assumption on the structure of the cluster. So it can generalize to more complex scenarios like intertwined spirals. Some recent works [16, 35, 45, 58] tried to incorporate spectral clustering with deep learning. Though spectral clustering has been applied extensively, combining it with CNN in person re-identification is still under investigation.

Re-ranking is a post-processing technique to refine the ranking of retrieval results. In essence, re-ranking methods aim at enhancing the original similarity metric by the information of local neighbors. Early works [19, 31] tried to explore k-reciprocal nearest neighbors for general image retrieval. Recently, Zhong *et al.* [65] introduced re-ranking technique into ReID task. They combined the Jaccard distance of k-reciprocal encodings and the Euclidean distance of original features in post-processing. Along this line, Sarfraz *et al.* [33] aggregated distances between expanded neighbors of image pairs to reinforce the original pairwise distance. Moreover, to take advantage of the diversity within a single feature, Yu *et al.* [61] further fused distances between different sub-features.

Graph convolutional networks generalize the vanilla convolution operator to non-Euclidean data. Due to the complementarity, it often acts as a feature aggregation component over the current CNN framework. GCN was first proposed by Kipf *et al.* [20] for semi-supervised classification. Currently, it is a rising research direction in computer vision. Yan *et al.* [59] modeled dynamics of human body skeletons via graph convolutional networks. Wang *et al.* [51, 52] exploited GCN and an equivalent view non-local feature aggregation to capture the spatial-temporal relations between convolutional features and object proposals in the video, respectively. GCN focuses on propagating and transforming information within the graph to generate better features. While the proposed SFT module aims at adjusting the supervision to guide the learning of the feature below it. The two methods differ in their motivations.

The two most related works to ours are [36, 37]. They both applied similarity transformation on the graph to achieve better results. However, there are obvious discrepancies in terms of the definition of the graph. For each image in the probe set, they construct one graph with probe-to-gallery similarities as nodes and gallery-to-gallery similarities as edges. While in our approach, each node directly corresponds to the feature of a sample and each edge is defined as the similarity of its endpoints. Consequently, in each mini-batch, they need to construct several subgraphs, while we treat the whole mini-batch as one single graph which is much conceptually simpler and faster.

3. Method

To capture thorough information from the data, we propose to formulate data points in the training batch as a graph. In the case, we focus on optimizing group-wise similarity on the graph. The inspiration is initially borrowed from spectral clustering which operates on the similarity graph of the data.

We first give a brief introduction of spectral clustering algorithm and its closely related concept graph cut in Section 3.1. We then elaborate on the proposed Spectral Feature Transformation (SFT) in Section 3.2. In Section 3.3, we extend the proposed feature transformation to the post-processing stage to further refine the retrieval result.

3.1. Graph Cut and Spectral Clustering

Under the setup of spectral clustering, data $X = \{x_i\}_{i=1,\dots,n}$ are represented as an undirected graph. Wherein, each vertex of the graph corresponds to a data point in X and each edge is weighted by the similarity between its endpoints $w_{ij} = \mathrm{sim}\,(x_i, x_j)$. For brevity, we take the 2-cluster problem as an example in the following formulation, and readers can refer to [41] for the multi-cluster extension.

To obtain the optimal clustering result on a graph, an intuitive way is to solve a minimum cut problem. For two disjoint subsets $A, B \subset X$, the cut between them is defined as

$$\mathrm{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}. \tag{1}$$

However, minimizing vanilla cuts often leads to a trivial solution where a single vertex is separated from the rest of the graph. To circumvent the issue, Shi *et al.* [38] proposed to normalize each subgraph by its volume:

$$\mathrm{Ncut}(A, B) = \frac{\mathrm{cut}(A, B)}{\mathrm{vol}(A)} + \frac{\mathrm{cut}(A, B)}{\mathrm{vol}(B)}, \tag{2}$$

where $\mathrm{vol}(A) = \sum_{i \in A, j \in X} w_{ij}$ is the total connection from nodes in A to all nodes in the graph.

3.2. Spectral Feature Transformation

Suppose $X \in \mathbb{R}^{n \times d}$ is the final embedding of a training batch. Wherein, n and d denote the number of data points and the dimension of the embedding vector, respectively. We adopt the cosine similarity with Gaussian function to measure the affinities between samples. Formally, each element of the affinity matrix W is defined as

$$w_{ij} = \exp\left(\frac{x_i^T x_j}{\sigma \cdot \|x_i\|_2 \|x_j\|_2}\right), \tag{3}$$

where σ is a hyper-parameter which reflects the decay rate of the affinity as the cosine similarity decreases. Now, we can define a similarity graph over all data points in the mini-batch as $G = (X, W)$. By normalizing the rows of W to 1, we can derive the transition probability matrix T:

$$T = D^{-1} W, \tag{4}$$

where D is a diagonal matrix whose elements are defined as $d_i = \sum_{j=1}^{n} w_{ij}$. In practice, the computation of T can be implemented by applying softmax function with temperature σ on affinity matrix W.

The most intriguing property we can derive from T is the escaping probability $P(A \to \bar{A})$. *It is proportional to the total transition probability from a subgraph $A \subset X$ to another $\bar{A} = X - A$* [26]. In ReID task, a subgraph A denotes the set of samples belonging to the same identity. So, the escaping probability is essentially the chance of an identity getting misclassified. In other words, it measures inverse group-wise similarities. It is straightforward that a small $P(A \to \bar{A})$ requires strong intra-cluster connections and weak inter-cluster connections, which is the desired property for spectral clustering. In fact, as proved in [26], the escaping probability is exactly equivalent to the Ncut metric,

$$\mathrm{Ncut}(A, \bar{A}) = P(A \to \bar{A}) + P(\bar{A} \to A). \tag{5}$$

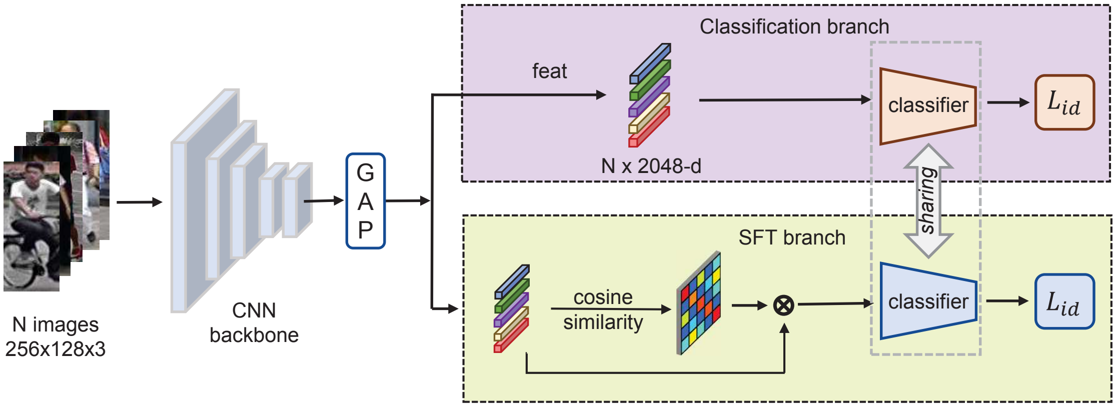


Figure 2. The overall architecture of the proposed model. We adopt the output of the final global average pooling layer as image embedding for retrieval. Spectral feature transformation is performed on the embeddings of the data batch. Subsequently, a classifier is imposed on the transformed feature. We also combine the model with an extra classification branch. Parameters are shared between the two classifiers.

From this perspective, Ncut metric can be readily derived from the transition probability matrix T.

Given the data, traditional spectral clustering aims at seeking optimal partitions w.r.t the Ncut metric. While in the fully supervised setting, the ground-truth partitions A, $\bar{A}$ are known. Moreover, the feature which can derive the transition probability matrix T is learned adaptively in the deep learning paradigm. In this case, the objective becomes to optimize T so that the Ncut metric of the given partitions is minimal. By doing this, we essentially minimize the probability of misclassifying a data sample from group A into group $\bar{A}$. Unfortunately, directly optimizing the transition probability is ill-conditioned. The hard constraint overlooks the potential connection between samples which degrades the performance. Alternatively, we utilize T to guide the transformation of feature X and apply supervision on the transformed feature. Specifically,

$$X' = TX, \tag{6}$$

where X' denotes the feature which has undergone the transformation. Subsequently, the supervision is imposed on the transformed feature using a classifier. In the case, implicit connections are considered and the necessity of hard constraint is also bypassed tactfully. The scheme can also be understood from the viewpoint of the spring model. As vertexes are optimized, springs (weighted by T) change accordingly.

To fully liberate the power of spectral clustering, it is necessary to satisfy the assumption that the input data obey the underlying cluster structure. In other words, there must be sufficient images for each identity in the training batch. Thus, we adopt the sampling strategy proposed by Hermans *et al.* [15] which is ubiquitous in deep metric learning. Specifically, a mini-batch in training contains P identities and each identity has K images. To further push the performance ahead, we combine an extra vanilla classification branch as in many existing methods. The two branches *share the same classifier* as supervision. Only then can we guarantee that the distribution of features are aligned before and after spectral feature transformation. Notably, *the proposed spectral feature transformation is just applied in the training process and would be discarded during inference*. The overall architecture of the proposed neural networks is displayed in Figure 2.

3.3. Local Blurring Re-ranking

In this section, we further extend the proposed spectral feature transformation to the offline post-processing stage. Given a probe image, images in the gallery are ranked according to the cosine similarity with it. Then, we collect features of top-n entries and perform spectral feature transformation on them. Finally, the top-n rank list is recomputed based on the similarity derived from transformed features. Since n is much smaller than the size of the gallery and the features are extracted in advance, the refinement process introduces negligible overhead.

The extension is based on the assumption that there underlies a cluster structure in the neighborhood of the probe images. This is exactly the case when the feature extractor has been properly trained on the training data. As expressed in the mathematical formulation of spectral feature transformation, the embedding of each data point will be blurred by the others according to the similarities between them. Each data point will be moved towards the high-density area (*i.e.* cluster center) which has more short paths to it. This process is equivalent to conduct a clustering operation on local neighbors of the probe image [2]. Therefore, it can make the cluster structure more compact and relieve the ambiguous issue in retrieval. In addition, as the evaluation protocol implies, the top ranking list has a larger impact on the fi-

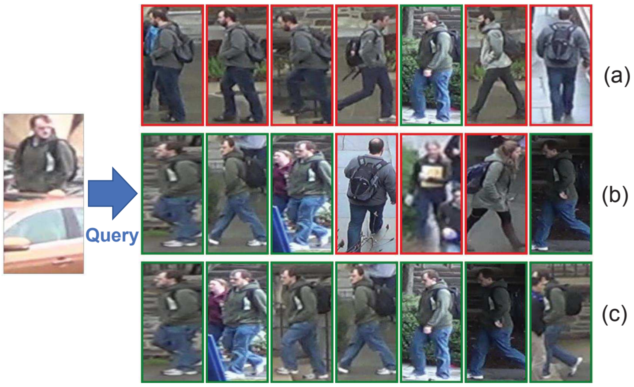


Figure 3. A retrieval example on DukeMTMC-reID. (a) is the result of the model with classification branch only. (b) is generated by the proposed model (*i.e*. SFT + classification). (c) is the refined result based on (b) using local blurring re-ranking.

nal performance. So we only refine the top-n ranking list to balance efficiency and performance gain. Compared with k-reciprocal re-ranking which is operated on the whole test set, the proposed re-ranking is much more efficient. Experiments show that this simple operation leads to prominent improvement.

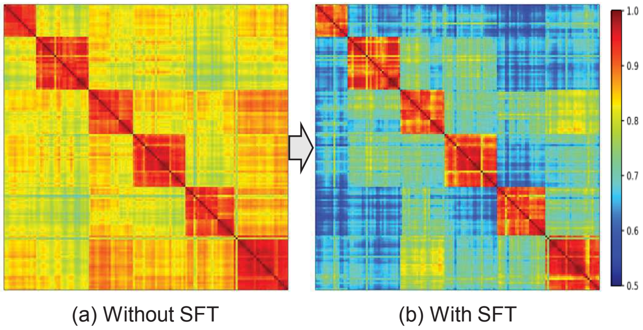
3.4. Discussion

In the sequel, we will analyze some appealing properties of our method which contribute to the improvement and connections to other techniques.

Relax assumptions and ease optimization Instead of applying direct constraints on pairwise similarities, our method relaxes the learning objective to optimize such similarities after the group-wise transformation of SFT. SFT moves the features towards the corresponding cluster center, thus it has enhanced the discrimination of features. This stabilizes the training process and finally leads to better performance.

Training diversity According to the definition of SFT, all samples in the mini-batch participate in the operation. The transformed features of the same sample differ because the composition of the data batch changes while training. This desired property introduces massive diversity which effectively alleviates the risk of over-fitting.

Connection to diffusion process Both diffusion process [18, 1, 9] and our SFT are based on Markov process. Meanwhile, they are different in motivation and implementation. In terms of motivation, diffusion process aims to obtain a more faithful similarity, while the objective of our method is to learn discriminative features. As for implementation, diffusion process is performed on the whole dataset, while our SFT process data in the form of mini-batch. They are applied to affinity matrices and features, respectively.



(a) Without SFT (b) With SFT

Figure 4. Visualization of the affinity matrix. We randomly sample 6 identities from DukeMTMC-reID and take all images belonging to them for visualization. For clarification, samples are arranged according to their identities. It can be seen that the proposed spectral feature transformation significantly suppresses the similarities among different identities.

4. Experiments

4.1. Datasets

To validate the effectiveness of the proposed method, we conduct extensive experiments on four popular person re-identification benchmarks, *i.e*., Market-1501 [62], DukeMTMC-reID [64, 32], CUHK03 [21] and MSMT17 [55]. In terms of CUHK03, we use manual annotations and follow the protocol proposed in [65].

Given the probe images, gallery images are sorted according to the *cosine similarity* with it. On the basis of generated ranking list, Cumulated Matching Characteristics (CMC) at rank-1, rank-5 and mean average precision (mAP) are calculated to evaluate the performance of the model.

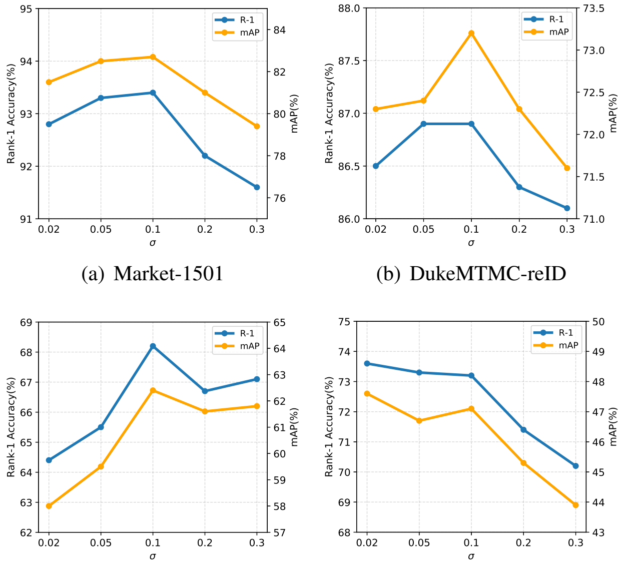
4.2. Implementation Details

We adopt ResNet-50 [14] pre-trained on ImageNet [7] as our backbone network. We use the output of global average pooling layer of ResNet as the embedding vector. In order to preserve more fine-grained information, the downsampling of the last stage of ResNet is discarded which leads to a total stride of 16. The hyper-parameter σ of SFT layer is set to 0.02 for MSMT17 and 0.1 for the remaining three datasets. As for the classifier, we follow a bottleneck design which has been proven effective by many works [44]. Specifically, a fully-connected layer is applied to reduce the dimension of the feature from 2048 to 512 which is followed by Batch Normalization [17] and PReLU [13]. The output is then l_2-normalized and fed into the loss function. To push the performance ahead, we adopt AM-Softmax [48] loss for the final classification. In all experiments, the margin and the scaling parameter of AM-Softmax are set to 0.3 and 15, respectively. In terms of data pre-processing, input images are resized into 256×128. Random horizontal flipping and random erasing [66] are utilized as data augmentation. In training, each mini-batch contains 16 persons and each person has 8 images which results in a batch size of 128. Stochastic Gradient Descent (SGD) with the momentum of 0.9 is applied for optimization. We train 140 epochs in total. The learning rate warms up from 0.001 to 0.1 linearly in the first 20 epochs. It is decayed to 0.01 and 0.001 at 80th and 100th epoch, respectively. As for local blurring re-ranking, we refine the top-50 ranking list for each probe image on Market-1501, DukeMTMC-reID and CUHK03. While for MSMT17, top-150 ranking list is refined, since it has a much larger gallery than the other datasets. Our implementation is based on MXNet [4] framework.

4.3. Ablation Study

Effectiveness of Spectral Feature Transformation. As shown in the first two rows of Table 1, consistent improvements are achieved on all four benchmarks. The improvement on Market-1501 is relatively marginal. There are many persons with few images on Market-1501, *e.g.* 161 persons have no more than 8 images. In such condition, the balanced sampling would re-sample frequently from the same images, which may limit the improvement. For convenience, we employ the same training setting for all datasets which makes the baseline overfit on CUHK03. While our approach is immune to overfitting as mentioned before. This results in significant improvement on CUHK03. In addition, we visualize the affinity matrix between images of 6 different identities with and without SFT module. It can be easily observed in Figure 4 that the affinity between different identities is obviously suppressed. Thus, the features extracted by our method are more discriminative for person ReID. It it noteworthy that the proposed SFT introduces negligible training overhead and no extra parameters. In our setting, it only leads to 0.0336 GFLOPs computation, while the overhead of the backbone network is 4.08 GFLOPs. The relative cost is less than 1%.

Effectiveness of Local Blurring Re-ranking. We also evaluate our method with and without the proposed local blurring re-ranking. As reported in rows 4-5 of Table 1, local blurring re-ranking could further improve the performance significantly. To further clarify its effectiveness, we make a comparison with the k-reciprocal encoding [65] method. As shown in the last two rows in Table 1, the proposed post-processing surpasses k-reciprocal encoding on all benchmarks in terms of Rank-1 accuracy which is the most considerable metric in the real scenario. As for mAP, our post-processing method demonstrates advantages only on the CUHK03 dataset. Note that k-reciprocal encoding takes massive resource to search for k-reciprocal nearest-neighbors of all items in the gallery. Suppose the gallery size is N, the computational complexity of k-reciprocal re-ranking is $O(N^2 \log N)$, while that of LBR is $O(N \log N)$. The gap of efficiency becomes significant when the gallery gets larger. This is also validated by the elapsed time on the three largest dataset reported in Table 3. Taking all these components together, the performance of our method improves dramatically. A qualitative illustration of the retrieval is represented in Figure 3. It is clear that the ranking result improves when components are added sequentially. And Local blurring re-ranking effectively corrects false matches.



(a) Market-1501 (b) DukeMTMC-reID

(c) CUHK03 (d) MSMT7

Figure 5. The influence of bandwidth σ.

Combination with vanilla classification branch. As shown in row 2 and row 4 in Table 1, the combination with classification branch leads to prominent improvement. We also implement a variant supervised by triplet hard loss [15] and classification loss as the counterpart. Our method shows consistent superiority over it. We find the participation of triplet loss even degrades the performance on MSMT17 slightly. We further investigate the necessity to share parameters between the two branches. As shown in Table 2, shared version outperforms unshared version significantly. It is reasonable since independent classifiers may optimize model to different directions which impairs the stability while training.

4.4. Parameter Analysis

Influence of hyper-parameter σ. The proper selection of affinity function is crucial for the success of spectral clustering. So, it is necessary to investigate the impact of σ on the learned features. To this end, we vary σ to five different values and evaluate the performance of the model trained under these settings. As visualized in Figure 5, our method is relatively robust to the value of λ.

Influence of the number of images per identity K. We investigate the trend of the performance when varying K. Given that Market-1501 and CUHK03 are relatively small

Variants	Market-1501		DukeMTMC		CUHK03		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Classification branch only (baseline)	77.3	91.2	63.9	82.0	40.6	44.9	37.3	66.7
SFT branch only	79.6	91.6	70.4	85.4	60.2	66.3	44.7	71.9
Triplet + Classification	80.0	92.2	68.2	83.7	27.6	59.0	60.2	65.6
SFT + Classification	82.7	93.4	73.2	86.9	62.4	68.2	47.6	73.6
SFT + Classification + LBR	87.5	94.1	79.6	90.0	71.7	74.3	58.3	79.0
SFT + Classification + k-reciprocal	90.6	93.5	83.3	88.3	68.7	71.7	60.8	76.1

Table 1. Ablation studies on Market-1501, DukeMTMC-reID, CUHK03(labeled) and MSMT17 dataset. LBR denotes the proposed local blurring re-ranking method.

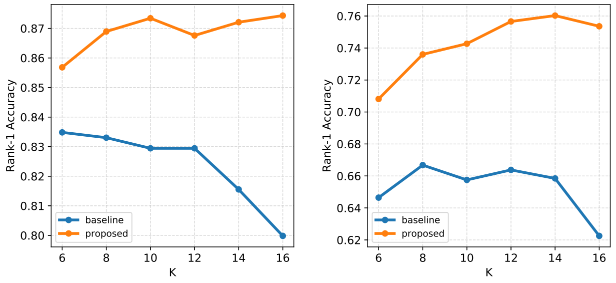
Dataset	unshared		shared	
	mAP	R-1	mAP	R-1
Market-1501	79.0	91.8	82.7	93.4
DukeMTMC	66.9	83.3	73.2	86.9
CUHK03	43.1	47.1	62.4	68.2
MSMT17	35.3	63.7	47.6	73.6

Table 2. Ablation study on shared/unshared classifier.

method	Market	DukeMTMC	MSMT17
k-reciprocal	209 s	152 s	11009 s
LBR (Ours)	41 s	24 s	423 s

Table 3. The elapsed time of re-ranking methods.

which can not satisfy the need of larger K. We only conduct experiments on MSMT17 and DukeMTMC-reID. Figure 6 shows that our approach can benefit from larger K, while the performance of vanilla baseline model even degrades when K increases. This phenomenon again validates our hypothesis that group-wise training is more advantageous with larger mini-batch. Because it can utilize holistic information of the whole batch for the training of a sample.



(a) DukeMTMC-reID (b) MSMT7

Figure 6. The trend of performance while K(#images per identity) varies.

4.5. Comparison with State-of-the-Art Methods

The proposed method is compared with state-of-the-art methods in this section. KR and LBR in the tables below denote k-reciprocal re-ranking and the proposed local blurring re-ranking, respectively.

Results on Market-1501 dataset. As shown in Table 4, our method achieves the best rank-1 accuracy among competitors, while mAP is slightly lower than SGGNN [37]. It must be highlighted that both SGGNN [37] and GSRW [36] undergo customized post-processing. After the refinement of LBR, our method outperforms them significantly. We further perform a comparison on the dataset with 500k distractors. The results are summarized in Table 5. As reported in the table, our method is robust to distractors. When disturbed by 100k distractors, the mAP/rank-1 accuracy of our method only decreases by 4.9%/2.5%. Note that the rank-1 accuracy is still over 90% in this case. While for the other four competitors, the degradations are much larger than ours. The performance gaps are even more significant when increasing the distractor size. Note that our method can still maintain over 90% rank-1 accuracy when disturbed by 100k distractors. This strongly demonstrates the robustness of our method.

Methods	Reference	Market-1501		
		mAP	R-1	R-5
GLAD [56]	ACMMM17	73.9	89.9	-
MLFN [3]	CVPR18	74.3	90.0	-
HA-CNN [22]	CVPR18	75.7	91.2	-
DuATM [39]	CVPR18	76.6	91.4	97.1
Part-aligned [42]	ECCV18	79.6	91.7	96.9
PCB [44]	ECCV18	77.4	92.3	97.2
GSRW [36]	CVPR18	82.5	92.7	96.9
SGGNN [37]	ECCV18	**82.8**	92.3	96.1
Mancs [47]	ECCV18	82.3	93.1	-
Proposed	-	82.7	**93.4**	**97.4**
Proposed(+ KR)	-	**90.6**	93.5	96.6
Proposed(+ LBR)	-	87.5	**94.1**	**97.5**

Table 4. Comparison with state-of-the-art methods on the Market-1501 dataset.

Results on DukeMTMC-reID dataset. The results on DukeMTMC-reID dataset are presented in Table 6. It can be seen that our method outperforms other state-of-the-arts significantly. Specifically, our approach gains 1.4% and 2% improvement over Mancs [47] in terms of mAP and rank-1 accuracy, respectively. After the refinement of LBR, our method even promotes rank-1 accuracy up to 90.0%.

Results on CUHK03 dataset. We only conduct experi-

Methods	Distractor Size							
	0		100k		200k		500k	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Zheng *et al.* [63]	59.9	79.5	$52.3_{\downarrow 7.6}$	$73.8_{\downarrow 5.7}$	$49.1_{\downarrow 10.8}$	$71.5_{\downarrow 8.0}$	$45.2_{\downarrow 14.7}$	$68.3_{\downarrow 11.2}$
APR [23]	62.8	84.0	$56.5_{\downarrow 6.3}$	$79.9_{\downarrow 4.1}$	$53.6_{\downarrow 9.2}$	$78.2_{\downarrow 5.8}$	$49.8_{\downarrow 13.0}$	$75.4_{\downarrow 8.6}$
TriNet [15]	69.1	84.9	$61.9_{\downarrow 7.2}$	$79.7_{\downarrow 5.2}$	$58.7_{\downarrow 10.4}$	$77.9_{\downarrow 7.0}$	$53.6_{\downarrow 15.5}$	$74.7_{\downarrow 10.2}$
Part-aligned [42]	79.6	91.7	$74.2_{\downarrow 5.4}$	$88.3_{\downarrow 3.4}$	$71.5_{\downarrow 8.1}$	$86.6_{\downarrow 5.1}$	$67.2_{\downarrow 12.4}$	$84.1_{\downarrow 7.6}$
Proposed	82.7	93.4	$77.8_{\downarrow 4.9}$	$90.9_{\downarrow 2.5}$	$75.5_{\downarrow 7.2}$	$89.3_{\downarrow 4.1}$	$71.9_{\downarrow 10.8}$	$87.1_{\downarrow 6.3}$

Table 5. Comparison with state-of-the-art methods on the Market-1501+500k dataset.

Methods	Reference	DukeMTMC		
		mAP	R-1	R-5
PSE [33]	CVPR18	62.0	79.8	89.7
HA-CNN [22]	CVPR18	63.8	80.5	-
MLFN [3]	CVPR18	62.8	81.0	-
DuATM [39]	CVPR18	64.6	81.8	90.2
GSRW [36]	CVPR18	66.4	80.7	88.5
SGGNN [37]	ECCV18	68.2	81.1	88.4
PCB+RPP [44]	ECCV18	69.2	83.3	-
Part-aligned [42]	ECCV18	69.3	84.4	92.2
Mancs [47]	ECCV18	71.8	84.9	-
Proposed	-	**73.2**	**86.9**	**93.9**
Proposed(+ KR)	-	**83.3**	88.3	92.0
Proposed(+ LBR)	-	79.6	**90.0**	**94.0**

Table 6. Comparison with state-of-the-art methods on the DukeMTMC-reID dataset.

Methods	Reference	CUHK03		
		mAP	R-1	R-5
SVDNet [43]	ICCV17	37.8	40.9	-
DPFL [6]	ICCV17	40.5	43.0	-
HA-CNN [22]	CVPR18	41.0	44.4	-
MLFN [3]	CVPR18	49.2	54.7	-
DaRe [54]	CVPR18	61.6	66.1	-
Proposed	-	**62.4**	**68.2**	**84.4**
Proposed(+ KR)	-	68.7	71.7	85.5
Proposed(+ LBR)	-	**71.7**	**74.3**	**85.6**

Table 7. Comparison with state-of-the-art methods on the CUHK03 dataset. We adhere to newly proposed evaluation protocol [65] and report results on manually labeled version of CUHK03.

Methods	Reference	MSMT17		
		mAP	R-1	R-5
GoogleNet [55]	CVPR18	23.0	47.6	65.0
PDC [55]	CVPR18	29.7	58.0	73.6
GLAD [55]	CVPR18	34.0	61.4	76.8
Proposed	-	**47.6**	**73.6**	**85.6**
Proposed(+ KR)	-	**60.8**	76.1	84.5
Proposed(+ LBR)	-	58.3	**79.0**	**85.8**

Table 8. Comparison with state-of-the-art methods on the MSMT17 dataset.

ments on the manually labeled subset of CUHK03 under the new protocol [65]. The results are reported in Table 7. It can be observed that our method achieves the best performance among compared methods. It outperforms DaRe [54] by 0.8% and 2.1% in terms of mAP and rank-1 accuracy, respectively.

Results on MSMT17 dataset. Since MSMT17 is released very recently, there is no other published work evaluated on it to our best knowledge. So we only compare our method with baselines reported by authors [55]. As shown in Table 8, our method outperforms these baselines dramatically. Specifically, it exceeds GLAD by 13.6% and 12.2% in terms of mAP and rank-1 accuracy, respectively. This verifies the scalability and the robustness of our method when applied in large scale scenarios. To clarify the superiority of our method, we remind readers that GLAD [56] performs pretty well on Market-1501 as recorded in Table 4.

5. Conclusion

Inspired by spectral clustering, we propose a novel feature transformation module to facilitate the learning of discriminative features which only involves several basic matrix operations. In contrast to most existing methods, our approach formulates the whole data batch as a similarity graph to capture potential relational structure. The emphasis is laid on optimizing group-wise similarities in our method. Furthermore, we extend the online operation to the post-processing stage. It conducts pre-clustering in the local neighborhood of the probe set which mitigates the ambiguity when retrieving. Though its simplicity, the proposed method brings prominent improvement over the strong baseline. Ablation studies on four benchmarks prove the effectiveness and scalability of our method.

Acknowledgement This work was supported in part by the National Key R&D Program of China (No. 2018YFB1004602), the Beijing Municipal Natural Science Foundation (No. Z181100008918010), the National Natural Science Foundation of China (No. 61836014, No. 61761146004, No. 61773375, No. 61602481). The authors would like to thank NVAIL for support.

References

[1] Song Bai, Zhichao Zhou, Jingdong Wang, Xiang Bai, Longin Jan Latecki, and Qi Tian. Ensemble diffusion for retrieval. In *ICCV*, 2017. 5

[2] Miguel Á Carreira-Perpiñán. Fast nonparametric clustering with gaussian blurring mean-shift. In *ICML*, 2006. 4

[3] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 7, 8

[4] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *NeurIPS Workshop*, 2016. 6

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017. 1

[6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2017. 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[8] WE Donath and AJ Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, pages 437–442, 1973. 2

[9] Michael Donoser and Horst Bischof. Diffusion processes for retrieval revisited. In *CVPR*, 2013. 5

[10] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. SphereReID: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, 2019. 1

[11] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 2

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 1, 4, 6, 8

[16] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 2016. 2

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[18] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *CVPR*, 2017. 5

[19] Herve Jegou, Hedi Harzallah, and Cordelia Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007. 2

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *ICCV*, 2014. 1, 5

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 2, 7, 8

[23] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv:1703.07220*, 2017. 8

[24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 2

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016. 2

[26] Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. In *AISTATS*, 2001. 2, 3

[27] Yair Movshovitzattias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *ICCV*, 2017. 1, 2

[28] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2002. 2

[29] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 1, 2

[30] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 2017. 2

[31] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR*, 2011. 2

[32] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop*, 2016. 5

[33] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018. 2, 8

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[35] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. SpectralNet: Spectral clustering using deep neural networks. In *ICLR*, 2018. 2

[36] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018. 3, 7, 8

[37] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018. 3, 7, 8

[38] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 2, 3

[39] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 7, 8

[40] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 1, 2

[41] X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, 2003. 3

[42] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 2, 7, 8

[43] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. SVDNet for pedestrian retrieval. In *ICCV*, 2017. 1, 8

[44] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1, 2, 5, 7, 8

[45] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised CNN segmentation. In *CVPR*, 2018. 2

[46] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. 2

[47] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018. 2, 7, 8

[48] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2, 5

[49] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 2

[50] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 2

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[52] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 3

[53] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *CVPR*, 2018. 2

[54] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018. 8

[55] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 5, 8

[56] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. GLAD: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017. 7, 8

[57] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 2

[58] Zhirong Wu, Alexei A. Efros, and Stella X. Yu. Improving generalization via scalable neighborhood component analysis. In *ECCV*, 2018. 2

[59] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 3

[60] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *ECCV*, 2018. 1

[61] Rui Yu, Zhichao Zhou, Song Bai, and Xiang Bai. Divide and fuse: A re-ranking approach for person re-identification. In *BMVC*, 2017. 2

[62] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 5

[63] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned CNN embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):13, 2017. 1, 8

[64] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, 2017. 1, 5

[65] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 2, 5, 6, 8

[66] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv:1708.04896*, 2017. 5