Interpretable and Generalizable Person Re-identification with Query-Adaptive Convolution and Temporal Lifting

Shengcai Liao[1(✉)] and Ling Shao[1,2]

[1] Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
{scliao,ling.shao}@ieee.org
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Abstract. For person re-identification, existing deep networks often focus on representation learning. However, without transfer learning, the learned model is fixed as is, which is not adaptable for handling various unseen scenarios. In this paper, beyond representation learning, we consider how to formulate person image matching directly in deep feature maps. We treat image matching as finding local correspondences in feature maps, and construct query-adaptive convolution kernels on the fly to achieve local matching. In this way, the matching process and results are interpretable, and this explicit matching is more generalizable than representation features to unseen scenarios, such as unknown misalignments, pose or viewpoint changes. To facilitate end-to-end training of this architecture, we further build a class memory module to cache feature maps of the most recent samples of each class, so as to compute image matching losses for metric learning. Through direct cross-dataset evaluation, the proposed Query-Adaptive Convolution (QAConv) method gains large improvements over popular learning methods (about 10%+ mAP), and achieves comparable results to many transfer learning methods. Besides, a model-free temporal cooccurrence based score weighting method called TLift is proposed, which improves the performance to a further extent, achieving state-of-the-art results in cross-dataset person re-identification. Code is available at https://github.com/ShengcaiLiao/QAConv.

1 Introduction

Person re-identification is an active research topic in computer vision. It aims at finding the same person as the query image from a large volume of gallery images. With the progress in deep learning, person re-identification has been largely advanced in recent years. However, when generalization ability becomes an important concern, required by practical applications, existing methods usually lack satisfactory performance, evidenced by direct cross-dataset evaluation

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58621-8_27) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12356, pp. 456–474, 2020.
https://doi.org/10.1007/978-3-030-58621-8_27

[10,53]. To address this, many transfer learning, domain adaptation, and unsupervised learning methods, performed on the target domain, have been proposed. However, these methods require heavy computations in deployment, limiting their application in practical scenarios where the deployment machine may have limited resources to support deep learning and users may cannot wait for a time-consuming adaptation stage. Therefore, improving the baseline model's generalization ability to support ready usage is still of urgent importance.

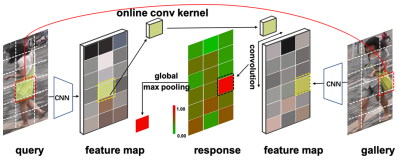

Fig. 1. QAConv constructs adaptive convolution kernels on the fly from query feature maps, and perform convolutions and max pooling on gallery feature maps to find the best local correspondences.

Fig. 2. Examples of the interpreted local correspondences from the outputs of the QAConv.

Most existing person re-identification methods compute a fixed representation vector, also known as a feature vector, for each image, and employ a typical distance or similarity metric (e.g. Euclidean distance or cosine similarity) for image matching. Without domain adaptation or transfer learning, the learned model is fixed as is, which is not adaptable for handling various unseen scenarios. Therefore, when generalization ability is a concern, it is expected to have an adaptive ability for the given model architecture.

In this paper, we focus on generalizable and ready-to-use person re-identification, through direct cross-dataset evaluation. Beyond representation learning, we consider how to formulate query-adaptive image matching directly in deep feature maps. Specifically, we treat image matching as finding local correspondences in feature maps, and construct query-adaptive convolution kernels on the fly to achieve local matching (see Fig. 1). In this way, the learned model benefits from adaptive convolution kernels in the final layer, specific to each image, and the matching process and result are interpretable (see Fig. 2), similar to traditional feature correspondence approaches [2,26]. Probably because finding local correspondences through query-adaptive convolution is a common process among different domains, this explicit matching is more generalizable than representation features to unseen scenarios, such as unknown misalignments, pose or viewpoint changes. We call this Query-Adaptive Convolution QAConv. To facilitate end-to-end training of this architecture, we further build a class memory module to cache feature maps of the most recent samples of each class, so as to compute image matching losses for metric learning.

Through direct cross-dataset evaluation without further transfer learning, the proposed method achieves comparable results to many transfer learning methods for person re-identification. Besides, to explore the prior spatial-temporal structure of a camera network, a model-free temporal cooccurrence based score weighting method is proposed, named Temporal Lifting (TLift). This is also computed on the fly for each query image, without statistical learning of a transition time model in advance. As a result, TLift improves person re-identification to a further extent, resulting in state-of-the-art results in cross-dataset evaluations.

To summarize, the novelty of this work include (i) a new deep image matching approach with query-adaptive convolutions, along with a class memory module for end-to-end training, and (ii) a model-free temporal cooccurrence based score weighting method. The advantages of this work are also two-fold. First, the proposed image matching method is interpretable, it is well-suited in handling misalignments, pose or viewpoint changes, and it also generalizes well in unseen domains. Second, both QAConv and TLift can be computed on the fly, and they are complementary to many other methods. For example, QAConv can serve as a better pre-trained model for transfer learning, and TLift can be readily applied by most person re-identification algorithms as a post-processing step.

2 Related Works

Deep learning approaches have largely advanced person re-identification in recent years [52]. However, due to limited labeled data and a big diversity in real-world surveillance, these methods usually have poor generalization ability in unseen scenarios. To address this, many unsupervised domain adaption (UDA) methods have been proposed [3,7,15,16,30,43,51,55,64], which show improved cross-dataset results than traditional methods, though requiring further training on the target domain. QAConv is orthogonal to transfer learning methods as it can provide a better baseline model for them (see Sect. 5.4 and Table 3).

There are many representation learning methods proposed to deal with viewpoint changes and misalignments in person re-identification, such as part-aligned feature representations [38,39,42,58], pose-adapted feature representations [33,57], human parsing based representations [14], local neighborhood matching [1,17], and attentional networks [18,24,25,31,35,49,50]. While these methods present high accuracy when trained and tested on the same dataset, their generalization ability to other datasets is mostly unknown. Besides, beyond representation learning, QAConv focuses on image matching via local correspondences.

Generalizable person re-identification was first studied in our previous works [10,53], where direct cross-dataset evaluation was proposed. More recently, Song et al. [36] proposed a domain-invariant mapping network by meta-learning, and Jia et al. [12] applied the IBN-Net [29] to improve generalizability, while QAConv is preliminarily reported in [19]. QAConv is orthogonal to methods of network design, for example, it can also be applied on the IBN-Net for improvements.

For deep feature matching, Kronecker-Product Matching (KPM) [34] computes a cosine similarity map by outer product for softly aligned element-wise

subtraction. Besides, Bilinear Pooling [22,37,40] and Non-local Neural Networks [44] also apply the outer product for part-aligned or self-attended representation learning. Different to the above methods, QAConv is a convolutional matching method but not simply outer product especially when its kernel size $s > 1$. It is explicitly designed for local correspondence matching, interpretation, and generalization, in a straightforward way without other branches.

For post-processing, re-ranking is a technique of refining matching scores, which further improves person re-identification [23,33,56,62]. Besides, temporal information is also a useful cue to facilitate cross-camera person re-identification [27,41]. While existing methods model transition times across different cameras but encounter difficulties in complex transition time distributions, the proposed TLift method applies cooccurrence constraint within each camera to avoid estimating transition times, and it is model-free and can be computed on the fly.

For memory based loss, ECN [64] proposed an exemplar memory which caches feature vectors of every instance for UDA. This makes the instance-level label inference convenient but limits its scalability. In contrast, class memory is independently designed [19], which is more efficient working in class level.

3 Query-Adaptive Convolution

3.1 Query-Adaptive Convolutional Matching

For face recognition and person re-identification, most existing methods do not explicitly consider the relationship between two input images under matching, but instead, like classification, they treat each image independently and apply the learned model to extract a fixed feature representation. Then, image matching is simply a distance measure between two representation vectors, regardless of the direct relationship between the actual contents of the two images.

In this paper, we consider the relationship between two images, and try to formulate adaptive image matching directly in deep feature maps. Specifically, we treat image matching as finding local correspondences in feature maps, and construct query-adaptive convolution kernels on the fly to achieve local matching. As shown in Fig. 1 and Fig. 3, to match two images, each image is firstly fed forward into a backbone CNN, resulting in a final feature map of size $[1, d, h, w]$, where d is the number of output channels, and h and w are the height and width of the feature map, respectively. Then, the channel dimension of both feature maps is normalized by the $\ell 2$-norm. After that, local patches of size $[s, s]$ at every location of the query feature map are extracted, and then reorganized into $[hw, d, s, s]$ as a convolution kernel, with input channels d, output channels hw, and kernel size $[s, s]$. This acts as a query-adaptive convolution kernel, with parameters constructed on the fly from the input, in contrast to fixed convolution kernels in the learned model. Upon this, the adaptive kernel can be used to perform a convolution on another feature map, resulting in $[1, hw, h, w]$ similarities.

Since feature channels are $\ell 2$-normalized, when $s = 1$, the convolution in fact measures the cosine similarity at every location of the two feature maps. Besides,

since the convolution kernel is adaptively constructed from the image content, these similarity values exactly reflect the local matching results between the two input images. Therefore, an additional global max pooling (GMP) operation will output the best local matches, and the maximum indices found by GMP indicate the best locations of local correspondences, which can be further used to interpret the matching result, as shown in Fig. 2. Note that GMP can also be done along the hw axis of the $[1, hw, h, w]$ similarity map. That is, seeking the best matches can be carried out from both sides of the images. Concatenating the output will result in a similarity vector of size $2hw$ for each pair of images.

3.2 Network Architecture

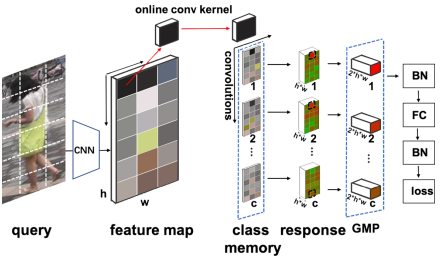

Fig. 3. Architecture of the QAConv. GMP: global max pooling. BN: batch normalization. FC: fully connection.

Fig. 4. Illustration of the proposed TLift approach.

The architecture of the proposed query-adaptive convolution method is shown in Fig. 3, which consists of a backbone CNN, the QAConv layer for local matching, a class memory layer for training, a global max pooling layer, a BN-FC-BN block, and, finally, a similarity output by a sigmoid function for evaluation in the test phase or loss computation in the training phase. The output size of the FC layer is 1, which acts as a binary classifier or a similarity metric, indicating whether or not one pair of images belongs to the same class. The two BN (batch normalization [11]) layers are all one-dimensional. They are used to normalize the similarity output and stabilize the gradient during training.

3.3 Class Memory and Update

We propose a class memory module to facilitate the end-to-end training of the QAConv network. Specifically, a $[c, d, h, w]$ tensor buffer is registered, where c is the number of classes. For each mini batch of size b, the $[b, d, h, w]$ feature map tensor of the mini batch will be updated into the memory buffer. We use a direct assignment update strategy, that is, each $[1, d, h, w]$ sample of class i from the mini batch will be assigned into location i of the $[c, d, h, w]$ memory buffer.

An exponential moving average update can also be used here. However, in our experience this is inferior to the direct replacement update. There might be two reasons for this. First, the replacement update caches feature maps of the most recent samples of each class, so as to reflect the most up-to-date state of the current model for loss computation. Second, since our task is to carry out image matching with local details in feature maps for correspondences, exponential moving average may smooth the local details of samples from the same class.

3.4 Loss Function

With a mini batch of size $[b, d, h, w]$ and class memory of size $[c, d, h, w]$, $b \times c$ pairs of similarity values will be computed by QAConv after the BN-FC-BN block. We use a sigmoid function to map the similarity values into $[0, 1]$, and compute the binary cross entropy loss. Since there are far more negative than positive pairs, to balance them and enable online hard example mining, we apply the focal loss [21] to weight the binary cross entropy. That is,

$$\ell(\theta) = -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{c}(1-\hat{p}_{ij}(\theta))^{\gamma} log(\hat{p}_{ij}(\theta)), \tag{1}$$

where θ is the network parameter, $\gamma = 2$ is the focusing parameter [21], and

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{if } y_{ij} = 1, \\ 1 - p_{ij} & \text{otherwise,} \end{cases} \tag{2}$$

where $y_{ij} = 1$ indicates a positive pair, while a negative pair otherwise, and $p_{ij} \in [0, 1]$ is the sigmoid probability.

4 Temporal Lifting

For person re-identification, to explore the prior spatial-temporal structure of a camera network, usually a transition time model is learned to measure the transition probability. However, for a complex camera network and various person transition patterns, it is not easy to learn a robust transition time distribution. In contrast, in this paper a model-free temporal cooccurrence based score weighting method is proposed, which is called Temporal Lifting (TLift). TLift does not model cross-camera transition times which could be variable and complex. Instead, TLift makes use of a group of nearby persons in each single camera, and find similarities between them.

Figure 4 illustrates the idea. A basic assumption is that people nearby in one camera are likely still nearby in another camera. Therefore, their corresponding matches in other cameras can serve as pivots to enhance the weights of other nearby persons. In Fig. 4, A is the query person. E is more similar than A' to A in another camera. With nearby persons B and C, and their top retrievals B' and C' acting as pivots, the matching score of A' can be temporally lifted since

it is a nearby person of B' and C', while the matching score of E will be reduced since there is no such pivot.

Formally, suppose A is the query person in camera Q, then, the set of nearby persons to A in camera Q is defined as $R = \{B|\Delta T_{AB} < \tau, \forall B \in Q\}$, where ΔT_{AB} is the within-camera time difference between persons A and B, and τ is a threshold on ΔT to define nearby persons. Then, for each person in R, cross-camera person retrieval will be performed on a gallery camera G by QAConv or other methods, and the overall top K retrievals for R are defined as the pivot set P. Next, each person in P acts as an ensemble point for 1D kernel density estimation on within-camera time differences in G, and the temporal matching probability between A and any person X in camera G will be computed as

$$p_t(A, X) = \frac{1}{|P|} \sum_{B \in P} e^{-\frac{\Delta T_{BX}^2}{\sigma^2}}, \tag{3}$$

where σ is the sensitivity parameter of the time difference. Then, this temporal probability is used to weight the similarity score of appearance models using a multiplication fusion as $p(A, X) = (p_t(A, X)+\alpha)p_a(A, X)$, where $p_a(A, X)$ is the appearance based matching probability (e.g. by QAConv), and α is a regularizer.

This way, true positives near pivots will be lifted, while hard negatives far from pivots will be suppressed. Note that this is also computed on the fly for each query image, without learning a transition time model in advance. Therefore, it does not require training data, and can be readily applied by many other person re-identification methods.

5 Experiments

5.1 Implementation Details

The proposed method is implemented in PyTorch, based upon an adapted version [61] of the open source person re-identification library (open-reid)[1]. Person images are resized to 384×128. The backbone network is the ResNet-152 [8] pre-trained on ImageNet, unless otherwise stated. The layer3 feature map of the backbone network is used, since the size of the layer4 feature map is too small. A 1×1 convolution with 128 channels is further appended to reduce the final feature map size. The batch size of samples for training is 32. The SGD optimizer is applied, with a learning rate of 0.001 for the backbone network, and 0.01 for newly added layers. They are decayed by 0.1 after 40 epochs, and the training stops at 60 epochs. The whole QAConv is end-to-end jointly trained, while class memory is updated only after the loss computation. Considering the memory consumption and the efficiency, the kernel size of QAConv is set to $s = 1$. Parameters for TLift are $\tau = 100$, $\sigma = 200$, $K = 10$, and $\alpha = 0.2$. They are not sensitive in a broad range, as analyzed in the Appendix.

A random occlusion module is implemented for data augmentation, which is similar to the random erasing [63] and cutout [5] methods (see Appendix for

[1] https://cysu.github.io/open-reid/.

comparisons). Specifically, a square area is generated with the size randomly sampled at most $0.8 \times width$ of the image. Then this square area is filled with white pixels. It is useful for QAConv because random occlusion forces QAConv to learn various local correspondences, instead of only saliency but easy ones. Beyond this, only a random horizontal flipping is used for data augmentation.

5.2 Datasets

Experiments were conducted on four large person re-identification datasets, Market-1501 [59], DukeMTMC-reID [6,60], CUHK03 [17], and MSMT17 [45]. The Market-1501 dataset contains 32,668 images of 1501 identities captured from 6 cameras. There are 12,936 images from 751 identities for training, and 19,732 images from 750 identities for testing . The DukeMTMC-reID is a subset of the multi-target and multi-camera pedestrian tracking dataset DukeMTMC [6]. It includes 1,812 identities and 36,411 images, where 16,522 images of 702 identities are used for training, and the remainings for test. The CUHK03 dataset includes 13,164 images of 1,360 pedestrians. We adopted the CUHK03-NP protocol provided in [62], where images of 767 identities were used for training, and other images of 700 identities were used for test. Besides, we used the detected subset for evaluation, which is more challenging. The MSMT17 dataset is the largest person re-identification dataset to date, which contains 4,101 identities and 126,441 images captured from 15 cameras. It is divided into a training set of 32,621 images from 1,041 identities, and a test set with the remaining images from 3,010 identities.

Cross-dataset evaluation was performed in these datasets, by training on the training subset of one dataset (except that in MSMT17 we used all images for training following [51,55]), and evaluating on the test subset of another dataset. The cumulative matching characteristic (CMC) and mean Average Precision (mAP) were used as the performance evaluation metrics. All evaluations followed the single-query evaluation protocol.

The Market-1501 and DukeMTMC-reID datasets are with frame numbers available, so that it is able to evaluate the proposed TLift method. The DukeMTMC-reID dataset has a good global and continuous record of frame numbers, and it is synchronized by providing offset times. In contrast, the Market-1501 dataset has only independent frame numbers for each session of videos from each camera. Accordingly we simply made a cumulative frame record by assuming continuous video sessions. After that, frame numbers were converted to seconds in time by dividing the Frames Per Second (FPS) in video records, where $\text{FPS} = 59.94$ for the DukeMTMC-reID dataset and $\text{FPS} = 25$ for the Market-1501 dataset.

5.3 Ablation Study

Some ablation studies have been conducted to understand the proposed method, in the context of direct cross-dataset evaluation between the Market-1501 and DukeMTMC-reID datasets. First, to understand the QAConv loss, several other

loss functions, including the classical softmax based cross entropy loss, the center loss [13,46], the Arc loss (derived from the ArcFace method [4] which is effective for face recognition), and the proposed class memory based loss, are evaluated for comparison. For these compared loss functions, the global average pooling of layer4 (better than layer3) of the ResNet-152 is used for feature representation, and the cosine similarity measure is adopted instead of the QAConv similarity. For the class memory loss, feature vectors are cached in memory instead of learnable parameters, and the same BN layer and Eq. (1) are applied after calculating the cosine similarity values between mini-batch features and memory features.

From results shown in Table 1, it is obvious that QAConv improves existing loss functions by a large margin, with 13.7%-19.5% improvements in Rank-1, and 9.6%-11.1% in mAP. Interestingly, large margin classifiers improves the softmax cross-entropy baseline when trained on the Market-1501 dataset, but do not have such improvements when trained on DukeMTMC-reID. This is probably due to many ambiguously labeled or closely walking persons in DukeMTMC-reID (see Sect. 5.5), which may confuse the strict large margin training. Note that the class memory based loss only performs comparable to other existing losses, indicating that the large improvement of QAConv is mainly due to the new matching mechanism, rather than the class memory based loss function. Besides, the Arc loss published recently is one of the best face recognition method, but it does not seem to be powerful when applied in person re-identification[2]. In our experience, the choice of loss functions does not largely influence person re-identification performance. Similar as in face recognition, existing studies [4,46] show that new loss functions do have improvements, but cannot be regarded as significant ones over the softmax cross entropy baseline. Therefore, we may conclude that the large improvement observed here is due to the new matching scheme, instead of different loss configurations (see Appendix for more analyses).

Table 1. Role of loss functions (%).

Method	Market→Duke		Duke→Market	
	Rank-1	mAP	Rank-1	mAP
Softmax cross-entropy	34.9	18.4	48.5	21.4
Arc loss [4]	35.3	17.1	48.9	21.4
Center loss [13,46]	38.9	22.1	48.8	22.0
Class memory loss	40.7	21.8	47.8	20.5
QAConv	**54.4**	**33.6**	**62.8**	**31.6**

Next, to understand the role of re-ranking (RR), the k-reciprocal encoding method [62] is applied upon QAConv. From results shown in Table 2, it

[2] We have tried different hyper parameters and reported the best results. The best margin values were found to be 0.5 on Market-1501 and 0.2 on DukeMTMC-reID.

can be seen that enabling re-ranking do improve the performance a lot, especially with mAP, which is increased by 18.8% under Market→Duke, and 19.6% under Duke→Market. This improvement is much more significant based on QAConv than that based on other methods as reported in [62]. This is probably because the new QAConv matching scheme better measures the similarity between images, which benefits the reverse neighbor based re-ranking method.

Furthermore, based on QAConv and re-ranking, the contribution of TLift is evaluated, compared to a recent method called TFusion (TF) [28], which is originally designed to iteratively improve transfer learning. From results shown in Table 2, it can be observed that employing TLift to explore temporal information further improves the results, with Rank-1 improved by 8.2%-10.2%, and mAP by 7.0%-8.8%. This improvement is complementary to re-ranking, so they can be combined. As for the existing method TFusion, it appears to be not stable, as a large improvement can be observed under Market→Duke, but little improvement can be obtained under Duke→Market, or even the mAP is clearly decreased[3]. This may be because TFusion is based on learning transition time distributions across cameras, which is not easy to deal with complex camera networks and person transitions as in the Market-1501 (various repeated presences per person in one camera). In contrast, the TLift method only depends on single-camera temporal information which is relatively more easy to handle. Note that TLift can also be generally applied to other methods for improvements, as shown in the Appendix. Besides, as shown in Table 2, directly applying TLift to QAConv without re-ranking also improves the performance a lot.

Table 2. Performance (%) of different post-processing methods.

Method	Market→Duke		Duke→Market	
	Rank-1	mAP	Rank-1	mAP
QAConv	54.4	33.6	62.8	31.6
QAConv + TLift	62.7	45.3	61.5	40.6
QAConv + RR [62]	61.8	52.4	68.5	51.2
QAConv + RR + TF [28]	**70.7**	**61.9**	68.6	47.2
QAConv + RR + TLift	70.0	61.2	**78.7**	**58.2**

Finally, to understand the effect of the backbone network, the QAConv results with the ResNet-50 as backbone are also reported in Tables 3 and 4, compared to the default ResNet-152 (denoted as QAConv_{50} and QAConv_{152}, respectively). As can be observed, a larger network ResNet-152 does have a better performance due to its larger learning capability. It can improve the Rank-1 accuracy over the QAConv_{50} by 1.3%–7.3%, and the mAP by 0.8%–5.5%. Besides, there are also consistent improvements in case of combining re-ranking and TLift. Hence, it seems

[3] Note that TFusion parameters were optimized on each dataset to get the best results, but for TLift we used fixed parameters for all datasets (see Appendix for analysis).

that this larger network, which contains more learnable parameters, does not have the overfitting problem when equipped with QAConv. Note that, though ResNet-152 is a very large network requiring heavy computation, in practice, it can be efficiently reduced by knowledge distillation [9].

5.4 Comparison to the State of the Arts

There are a great number of person re-identification methods since this is a very active research area. Here we only list recent results for comparison due to limited space. The cross-dataset evaluation results on the four datasets are listed in Tables 3 and 4. Considering that many person re-identification methods employ the ResNet-50 network, for a fair comparison, the following analysis is based on the QAConv_{50} results. Note that this paper mainly focuses on cross-dataset evaluation. Therefore, some recent methods performing unsupervised learning on the target dataset are not compared here, such as the TAUDL [15], UTAL [16], and UGA [48], and also partially due to the fact that they use single-camera target identity labels for training. There are mainly two groups of methods listed in Table 3, namely unsupervised transfer learning based methods, and direct cross-dataset evaluation based methods. The first group of methods require images from the target dataset for unsupervised learning, which are not directly comparable to the second one that directly evaluates on the target

Table 3. Comparison of the state-of-the-art cross-dataset evaluation results (%) with DukeMTMC-reID and Market-1501 as the target datasets.

Method	Training		Test: Duke		Training		Test: Market	
	Source	Target	R1	mAP	Source	Target	R1	mAP
PUL, TOMM18 [7]	Market	Duke	30.4	16.4	Duke	Market	44.7	20.1
TJ-AIDL, CVPR18 [43]	Market	Duke	44.3	23.0	Duke	Market	58.2	26.5
MMFA, BMVC18 [20]	Market	Duke	45.3	24.7	Duke	Market	56.7	27.4
CFSM, AAAI19 [3]	Market	Duke	49.8	27.3	Duke	Market	61.2	28.3
DECAMEL, TPAMI19 [54]	-	-	-	-	Multi	Market	60.2	32.4
PAUL, CVPR19 [51]	Market	Duke	56.1	35.7	Duke	Market	66.7	36.8
ECN, CVPR19 [64]	Market	Duke	63.3	40.4	Duke	Market	75.1	43.0
CDS, ICME19 [47]	Market	Duke	67.2	42.7	Duke	Market	71.6	39.9
ECN baseline, CVPR19 [64]	Market		28.9	14.8	Duke		43.1	17.7
PN-GAN, ECCV18 [32]	Market		29.9	15.8	-		-	-
QAConv_{50}	Market		48.8	28.7	Duke		58.6	27.2
QAConv_{152}	Market		54.4	33.6	Duke		62.8	31.6
QAConv_{50} + RR + TLift	Market		64.5	55.1	Duke		74.6	51.5
QAConv_{152} + RR + TLift	Market		70.0	61.2	Duke		78.7	58.2
MAR, CVPR19 [55]	MSMT	Duke	67.1	48.0	MSMT	Market	67.7	40.0
PAUL, CVPR19 [51]	MSMT	Duke	72.0	53.2	MSMT	Market	68.5	40.1
MAR baseline, CVPR19 [55]	MSMT		43.1	28.8	MSMT		46.2	24.6
PAUL baseline, CVPR19 [51]	MSMT		65.7	45.6	MSMT		59.3	31.0
QAConv_{50}	MSMT		69.4	52.6	MSMT		72.6	43.1
QAConv_{152}	MSMT		72.2	53.4	MSMT		73.9	46.6
QAConv_{50} + RR + TLift	MSMT		80.3	77.2	MSMT		86.5	72.2
QAConv_{152} + RR + TLift	MSMT		82.2	78.4	MSMT		88.4	76.0

dataset in consideration of real applications. The proposed QAConv method belongs to the second group. There are very few existing results of the same setting for the second group, except some baselines of other recent methods and the PN-GAN [32] which aims at augmenting source training data by GAN. For the comparison to the transfer learning methods, we consider that QAConv can serve as a better pre-trained model for them, and computing the RR + TLift on the fly is also more efficient than training on target dataset.

DukeMTMC-reID Dataset. As can be observed from Table 3, when trained on the Market-1501 dataset, QAConv achieves the best performance in the direct evaluation group with a large margin. When compared to transfer learning methods, QAConv also outperforms many of them except some very recent methods, indicating that QAConv enables the network to learn how to match two person images, and the learned model generalizes well in unseen domains without transfer learning. Besides, by enabling re-ranking and TLift, the proposed method achieves the best result among all except Rank-1 of CDS. Note that the re-ranking and TLift methods can also be incorporated into other methods, though. Therefore, we list their results separately. However, both of these are calculated on the fly without learning in advance, so together with QAConv, it appears that a ready-to-use method with good generalization ability can also be achieved even without further UDA, which is a nice solution considering that UDA requires heavy computation for deep learning in deployment phase.

When trained on MSMT17, QAConv itself beats all other methods except the transfer learning method PAUL. This is also the second best result among all existing methods taking DukeMTMC-reID as the target dataset, regardless of the training source. This clearly indicates QAConv's superiority in learning from large-scale data. It is preferred in practice in the sense that, when trained with large-scale data, there may be no need to adapt the learned model in deployment.

Market-1501 Dataset. With Market-1501 as the target dataset as shown in Table 3, similarly, when trained with MSMT17, QAConv itself also achieves the best performance among others except Rank-1 of ECN. This can be considered a large advancement in cross-dataset evaluation, which is a better evaluation strategy for understanding the generalization ability of algorithms. Besides, when equipped with RR+TLift, the proposed method achieves the state of the art, with Rank-1 accuracy of 86.5% and mAP of 72.2%. Note that this comparison is not in a sense of *fair*. We would like to share that beyond many recent efforts in UDA, enlarging the training data and exploiting on-the-fly computations in re-ranking and temporal fusion may also lead to good performance in unknown domain, with the advantage of no cost in training deep models everywhere.

CUHK03 Dataset. The CUHK03 and MSMT17 datasets present large domain gaps to others. For CUHK03, it can be observed from Table 4 that, with either Market-1501 or DukeMTMC-reID dataset as training set, QAConv without UDA performs better than a UDA method PUL [7], and fairly comparable to another recent transfer learning method CDS [47]. However, all methods perform not well

Table 4. Comparison of the state-of-the-art cross-dataset evaluation results (%) with CUHK03-NP (detected) and MSMT17 as the target datasets.

Method	Training		Test: CUHK03		Training		Test: MSMT	
	Source	Target	R1	mAP	Source	Target	R1	mAP
PUL, TOMM18 [7]	Market	CUHK03	7.6	7.3	-	-	-	-
CDS, ICME19 [47]	Market	CUHK03	9.1	8.7	-	-	-	-
PTGAN [45], CVPR18	-	-	-	-	Market	MSMT	10.2	2.9
ECN, CVPR19 [64]	-	-	-	-	Market	MSMT	25.3	8.5
QAConv_{50}	Market		9.9	8.6	Market		22.6	7.0
QAConv_{152}	Market		14.1	11.8	Market		25.6	8.2
PUL, TOMM18 [7]	Duke	CUHK03	5.6	5.2	-	-	-	-
CDS, ICME19 [47]	Duke	CUHK03	8.1	7.1	-	-	-	-
PTGAN [45], CVPR18	-	-	-	-	Duke	MSMT	11.8	3.3
ECN, CVPR19 [64]	-	-	-	-	Duke	MSMT	30.2	10.2
QAConv_{50}	Duke		7.9	6.8	Duke		29.0	8.9
QAConv_{152}	Duke		11.0	9.4	Duke		32.7	10.4
QAConv_{50}	MSMT		25.3	22.6	-	-	-	-
QAConv_{152}	MSMT		32.6	28.1	-	-	-	-

on the CUHK03 dataset. Only with the large MSMT17 data set as the source training data, the proposed method performs relatively better.

MSMT17 Dataset. With the MSMT17 as target, only QAConv does not require adaptation in Table 4. However, it performs better than PTGAN [45] and in part comparable to ECN [64]. This further confirms the generalizability of QAConv under large domain gaps, since without UDA it is already in part comparable to the state-of-the-art UDA methods. Note that TLift is not applicable on CUHK03 and MSMT17 due to no temporal information provided.

5.5 Qualitative Analysis and Discussion

A unique characteristic of the proposed QAConv method is its interpretation ability of the matching. Therefore, we show some qualitative matching results in Fig. 5 for a better understanding of the proposed method. The model used here is trained on the MSMT17 dataset, and the evaluations are done on the query subsets of the Market-1501 and DukeMTMC-reID datasets. Results of both positive pairs and hard negative pairs are shown. Note that only reliable correspondences with matching scores over 0.5 are shown, and the local positions are coarse due to the 24×8 size of the feature map. As can be observed from Fig. 5, the proposed method is able to find correct local correspondences for positive pairs of images, even if there are notable misalignments in both scale and position, pose/viewpoint changes, occlusions, and mix up of other persons, thanks to the local matching mechanism of QAConv instead of global feature representations. Besides, for hard negative pairs, the matching of QAConv still appears to be mostly reasonable, by linking visually similar parts or even the same person (may be ambiguously labeled or walking closely to other persons).

Note that the QAConv method gains the matching capability by automatic learning, from supervision of only class labels but not local correspondence labels.

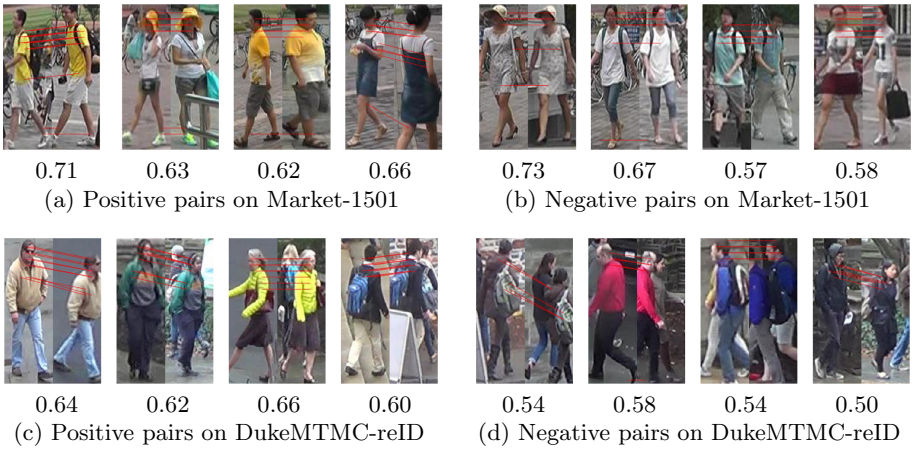

Fig. 5. Examples of qualitative matching results by the proposed QAConv method using the model trained on the MSMT17 dataset. Numbers represent similarity scores.

The QAConv network was trained on an NVIDIA DGX-1 server, with two V100 GPU cards. With the backbone network ResNet-50, the training time of QAConv on the DukeMTMC-reID dataset was 1.22 h. In contrast, the most efficient softmax baseline took 0.72 h for training. For deployment, the ECN [64] reported 1 h of transfer learning time with DukeMTMC-reID as target, while MAR [55] and DECAMEL [54] reported 10 and 35.2 h of total learning time, respectively, compared to the ready-to-use QAConv. For inference, with the DukeMTMC-reID dataset as target, QAConv took 26 s for feature extraction and 26 s for similarity computation. In contrast, the softmax baseline took 26 s for feature extraction and 0.2 s for similarity computation. Besides, the proposed method took 303 s for reranking, and 67 s for TLift. This is still efficient, especially for RR+TLift, compared to transfer learning for deployment. Therefore, the overall solution of QAConv+RR+TLift is promising in practical applications.

For further analysis on memory usage, please see the Appendix. As for the TLift, it can only be applied on datasets with good time records. Though this information is easy to obtain in real surveillance, most existing person re-identification datasets do not contain it. Another drawback of TLift is that it cannot be applied to arbitrary query images beyond a camera network, though once an initial match is found, it can be used to refine the search. Besides, it cannot help when there is no nearby person with the query.

6 Conclusion

In this paper, through extensive experiments we show that the proposed QAConv method is quite promising for person matching without further transfer learning, and it has a much better generalization ability than existing baselines. Though QAConv can also be plugged into other transfer learning methods as a better pre-trained model, in practice, according to the experimental results of this paper, we suggest a ready-to-use solution which works in the following principles. First, a large-scale and diverse training data (e.g. MSMT17) is required to learn a generalizable model. Second, a larger network (e.g. ResNet-152) benefits a better overall performance, which could be further distilled into smaller ones for efficiency. Finally, score re-ranking and temporal fusion model such as TLift can be computed on the fly in deployment, which can largely improve performance and they are more efficient to use than transfer learning.

Acknowledgements. This work was partly supported by the NSFC Project #61672521. The authors would like to thank Yanan Wang who helped producing several illustration figures in this paper, Jinchuan Xiao who optimized the TLift code, and Anna Hennig who helped proofreading the paper.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
2. Bay, H., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
3. Chang, X., Yang, Y., Xiang, T., Hospedales, T.M.: Disjoint label space transfer learning with common factorised space. Proc. AAAI Conf. Artif. Intell. **33**, 3288–3295 (2019)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
5. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
6. Ergys, R., Francesco, S., Roger, Z., Rita, C., Carlo, T.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV workshop on Benchmarking Multi-Target Tracking (2016)
7. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: clustering and fine-tuning. TOMM **14**(4), 83 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Hu, Y., Yi, D., Liao, S., Lei, Z., Li, S.Z.: Cross dataset person Re-identification. In: ACCV Workshop on Human Identification for Surveillance (HIS), pp. 650–664 (2014)

11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
12. Jia, J., Ruan, Q., Hospedales, T.M.: Frustratingly easy person re-identification: Generalizing person re-id in practice. In: British Machine Vision Conference (2019)
13. Jin, H., Wang, X., Liao, S., Li, S.Z.: Deep person re-identification with improved embedding and efficient training. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 261–267. IEEE (2017)
14. Kalayeh, M.M., Emrah, B., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1062–1071 (2018)
15. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 737–753 (2018)
16. Li, M., Zhu, X., Gong, S.: Unsupervised tracklet person re-identification. TPAMI **42**(7), 1770–1782 (2019)
17. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
18. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
19. Liao, S., Shao, L.: Interpretable and generalizable deep image matching with adaptive convolutions. CoRR abs/1904.10424v1 (23, April 2019), http://arxiv.org/abs/1904.10424v1
20. Lin, S., Li, H., Li, C.T., Kot, A.C.: Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In: The British Machine Vision Conference (BMVC) (2018)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
22. Lin, T.Y., Roychowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: IEEE International Conference on Computer Vision (2015)
23. Liu, C., Loy, C.C., Gong, S., Wang, G.: Pop: person re-identification post-rank optimisation. In: International Conference on Computer Vision (2013)
24. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Process. **26**(7), 3492–3506 (2017)
25. Liu, X., et al.: Hydraplus-net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 350–359 (2017)
26. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
27. Lv, J., Chen, W., Li, Q., Yang, C.: Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7948–7956 (2018)

28. Lv, J., Chen, W., Li, Q., Yang, C.: Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 7948–7956 (2018)
29. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 464–479 (2018)
30. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1306–1315 (2016)
31. Qian, X., Fu, Y., Jiang, Y.G., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5399–5408 (2017)
32. Qian, X., et al.: Pose-normalized image generation for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 650–667 (2018)
33. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 420–429 (2018)
34. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: End-to-end deep kronecker-product matching for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6886–6895 (2018)
35. Si, J., et al.: Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5372 (2018)
36. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 719–728 (2019)
37. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 402–419 (2018)
38. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 402–419 (2018)
39. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
40. Ustinova, E., Ganin, Y., Lempitsky, V.: Multi-Region bilinear convolutional neural networks for person re-identification. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (2017)
41. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: AAAI Conference on Artificial Intelligence (2019)
42. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, pp. 274–282. ACM (2018)
43. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2275–2284 (2018)

44. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
45. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)
46. Wen, Yandong., Zhang, Kaipeng., Li, Zhifeng, Qiao, Yu.: A discriminative feature learning approach for deep face recognition. In: Leibe, Bastian, Matas, Jiri, Sebe, Nicu, Welling, Max (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
47. Wu, J., Liao, S., Wang, X., Yang, Y., Li, S.Z., et al.: Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 886–891. IEEE (2019)
48. Wu, J., Yang, Y., Liu, H., Liao, S., Lei, Z., Li, S.Z.: Unsupervised graph association for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8321–8330 (2019)
49. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2119–2128 (2018)
50. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4733–4742 (2017)
51. Yang, Q., Yu, H.X., Wu, A., Zheng, W.S.: Patch-based discriminative feature learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3633–3642 (2019)
52. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep Learning for Person Re-identification: A Survey and Outlook. arXiv preprint arXiv:2001.04193 (2020)
53. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: International Conference on Pattern Recognition, pp. 34–39 (December 2014)
54. Yu, H.X., Wu, A., Zheng, W.S.: Unsupervised person re-identification by deep asymmetric metric embedding. In: IEEE Transactions on Pattern Analysis and Machine intelligence (2019)
55. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2148–2157 (2019)
56. Yu, R., Zhou, Z., Bai, S., Bai, X.: Divide and fuse: a re-ranking approach for person re-identification. In: The British Machine Vision Conference (BMVC) (2017)
57. Zhao, H., et al.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1077–1085 (2017)
58. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3219–3228 (2017)
59. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of IEEE International Conference on Computer Vision (2015)
60. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: International Conference on Computer Vision, pp. 3774–3782 (2017)

61. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camstyle: A novel data augmentation method for person re-identification. In: IEEE Transactions on Image Processing (2018)
62. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327 (2017)
63. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2020)
64. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 598–607 (2019)