Interpretable and Generalizable Deep Image Matching with Query-adaptive Convolutions

Anonymous Author(s)
Affiliation
Address
email

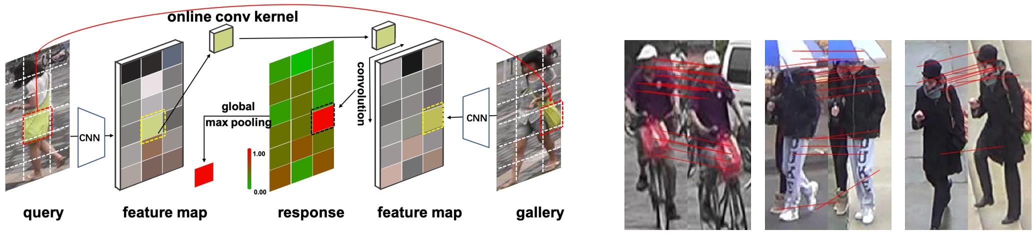
Abstract

For image matching tasks, such as face recognition and person re-identification, existing deep networks often focus on representation learning. However, without domain adaptation or 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 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 result is 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 image matching 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. The proposed method is preliminarily validated on the person re-identification task. Through direct cross-dataset evaluation without further transfer learning, it achieves better results than many transfer learning methods. Besides, a model-free temporal cooccurrence based score weighting method is proposed, which improves the performance to a further extent, resulting in state-of-the-art results in cross-dataset evaluations.

1 Introduction

Image matching is an important and fundamental task in computer vision, which has various applications ranging from image registration, 3D scene reconstruction, image retrieval, and so on. Traditionally, feature detection (e.g. SIFT [28], SURF [3]) and feature matching have been actively studied. This is particularly useful for matching unaligned images.

In terms of matching aligned or roughly aligned images, such as face or person images, representing images as canonical features is generally regarded as more useful than carrying out local feature matching following traditional computer vision methods. This is supported from traditional holistic features like PCA [43] and LDA [57], statistical representation of local features like LBP [2] and Gabor [23], to modern deep representations [42, 40, 31]. However, a precondition is that the images being represented should be well aligned, especially for traditional methods. Therefore, keypoint or landmark detection is usually applied in advance of the feature representation, but misalignment still occasionally occurs. Regarding person re-identification, automatic person detection still lacks satisfactory solutions. Researchers have also applied keypoint detection to facilitate person re-identification [56, 36], but person keypoint detection in surveillance is itself a challenge. On the other hand, attention based deep representation learning [30, 51, 25, 33, 49, 18, 27, 52, 44, 62, 37, 18, 50] goes beyond explicit alignment for feature extraction. It is able to find salient features in images for feature representation, so as to alleviate the requirement of precise alignment.



(a) query-adaptive convolution

(b) local correspondences

Figure 1: (a) Illustration of the proposed query-adaptive convolution (QAConv) based image matching method. We construct adaptive convolution kernels on the fly from query feature maps, and perform convolutions on gallery feature maps, followed by global max pooling to find the best local correspondences. (b) Examples of the interpreted local correspondences from the output of the proposed QAConv image matching method.

However, most existing methods compute a fixed representation vector, also known as a feature vector, for each image, and employ an ad-hoc 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. Even with an attention model, the learned attention mechanism might be limited in the source domain, but not very responsive to unseen image content. Therefore, when generalization ability is a concern, it is expected to have an adaptive ability for the given model architecture.

In this paper, 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 (a)). 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. 1 (b)), similar as the traditional feature matching [28, 3]. 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 thisarchitecture, we further build a class memory module to cache feature maps of the most recent samples of each class (see Fig. 2 (a)), so as to compute image matching losses for metric learning.

The proposed method is preliminarily validated on the person re-identification task, where domain adaptation is an active research area. Through direct cross-dataset evaluation without further transfer learning, the proposed method achieves better results than many transfer learning methods. Besides, to explore the prior spatial-temporal structure of a camera network, a model-free temporal cooccurrence based score weighting method is proposed, which we call Temporal Lifting (TLift, see Fig. 2 (b)). 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 evaluation.

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 does not require transfer learning, but can serve as a better pre-learned model for transfer learning methods, 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 [41, 46, 64, 36, 59, 15, 39, 44, 62, 37, 18, 50, 55, 4]. However, due to limited labeled data and a big diversity in real-world surveillance, deep person re-identification methods usually have poor generalization ability in unseen scenarios. To address this, a number of unsupervised transfer learning or domain adaption approaches have been proposed for person re-identification [32, 63, 7, 5, 53, 47, 16, 9, 17, 22]. However, they require further training on the target domain, though mostly unsupervised or self-supervised. In practical applications, the end-device may have limited computational power to support deep learning. Therefore, improving the baseline model's generalization ability is still of urgent importance.

There are a number of person re-identification approaches proposed to deal with pose or viewpoint changes, and misalignment, for example, part-based feature representations [41, 46, 39], pose-adapted feature representations [56, 36], human parsing based representations [15], local neighborhood matching [1], and attentional networks [30, 51, 25, 33, 27, 49, 52, 44, 62, 37, 18, 50]. While these methods present high accuracy when trained and tested on the same dataset, their generalization ability to other datasets is mostly unknown.

For post-processing, re-ranking is a technique of refining matching scores, which further improves person re-identification accuracy [24, 54, 65, 36]. Besides, temporal information is also a useful cue to facilitate cross-camera person re-identification [29, 45]. 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.

3 Query-adaptive Convolution

3.1 Image Matching via Query-adaptive Convolution

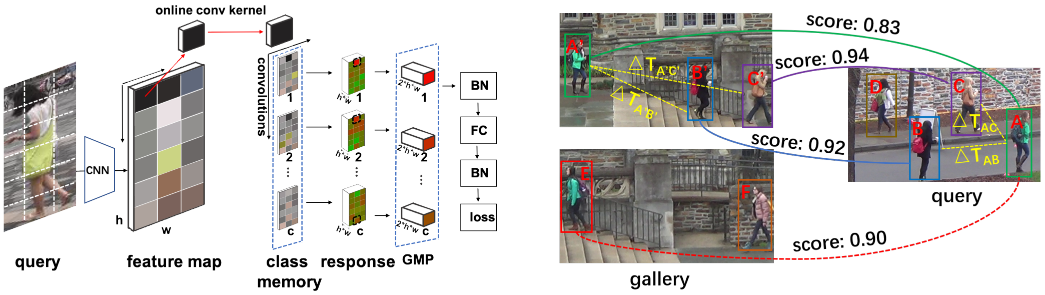
For face recognition and person re-identification, most existing methods do not directly consider the relationship between the two input images under matching, but instead, like classification, they treat each image independently and apply the learned model to extract a fixed feature representation. 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.

Therefore, 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 (a), each input image is firstly fed forwarded 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 under matching is normalized by the $\ell 2$-norm. After that, one of the feature maps is permuted and reshaped into $[h \times w, d, 1, 1]$ as a normalized convolution kernel, with input channels d, output channels $h \times w$, and kernel size 1×1. This acts as an 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 query-adaptive kernel can be used to perform a convolution on the normalized feature map of another image. The result is $[1, h \times w, h, w]$. Since feature channels are $\ell 2$-normalized, the convolution in fact measures the Cosine similarity on 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 illustrated in Fig. 1.

Note that, by matching only two images, the above process can also be done by matrix multiplication. However, when a batch of images is considered, a convolution is more suited and more efficient. Also note that GMP can also be done by reshaping the $[1, h \times w, h, w]$ similarities and maximizing along the $h \times w$ channels. That is, seeking the best matches can be carried out from both sides of the images. Concatenating the output will result in a $2 \times h \times w$ similarity vector for each pair of images.

3.2 Network Architecture

The architecture of the proposed query-adaptive convolution based image matching method is shown in Fig. 2 (a), which consists of a backbone CNN, the QAConv layer for local matching, a class



(a) architecture of QAConv

(b) temporal lifting

Figure 2: (a) Architecture of the proposed QAConv. GMP means global max pooling. (b) Illustration of the proposed TLift approach. 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 score of A' can be temporally lifted since it is a nearby person of B' and C', while the score of E will be reduced since there is no such pivot.

memory layer for training (introduced below), 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 [14]) 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

To train the QAConv image matching architecture, we need to form sufficient training image pairs. A natural way to do this is to use mini batches for training, and form image pairs within each mini batch. However, this is not efficient for sampling the whole training set and the convergence is slow, since there are N^2 possible combinations of all sample pairs, where N is the number of training images. Therefore, 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, after the loss computation (introduced below), 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. Considering that there are far more negative than positive pairs, to enable online hard negative mining while reducing the influence of the mass of negative pairs, we also apply the focal loss [21] to weight the binary cross entropy. That is,

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

where θ is the network parameter, γ is the focusing parameter (by default $\gamma = 2$ as in [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 probability output of the sigmoid function.

4 Temporal Lifting

For the task of person re-identification, to explore the prior spatial-temporal structure of a camera network, a model-free temporal cooccurrence based score weighting method is proposed, which we call Temporal Lifting (TLift). Fig. 2 (b) 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. 2 (b), 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

$$R = \{B | \Delta T_{AB} < \tau, \forall B \in Q\}, \tag{3}$$

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 with the QAConv similarity measures, and the top K retrievals are defined as the pivot set P. Then, each person in the pivot set P acts as an ensemble point for one-dimensional kernel density estimation on within-camera time differences on G, and the matching probability between A and any person X in camera G will be computed as

$$p_{AX} = \frac{1}{|P|} \sum_{B \in P} e^{-\frac{\Delta T_{BX}^2}{2\sigma^2}}, \tag{4}$$

where σ is the sensitivity parameter of the time difference. Then, this temporal probability is used to weight the similarity score of the QAConv using a multiplication fusion. In this way, true positives near to 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 statistical learning of 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. The QAConv has no hyper parameters. Parameters for TLift are $K = \tau = \sigma = 100$. We used an adapted version [64] of the open source person re-identification library (open-reid) package [1]. Person images are resized to 384×128. A random block module (see supplementary) is implemented for data augmentation, similar to the random erasing method [66]. The backbone network is the Resnet152 [10], pre-trained on ImageNet. We used the layer3 feature map for all the experiments, since the size of the layer4 feature map is too small. A 128-channel convolution 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. It is decayed by 0.1 once, monitored by the ReduceLROnPlateau module in PyTorch until convergence.

5.2 Datasets

Experiments were conducted on two large person re-identification datasets, Market-1501 [58] and DukeMTMC-reID [8, 60], with frame numbers available so that we were able to evaluate the proposed TLift method. 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

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

Table 1: Comparison of different backbone networks.

Backbones		Duke→Market		Market→Duke	
		Rank-1	mAP	Rank-1	mAP
Resnet50	H256	56.3	27.5	42.9	25.6
	H384	57.1	26.1	48.1	29.1
Resnet152	H256	55.0	**25.9**	47.4	27.3
	H384	**61.2**	**30.5**	**54.2**	**33.3**

pedestrian tracking dataset DukeMTMC [8]. It includes 1,812 identities and 36,411 images in which 16,522 images of 702 identities are used for training, 2,228 images of another 702 identities are used as query images, and the remaining 17,661 images are used as gallery images. Cross-dataset evaluation was performed in these two datasets, by training on the training subset of one dataset, and evaluating the performance 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 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 only has independent frame numbers for each session of videos from each camera. For several sessions from each camera, we roughly calculated the overall time frames of each session as offset, and made a cumulative record by assuming the video sessions were continuously recorded. After that, frame numbers were converted to seconds in time by dividing the Frames Per Second (FPS) in video records, where FPS=60 for the DukeMTMC-reID dataset and FPS=25 for the Market-1501 dataset.

5.3 Ablation Study

We first evaluated the influence of the backbone networks and the input image size. The results are shown in Table 1, where H256 means the input image size of 256×128, while H384 means 384×128. From the results, it can be observed that H384 performs slightly better than H256, especially for Market→Duke, and with the Resnet152 backbone. Regarding the backbones, for Duke→Market, the performances of all results are comparable. However, for the Market→Duke experiments, there is a good improvement of Resnet152 over Resnet50. Therefore, we used Resnet152 in the following experiments, with a 384×128 input image size. Note that, though Resnet152 is a very large network requiring heavy computation, in practice, it can be efficiently reduced by knowledge distillation [12].

With the same backbone network Resnet152 and input image size 384×128, we also implemented three baseline methods for comparison, including the classical softmax based cross entropy (CE) loss, the center loss [48], and the proposed class memory based loss with the Cosine similarity measure of the 128-dimensional feature vectors after global average pooling, instead of the QAConv similarity. The comparison results to the proposed QAConv are shown in Table 2. From these results, it is obvious that the proposed QAConv method improves the baselines by a large margin. Note that the class memory based loss only contributes slight improvements over other baselines, indicating that the large improvement of QAConv is mainly due to the new matching mechanism, rather than the class memory based loss function. One may argue that there are still other better choices than the baseline losses compared here, like Sphereface [26] and ArcFace [6]. However, from their studies, these losses do not provide significant improvements over the softmax cross entropy baseline, and, in our experience, the choice of loss functions does not markedly influence performance in person re-identification. Therefore, we may conclude that the large improvement observed here is due to the new method for image matching, instead of different loss configurations.

Furthermore, to understand the role of re-ranking (RR), we also applied the k-reciprocal encoding based re-ranking method [65]. The results are also listed in Table 2. As can be seen, all methods observe a large improvement when enabling the re-ranking. Additionally, it appears that the QAConv method gains much larger improvements than the other baselines. This is probably because the image matching mechanism of QAConv better measures the similarity between images, which benefits the reverse neighbor based re-ranking method.

Next, we evaluated the contribution of TLift, with results shown in the last row of Table 2. Again, we can observe a large improvement when employing TLift to explore temporal information. This improvement is complementary to re-ranking, so they can be combined.

Table 2: Comparison of different metric/loss layers and role of post-processing methods.

Metric / Loss	Re-rank	TLift	Duke→Market		Market→Duke	
			Rank-1	mAP	Rank-1	mAP
Softmax-CE			42.0	17.9	31.1	15.7
Center loss			40.6	18.3	33.5	17.6
Class memory			41.7	17.5	35.9	18.5
QAConv			**61.2**	**30.5**	**54.2**	**33.3**
Softmax-CE	✓		47.2	28.8	39.4	27.2
Center loss	✓		45.5	27.9	39.7	29.0
Class memory	✓		47.9	28.9	43.0	32.1
QAConv	✓		**66.6**	**50.3**	**61.4**	**52.5**
QAConv	✓	✓	**79.6**	**57.6**	**82.6**	**66.1**

Table 3: Comparison of state-of-the-art cross-dataset evaluation results (%). Transfer learning methods used the training set of the target dataset.

Method	Publication	Transfer learning	Duke→Market		Market→Duke	
			Rank-1	mAP	Rank-1	mAP
UMDL [32]	CVPR 2016	✓			18.5	7.3
Ver+ID [61]	TOMM 2017				25.7	12.8
PN-GAN [34]	ECCV 2018				29.9	15.8
PUL [9]	TOMM 2018	✓	44.7	20.1	30.4	16.8
CAMEL [53]	ICCV 2017	✓	54.5	26.3		
TJ-AIDL [47]	CVPR 2017	✓	58.2	26.5	44.3	23.0
MMFA [20]	BMVC 2018	✓			45.3	24.7
SPGAN [7]	CVPR 2018	✓	58.1	26.9	46.9	26.4
HHL [63]	ECCV 2018	✓	62.2	31.4	46.9	27.2
CFSM [5]	AAAI 2019	✓			49.8	27.3
BUC [22]	AAAI 2018	✓	66.2	38.3	47.4	27.5
ARN [19]	CVPRW 2018	✓			60.2	33.4
TAUDL [16]	ECCV 2018	✓	63.7	41.2	61.7	43.5
UTAL [17]	TPAMI 2019	✓	69.2	46.2	62.3	44.6
UDARTP [38]	arXiv 2018	✓	75.8	53.7	68.4	49.0
QAConv			61.2	30.5	54.2	33.3
QAConv+RR			66.6	50.3	61.4	52.5
QAConv+RR+TLift			**79.6**	**57.6**	**82.6**	**66.1**

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. The cross-dataset evaluation results of Duke→Market are listed in Table 3. As can be observed, our QAConv method without transfer learning performs better than four transfer learning methods, indicating that QAConv enables the network to learn how to match two images, and the learned model generalizes well in unseen domains. Besides, by enabling re-ranking and TLift, we achieve the state of the art, with 3.8% improvement in rank-1 and 3.9% in mAP. 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 domain adaptation.

For the cross-dataset evaluation of Market→Duke, the results are also listed in Table 3. As can be observed, the QAConv method without transfer learning performs better than 8 out of 12 recent transfer learning methods. This can be considered a large improvement in cross-dataset evaluation, which is a better evaluation strategy for understanding the generalization ability of algorithms. Besides, the final performance of QAConv with re-ranking and TLift also achieves the new state of the art on Market→Duke, with 14.2% improvement in rank-1 and 17.1% improvement in mAP.

Table 4: Comparison of state-of-the-art within-dataset evaluation results (%).

Method	Publication	Market-1501		DukeMTMC-reID	
		Rank-1	mAP	Rank-1	mAP
DSR [11]	CVPR 2018	83.6	64.2		
DML [55]	CVPR 2018	87.7	68.8		
CamStyle [64]	CVPR 2018	89.5	71.55	78.3	57.6
dMpRL [13]	TIP 2018			76.8	58.6
AACN [50]	CVPR 2018			76.8	59.3
MLFN [4]	CVPR 2018			81.2	62.8
AWTL [35]	CVPR 2018	89.5	75.7	79.8	63.4
HA-CNN [18]	CVPR 2018	91.2	75.7	80.5	63.8
DuATM [37]	CVPR 2018			81.8	64.6
PAN+RR [62]	TCSVT 2018			75.9	66.7
PCB [41]	ECCV 2018	93.8	81.6	83.3	69.2
Mancs [44]	ECCV 2018	93.1	82.3		
Part-aligned [39]	ECCV 2018			84.4	69.3
SPreID [15]	CVPR 2018	92.5	81.3	86.0	73.3
DGNet [59]	CVPR 2019			86.6	74.8
MGN [46]	ACMMM 2018	95.7	86.9	88.7	78.4
PSE+RR [36]	CVPR 2018	90.3	84.0	85.2	79.8
QAConv		93.7	83.3	88.3	76.7
QAConv+RR		95.4	94.1	91.5	90.5
QAConv+RR+TLift		**97.7**	**95.0**	**96.6**	**93.8**

Finally, for reference, we also list the within-dataset evaluation results on Market-1501 and DukeMTMC-reID in Table 4. As can be observed, the QAConv method is not the best one in within-dataset evaluation, indicating that good within-dataset evaluation performance may not be necessary for understanding an algorithm's generalization ability. We achieve the best result among compared methods with QAConv+RR+TLift, though.

5.5 Qualitative Analysis and Discussion

The unique characteristic of the proposed QAConv method is its interpretable matching results. Therefore, we show some qualitative matching results in Fig. 1 (b) for a better understanding of the proposed method. As can be observed, the proposed method is able to find correct local correspondences for positive image pairs, even if there are notable misalignments or pose/viewpoint changes. For more examples including negative pairs, please see the supplementary file.

One drawback of QAConv is that it requires more memory to run than other methods, and it needs to restore feature maps of images, rather than features, where feature maps are generally larger in size than representation features. Besides, TLift 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.

6 Conclusion

In this paper, beyond representation learning, we formulate image matching directly in deep feature maps and develop a deep image matching method called QAConv. It is able to find local correspondences in feature maps by constructing query-adaptive convolution kernels on the fly for local matching. A good property of QAConv is that its matching result is interpretable, and this explicit matching is more generalizable than representation features to unseen scenarios. A model-free temporal cooccurrence based score weighting method called TLift is also proposed, which achieves a large improvement with time frames. The proposed method is preliminarily validated on the person re-identification task, resulting in state-of-the-art results in cross-dataset evaluation. In future research, it would be interesting to apply QAConv to other image matching scenarios, such as face recognition.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[4] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[5] Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M Hospedales. Disjoint label space transfer learning with common factorised space. *arXiv preprint arXiv:1812.02605*, 2018.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[8] R. Ergys, S. Francesco, Z. Roger, C. Rita, and T. Carlo. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on Benchmarking Multi-Target Tracking*, 2016.

[9] H. Fan, L. Zheng, C. Yan, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMM*, 14(4):83, 2018.

[10] K. He, X. zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. Multi-pseudo regularized label for generated data in person re-identification. *IEEE Transactions on Image Processing*, 28(3):1391–1403, 2019.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[15] Mahdi M Kalayeh, B. Emrah, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[16] M. Li, X. Zhu, and S. Gong. Unsupervised person re-identification by deep learning tracklet association. *arXiv preprint arXiv:1809.02874*, 2018.

[17] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *TPAMI*, 2019.

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[19] Y. Li, F. Yang, Y. Liu, Y. Yeh, X. Du, and Y. Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. *arXiv preprint arXiv:1804.09347*, 2018.

[20] S. Lin, H. Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*, 2018.

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[22] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 2, 2019.

[23] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

[24] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *International Conference on Computer Vision*, 2013.

[25] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017.

[26] W. Liu, Y. Wen, Z. Yu, M. Li, R. Bhiksha, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, page 1, 2017.

[27] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[28] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[29] J. Lv, W. Chen, Q. Li, and C. Yang. 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*, pages 7948–7956, 2018.

[30] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[31] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, volume 1, page 6, 2015.

[32] P. Peng, T. Xiang, Y. Wang, P. Massimiliano, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

[33] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5399–5408, 2017.

[34] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–667, 2018.

[35] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6036–6046, 2018.

[36] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[37] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, 2018.

[38] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *arXiv preprint arXiv:1807.11334*, 2018.

[39] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018.

[40] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[41] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[42] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[43] Matthew A. Turk and Alex P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, March 1991.

[44] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018.

[45] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *AAAI Conference on Artificial Intelligence*, 2019.

[46] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 274–282. ACM, 2018.

[47] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *arXiv preprint arXiv:1803.09786*, 2018.

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[49] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. What-and-where to match: deep spatially multiplicative integration networks for person re-identification. *Pattern Recognition*, 76:727–738, 2018.

[50] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018.

[51] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[52] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2017.

[53] H. Yu, A. Wu, and W. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of IEEE International Conference on Computer Vision*, 2017.

[54] Rui Yu, Zhichao Zhou, Song Bai, and Xiang Bai. Divide and fuse: A re-ranking approach for person re-identification. *arXiv preprint arXiv:1708.04169*, 2017.

[55] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[56] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 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*, pages 1077–1085, 2017.

[57] Wenyi Zhao, Arvindh Krishnaswamy, Rama Chellappa, Daniel L Swets, and John Weng. Discriminant analysis of principal components for face recognition. In *Face Recognition*, pages 73–85. Springer, 1998.

[58] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[59] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint Discriminative and Generative Learning for Person Re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *international conference on computer vision*, pages 3774–3782, 2017.

[61] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018.

[62] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[63] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, 2018.

[64] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 2018.

[65] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

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