A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking

M. Saquib Sarfraz[1], Arne Schumann[2,1], Andreas Eberle[1], Rainer Stiefelhagen[1]
[1]Karlsruhe Institute of Technology, [2]Fraunhofer IOSB
Karlsruhe, Germany
{firstname.lastname}@kit.edu

Abstract

*Person re-identification is a challenging retrieval task that requires matching a person's acquired image across non-overlapping camera views. In this paper we propose an effective approach that incorporates both the fine and coarse pose information of the person to learn a discriminative embedding. In contrast to the recent direction of explicitly modeling body parts or correcting for misalignment based on these, we show that a rather straightforward inclusion of acquired camera view and/or the detected joint locations into a convolutional neural network helps to learn a very effective representation. To increase retrieval performance, re-ranking techniques based on computed distances have recently gained much attention. We propose a new unsupervised and automatic re-ranking framework that achieves state-of-the-art re-ranking performance. We show that in contrast to the current state-of-the-art re-ranking methods our approach does not require to compute new rank lists for each image pair (*e.g.*, based on reciprocal neighbors) and performs well by using simple direct rank list based comparison or even by just using the already computed euclidean distances between the images. We show that both our learned representation and our re-ranking method achieve state-of-the-art performance on a number of challenging surveillance image and video datasets. Code is available at* https://github.com/pse-ecn.

1. Introduction

Person re-identification (re-id) in non-overlapping camera views poses a difficult matching problem. Most previous solutions try to learn the global appearance of a persons using Convolutional Neural Networks (CNNs) by either applying a straightforward classification loss or using a metric learning loss. To better learn local statistics, the same has been applied to image regions, *e.g.* by using horizontal stripes or grids [19, 4]. Because of the inherent challenge of matching between different views and poses of a person, there is no implicit correspondence between the local regions of the images (see Figure 1). This correspondence can be established by explicitly using full body pose information for alignment [32] or locally through matching corresponding detected body parts [41, 42]. Using this local or global person description by incorporating the body pose or body parts information can strongly benefit person re-id.

Figure 1. Camera view and body pose can significantly alter the visual appearance of a person. A different view might show different aspects of the clothing (*e.g.* a backpack) while an altered pose may lead to body parts (*e.g.* arms or legs) being located at different positions of the image. Pose information can also help guide the attention of a re-id approach in case of mis-aligned detections.

In this work we show that incorporating a simple cue of the person's coarse pose (*i.e.* the captured view with respect to the camera) or the fine body pose (*i.e.* joint locations) suffice to learn a very discriminative representation with a simple classification loss. We present an appealing design choice to incorporate these cues and show its benefit in the performance gain over state-of-the-art methods on large and challenging surveillance benchmarks. We demonstrate that learning and combining view specific feature maps on a standard underlying CNN architecture results in a significantly better re-id embedding. Similarly an incorporation of body joint locations as additional input channels helps to increase the re-id accuracy.

For improving person retrieval, after computing the initial distances, a re-ranking step can often improve ranking quality by a good margin. Re-ranking has seen a renewed

interest in recent years [23, 9, 15, 37, 48]. The re-ranking problem is formulated as re-estimating the distances between probe and gallery images such that more correct results are ranked at the top of the returned lists. In recent proposals this was generally achieved by exploiting the similarity of the lists of top k nearest neighbors of both the probe and gallery image in question. Among the state-of-the-art re-ranking methods these neighborhood lists are often recomputed for each image pair, based on the common or reciprocal neighbors [36, 1, 48]. This makes it more computationally demanding to recompute the distances between these varying length lists.

A second contribution of this work is a new re-ranking method that introduces the concept of expanded cross neighborhood distance. The method aggregates the distances of close neighbors of the probe and the gallery image, where the distance can simply be the direct euclidean distance or the distance based on the rank lists. We show that within this more general framework of re-ranking simple rank list comparison based on the directly obtained rank lists achieves state-of-the-art re-ranking performance without the requirement to recompute new rank lists.

In summary, our contributions are threefold: 1) We propose a new CNN embedding which incorporates coarse and fine-grained person pose information. 2) We propose a new unsupervised and automatic re-ranking method that achieves larger re-ranking improvements than previous methods. 3) Our pose-sensitive person re-id model and our re-ranking method set a new state-of-the-art on four challenging datasets. We also demonstrate the scalability of our approach with very large gallery sizes and its performance for person search in full camera images.

2. Related Work

In recent years many state-of-the-art re-id results have been achieved by approaches relying on feature embeddings learned through CNNs [41, 10, 40, 20]. We focus our discussion of related approaches to those which include a degree of pose information, as well as re-ranking methods.

Re-Id using Pose A person's body pose is an important cue for successful re-identification. The popular SDALF feature by Farenza *et al*. [8] uses two axes dependent on the body's pose to derive a feature description with pose invariance. Cho *et al*. [6] define four view angles (front, left, right, back) and learn corresponding matching weights to emphasize matching of same-view person images. A more fine-grained pose representation based on Pictorial Structures was first used in [5] to focus on matching between individual body parts. More recently, the success of deep learning architectures in the context of re-id has lead to several works that include pose information into a CNN-based matching. In [43] Zheng *et al*. propose to use a CNN-based external pose estimator to normalize person images based on their pose. The original and normalized images are then used to train a single deep re-id embedding. A similar approach is described by Su *et al*. in [32]. Here, a sub-network first estimates a pose map which is then used to crop the localized body parts. A local and a global person representation are then learned and fused. Pose variation has also been addressed by explicitly detecting body parts through detection frameworks [41], by relying on visual attention maps [25], or body part specific attention modeling [42].

In contrast to our proposed method, these works mostly rely only on fine-grained pose information. Furthermore, these approaches either include pose information by explicitly normalizing their input images or by explicitly modeling part localization and matching these in their architecture. In contrast to this, our approach relies on confidence maps generated by a pose estimator which are added as additional channels to the input image. This allows a maximum degree of flexibility in the learning process of our CNN and leaves it to the network to learn which body parts are relevant and reliable for re-id. Apart from this fine-grained pose information we show that a more coarse pose cue turns out to be even more important and can be effectively used to improve the re-id performance.

Re-Ranking In the recent years, re-ranking techniques are drawing more and more attention in the area of person re-id. Shen *et al*. [30] use k-nearest neighbors (k-NN) to produce new rank lists and recompute distances based on these. Garcia *et al*. [9] propose to jointly learn the context and content information in the ranking list to remove candidates in the top neighbors and improve performance of person re-id. [15] extends this to revise the initial ranking list with a new similarity obtained from fusion of content and contextual similarity.

Li *et al*. [18] first proposed to use the relative information of common nearest neighbors of each image to improve re-ranking. Ye *et al*. [36] combined the common nearest neighbors of global and local features as new queries and revise the initial ranking list by aggregating these into new ranking lists. Using the similarity and dissimilarity cues from the neighbors of different baseline methods [37] proposes a ranking aggregation algorithm to improve person re-id. In contrast to common neighbors, Jegou *et al*. [14] use reciprocal neighbors (*i.e*. common neighbors that reciprocate in a k-neighborhood sense) and propose to compute a contextual dissimilarity measure (CDM). [24] formally uses the k-reciprocal neighbors to compute ranking lists. Most recent state-of-the-art re-ranking methods are based on computing these rank list comparisons using a generalized Jaccard distance. To overcome the associated complexity of computing intersection and unions of underlying variable length lists, Sparse Contextual Activation (SCA) [1] encodes the neighborhood set into a sparse vector and then computes the dis-

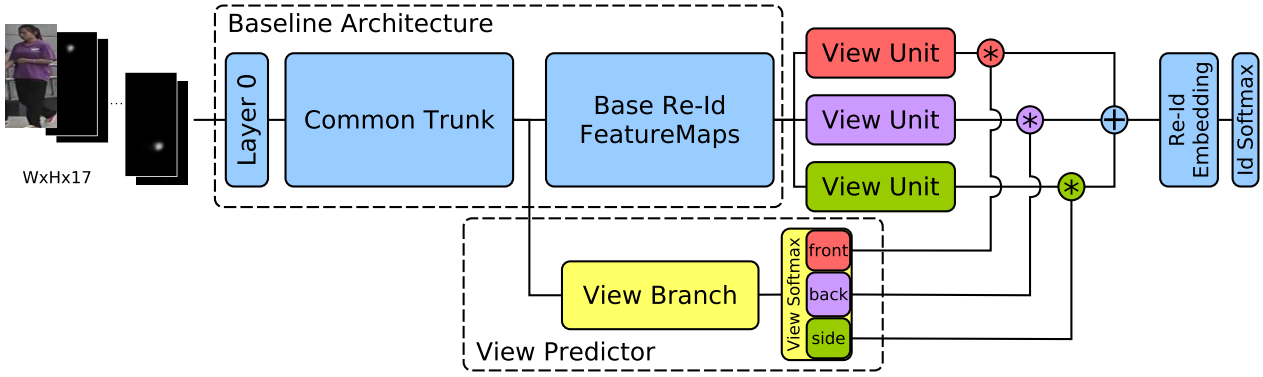

Figure 2. Overview of our pose sensitive embedding (PSE) architecture. As baseline architecture we employ either ResNet-50 or Inception-v4. Pose information is included through detailed body joint maps in the input, as well as through a coarse view predictor.

tance. To reduce the false positives and noise in the original ranked lists, more context is included by forming new rank lists based on reciprocal neighbors [14][24][48]. Zhong et al [48] use the k-reciprocal lists and compute the Jaccard distance by using SCA encoding. They then propose to fuse this distance with the original distance to obtain the final re ranking. Note, while reciprocal list based comparisons provides the current best re-ranking scores, it requires an additional complexity of recomputing the reciprocal rank lists for each image pair.

In contrast to common or reciprocal neighbors and producing new rank lists based on these, we propose the concept of expanded neighbors and aggregating their cross distances among the images in a pair. We show that this results in a more effective re-ranking framework while not requiring to re-compute new rank-lists for each image pair.

3. Pose-Sensitive Embedding

A person's pose and orientation to the camera can greatly affect the visual appearance in the image. Explicitly including this information into the learning process of a re-id model can thus often increase the resulting accuracy. Previous works have relied on either fine-grained pose information (*e.g.* joint keypoints) or coarse information (*e.g.* orientation to the camera). In this section we describe two new methods for including both levels of granularity into a pose-sensitive embedding. Both methods can be simultaneously incorporated into the same baseline CNN architecture and our experiments show that a combination of the two achieves a higher accuracy than either one alone. An overview of our CNN architecture with both types of pose information is depicted in Figure 2.

3.1. View Information

We use the quantization ['front', 'back', 'side'] of a person's orientation to the camera as coarse pose information. Since this information is dependent on the camera, as well as the person, we call it *view information* in the remainder of this work.

Our inclusion of view information into the re-id embedding is based on our prior work [27] on semantic attribute recognition. A ternary view classifier is added as a side-branch of our main re-id CNN. The tail part of the main CNN is then split into three equivalent units by replicating existing layers. The view classifier's three view prediction scores are used to weight the output of each of these units. This modulates the gradient flowing through the units, *e.g.* for a training sample with a strong 'front' prediction, mainly the unit weighted by the front-weight will contribute to the final embedding and thus only this unit will receive a strong gradient update for the current training sample. This procedure allows each unit to learn a feature map specialized for one of the three views. Importantly, and in contrast to [27], we do not weight and fuse final embeddings or prediction vectors but apply the weights to full feature maps which are then combined into the final embedding. This achieves a more robust representation.

We cannot generally assume to have view annotations available on the re-id dataset we want to train our embedding on. Thus, we pretrain a corresponding view classifier on the separate RAP [17] pedestrian dataset which provides such annotations. We then directly transfer the classifier to our re-id model. Low-level features (*i.e.* early layers) can be shared between the view predictor and the re-id network in order to reduce model complexity.

In our default ResNet-50 architecture the view predictor branch is split off from the main network after the third dimensionality reduction step (*i.e.* at feature map dimensions $28 \times 28 \times 256$). We then apply three consecutive convolutions with step sizes 2, 2, and 5 to reduce the dimension further (to $1 \times 1 \times 1024$). The resulting feature vector is used to predict the view using a three-way softmax. As view units we use three replications of the ResNet Block-

4. The $7 \times 7 \times 2048$ dimensional fused output of the units is pooled and fed to a fully connected layer which yields our 1536 dimensional embedding.

3.2. Full Body Pose

As fine-grained representation of a person's pose we use the locations of 14 main body joint keypoints. To obtain this information we use the off-the-shelf DeeperCut [12] model. In contrast to prior use of pose information for re-id, we do not use this information to explicitly normalize the input images. Instead, we include the information into the training process by adding an additional input channel for each of the 14 keypoints. These channels serve to guide the CNN's attention and help learn how to best apply the body joint information into the resulting embedding. To further increase this flexibility, we do not rely on the final keypoint decisions of the DeeperCut approach, but instead provide our re-id CNN with the full confidence map for each keypoint. This prevents any erroneous input based on hard keypoint decisions and leaves our model the chance to compensate for, or at least recognize, unreliable pose information.

3.3. Training Details

We initialize all our CNNs with weights pretrained for ImageNet classification. In order to train a model with view information (Section 3.1) we start by fine-tuning only the view-predictor branch on the RAP dataset [17]. Next we train only the view units and the final person identity classification layer on the target re-id dataset. The weights of the view predictor and all layers before the view units are fixed for this stage. This allows the randomly initialized view units and final layers to adapt to the existing weights of earlier layers.

When training an embedding including full body pose information (Section 3.2) the ImageNet weights do not match the size of our input, due to the additional 14 keypoint channels. To adapt the network for 17 channel inputs we start our training by fine-tuning only the first layer (Layer 0 in Figure 2), and the final person identity classification layer which are both initialized randomly. The remainder of the network remains fixed. Once these two layers are adapted to the rest of the network (*i.e.* convergence is observed), we proceed by fine-tuning the entire network.

For our final pose sensitive embedding (PSE) we combine both types of pose information into one network as depicted in Figure 2. We initialize our training with the full body pose model described in the previous section and add the view predictor onto it. The view predictor is fine-tuned on the RAP dataset with pose maps and can benefit from the additional full body pose information. Further fine-tuning of the re-id elements of the network is then performed on the target re-id dataset as described above.

For all our CNN embeddings we employ the same training protocol. Input images are normalized to channel-wise zero-mean and a standard variation of 1. Data augmentation is performed by resizing images to 105% width and 110% height and randomly cropping the training sample, as well as random horizontal flip (this is the main reason why we do not differentiate between left and right side views). Training is performed using the Adam optimizer at recommended parameters with an initial learning rate of 0.0001 and a decay of 0.96 every epoch.

4. Expanded Cross Neighborhood Distance based Re-Ranking

In this section we introduce the concept of Expanded Cross Neighborhood (ECN) distance which can provide a very high boost in performance while not strictly requiring rank list comparisons. We show that, for an image pair, accumulating the distances of only the immediate two-level neighbors of each image with the other image results in a promising re-ranking. Within this cross neighborhood based distance framework, the underlying accumulated distances can be just the original euclidean distances or the re-calculated rank-list based distances. We also show that within this framework using a simple list comparison measure on the initially obtained rank lists achieves the state of the art re-ranking performance. Our proposal is fully automatic, unsupervised and can work well without requiring to compute new rank lists.

Formally, given a probe image p and a gallery set G with B images $G = \{g_i \mid i = 1, 2, \cdots, B\}$, the euclidean distance between p and each of the gallery g_i is $\|\mathbf{p} - \mathbf{g_i}\|_2^2$. Computing pairwise distance between all images in the gallery and probe sets, the initial ranking $\mathcal{L}(p, G) = \{g_1^o, \cdots, g_B^o\}$ for each image is then obtained by sorting this distance in an increasing order.

Given such initial rank lists $\mathcal{L}$ of all the images in the gallery and probe sets, we define the expanded neighbors of the probe p as the multiset $N(p, M)$ such that:

$$N(p, M) \leftarrow \{N(p, t), N(t, q)\} \tag{1}$$

where $N(p, t)$ are the top t immediate neighbors of probe p and $N(t, q)$ contains the top q neighbors of each of the elements in set $N(p, t)$:

$$\begin{aligned} N(p, t) &= \{g_i^o \mid i = 1, 2, \cdots, t\} \\ N(t, q) &= \{N(g_i^o, q), \cdots, N(g_t^o, q)\} \end{aligned} \tag{2}$$

A similar expanded neighbors multiset can be obtained for each of the gallery images $N(g_i, M)$ in terms of its immediate neighbors and their neighbors. The total number of neighbors M in the set $N(p, M)$ or $N(g_i, M)$ is $M = t + t \times q$. Finally the Expanded Cross Neighborhood

(ECN) distance of an image pair (p, g_i) is defined as

$$ECN(p, g_i) = \frac{1}{2M} \sum_{j=1}^{M} d(pN_j, g_i) + d(g_i N_j, p) \quad (3)$$

where pN_j is the jth neighbor in the probe expanded neighbor set $N(p, M)$ and $g_i N_j$ is the jth neighbor in the ith gallery image expanded neighbor set $N(g_i, M)$. The term $d(\cdot)$ is the distance between that pair. One can see that the ECN distance, above, just aggregates the distances of the expanded neighbors of each of the image in pair with the other. While we show in our evaluation that using the direct euclidean distance in Equation 3 results in a similar improvement in the rank accuracies, one can also use a more robust rank-list based distance to further enhance the performance, especially in terms of the mean average precision (mAP). These distances can be computed directly from the initial paired distance matrix or the resulting initial rank lists. Recent re-ranking proposals use the Jaccard distance for the list comparison which is computationally expensive, here we propose to use a rather simple list comparison similarity measure proposed by Jarvis and Patrick [13], and also successfully employed in a face verification task in [28]. The list similarity is measured in terms of the position of top *K* neighbors of the two lists. For a rank list with B images, let $pos_i(b)$ denote the position of image b in the ordered rank list $\mathcal{L}_i$. In terms of considering only the first K neighbors in the list, the Rank-list similarity R is given by:

$$R(\mathcal{L}_i, \mathcal{L}_j) = \sum_{b=1}^{B} [K + 1 - pos_i(b)]_+ \times [K + 1 - pos_j(b)]_+ \quad (4)$$

Here, $[\cdot]_+ = max(\cdot, 0)$. This measure ensures to base similarity in terms of top K neighbors while taking into account their position in the list. From an implementation point, this rank list similarity can effectively be computed from the initially obtained rank lists by single matrix multiplication and addition operations. To use this in Equation 3, we convert it into the distance $d = 1 - R^*$ where R^* denotes the minmax scaling of values in R. Finally the parameters t, q and K (in case of using the rank-list distances) for computing the final ECN distance are set to $t = 3$, $q = 8$ and $K = 25$. while we show that these parameters choices are very stable in terms of performance on a number of different sized datasets, one can intuitively also see that using the strongest top neighbors in the first level (t) and expanding these to few more at the second level (q) makes sense. Note since our neighbors' of neighbor expansion only looks for the first and second level of neighbors, we do not need to compute an expensive KD-tree or neighborhood graphs to get these expanded sets in Equation 1, we can readily obtain these from the initially computed ordered rank list matrix.

5. Evaluation

We report results using the standard cross camera evaluation in the single-query setting. Accuracy is measured by rank scores, obtained from cumulated matching characteristics (CMC), and mean average precision (mAP).

Datasets: We evaluate our approach on four datasets, Market-1501 [44] (Market), Duke-MTMC-reID [26] (Duke), MARS [31] and PRW [45].

The Market-1501 (Market) dataset consists of 32,668 bounding boxes of 1,501 distinct persons generated by a person detector on videos from six cameras. 751 persons are used for training and 750 for testing. The training set contains 12,936 images, the gallery set 19,732 images, and the query set has 3,368 images.

The Duke-MTMC-reID (Duke) dataset is created from data of eight cameras. Of 1,812 people in the data 1,404 occur in more than one camera. Training and test sets both consist of 702 persons. The training set includes 16,522 images, the gallery 17,661 images, and the query set 2,228 image. Person bounding boxes in the Duke dataset are manually annotated.

The MARS dataset consists of 20,478 tracklets of 1,261 re-occurring persons. Including 3,248 distractor tracklets this brings the total number of person images in the dataset to 1,191,993 with a train/test split of 509,914/681,089 images of 625 and 636 persons, respectively. This dataset is well suited to evaluate the performance of a re-id approach for person track retrieval.

The PRW dataset consists of 11,816 frames of video data. The images are annotated with 43,110 person bounding boxes of which 34,304 are assigned one of 932 person IDs. For training 5,134 frames including 482 different persons are available. At test time 2,057 cropped query images of persons must be found in a gallery of 6,112 full images. The PRW dataset allows for an evaluation of the robustness of a re-id method to false positive or mis-aligned person detections.

In order to compare to related approaches we split our evaluation into three parts. In Sections 5.1 and 5.2 we investigate key components of our pose-sensitive embedding and re-ranking, respectively. In Section 5.3 we compare our proposed embedding and re-ranking with state-of-the-art approaches. We also demonstrate the robust performance of our approach against detector errors and its scalability for very large galleries.

5.1. Study of Pose Information

We investigate the usefulness of including different granularities of pose information into the CNN by performing separate experiments with only view information, only pose information, and a combination of both. Experiments are performed on Market and Duke. To show that our proposal is not strictly dependent on the underlying CNN ar-

CNN	Method	Market-1501					Duke				
		mAP	R-1	R-5	R-10	R-50	mAP	R-1	R-5	R-10	R-50
Inception-v4	Baseline	51.9	75.9	89.8	92.5	97.3	36.6	61.8	74.8	79.8	89.4
	Views only	61.9	81.5	92.3	94.9	98.1	40.3	62.7	76.6	81.1	90.3
	Pose only	60.9	81.7	91.8	94.4	97.9	48.2	70.5	81.9	86.1	92.7
	PSE	**64.9**	**84.4**	**93.1**	**95.2**	**98.4**	**50.4**	**71.7**	**83.5**	**87.1**	**93.1**
ResNet-50	Baseline	59.8	82.6	92.4	94.9	98.2	50.3	71.5	83.1	87.0	94.1
	Views only	66.9	**88.2**	**95.4**	**97.2**	98.9	56.7	76.9	87.3	90.7	95.7
	Pose only	61.6	82.8	93.1	95.5	98.3	53.1	73.4	84.5	88.1	94.3
	PSE	**69.0**	87.7	94.5	96.8	**99.0**	**62.0**	**79.8**	**89.7**	**92.2**	**96.3**

Table 1. Comparison of different types of pose information. While views and full body pose individually lead to notable improvements, a combination of both often results in further improvements.

chitecture, besides using our main ResNet-50 base CNN, we also show results on the popular Inception-v4 CNN. For Inception-v4, the view predictor is branched out at the earlier Reduction-A block and view units are similarly added by using three Inception-C blocks at the end. Results of our experiments are given in Table 1.

Compared to a baseline without any explicitly modeled pose information, inclusion of either views or pose significantly increases the accuracy of the resulting feature embedding. This observation holds across both datasets, as well as both network architectures. For the ResNet model, the view information results in a larger absolute improvement of about 6-7% in mAP on both datasets while the pose information leads only to an improvement of about 2-3% in mAP. Results for the Inception-v4 model are less consistent. Both types of information still achieve large improvements but on the Market dataset the absolute improvement for both types lies around 10% in mAP while on Duke the 11% mAP improvement through pose information clearly outperforms the 4% gained by including view information.

Finally, a combination of the two types of information leads to a further consistent increase in mAP compared to the best result of either individual pose information. For instance, on the base ResNet-50 model, the combination achieves a further improvement in mAP of 2.1% and 5.3% on Market and Duke, respectively. Similarly, on the base Inception-v4 model the combination further improves the mAP by 3% on Market and 2.2% on Duke. This clearly indicates that our methods of including different degrees of pose information complement each other.

View-predictor performance: The performance of the trained ResNet-50 view predictor on the annotated test set of RAP dataset is 82.2%, 86.9% and 81.9% on front, back, and side views, respectively. In order to illustrate its performance on our target re-id dataset we display mean images in Figure 3. These are obtained by averaging all images, on the test set of the target dataset, which are classified as front, left, or side. This visualization gives an impression of the view prediction accuracy on the target re-id data in the absence of annotated view labels. In the frontal mean image

Figure 3. Mean images of Market-1501 (left) and Duke (right) test sets using predictions of the PSE model's view predictor. The images show front, back and side view from left to right.

a skin-colored face region is clearly discernible, indicating that the majority of images were in fact frontal ones. Similarly, the back mean image correctly shows the backside of a person. The side view is more ambiguous, aside from the possible view predictor errors, mainly because we group left and right side into one combined class.

5.2. Study of Re-Ranking

In Table 2 we compare several configurations of our proposed ECN re-ranking with other popular re-ranking methods across the Market, CUHK03 (detected) [19] and MARS datasets. Note that the CUHK03 includes both the labeled and detected (using a person detector) person bounding boxes. We chose the CUHK03 (detected) as it is more challenging. We evaluate CHUK03 under the new fixed train/test protocol as used in [48] [39]. To compare with the published results of several re-ranking methods on these datasets, we use the same baseline features, 2,048-dim ID-discriminative embedding provided by [48]. We compare with the previous re-ranking techniques for object retrieval and person re-id including contextual dissimilarity measure (CDM) [14], spatially constrained (k-NN) re-ranking [30], Average query expansion (AQE) [7] and the current state-of-the-art Sparse Contextual Activation (SCA) [1] , k-reciprocal encoding (k-reciprocal) [48] and its direct multiplicative application Divide and Fuse (DaF) [39]. As shown our ECN re-ranking achieves a consistent improvement in performance across all three datasets on both mAP and rank-1 metrics.

We provide the performance of the different components

Re-Ranking		Market-1501 IDE-R		CUHK03 IDE-R		MARS IDE-C	
		mAP	R-1	mAP	R-1	mAP	R-1
None		55.0	78.9	19.7	21.3	41.2	61.7
AQE [7]		-	-	-	-	47.0	61.8
CDM [14]		56.7	79.8	20.6	22.9	44.2	62.1
K-NN [30]		60.3	79.5	22.9	24.3	-	-
SCA [1]		68.9	79.8	26.6	24.7	-	-
k-reciprocal [48]		70.4	81.4	27.3	24.9	51.5	62.8
DaF [39]		**72.4**	82.3	30.0	26.4	-	-
Our	Rank dist (Eq. 4)	66.1	80.3	25.0	25.3	48.7	62.2
Our	ECN (orig-dist)	66.7	81.7	27.5	25.9	50.1	64.7
Our	**ECN (rank-dist)**	71.1	**82.3**	**30.2**	**27.3**	**53.2**	**64.6**

Table 2. Comparison of the proposed ECN re-ranking method with state-of-the-art on three datasets, Market-1501, CHUK03 (detected) and MARS. Baseline features: 2,048-dim ID-discriminative Embedding fine tuned on Resnet (IDE-R) and CaffeNet (IDE-C) [48].

of our ECN framework. As shown in Table 2, only using the rank-list distance of Equation 4 (rank-dist) still provides meaningful performance gains. Within the ECN framework just using the direct euclidean distances in Equation 3 'ECN (orig-dist)' results in similar high performance gains in the rank-1 scores, in fact better than the state-of-the-art k-reciprocal [48] method that uses the reciprocal list comparisons with local query expansion and fusion of rank and euclidean distances. As this does not involve computing any rank list based comparison, this result is an additional very attractive outcome of our proposal. Finally our ECN re-ranking using the simple rank-list comparison of Equation 4 as distance in the ECN Equation 3 provides the best results and improves the mAP further.

Parameters impact: In all of our evaluations presented in Table 2 as well as in Table 3, the ECN parameters are set to t=3 and q=8. Given the very different number of images in query and test sets of the used datasets, the results show the stability of these parameters. We studied the impact of changing these on Market and Duke datasets and found that it is subtle in the range for $t \in [2, 4]$ and $q \in [4, 10]$, the performance drops between $\sim$0.2-0.8% on different combinations within this range. Similarly the impact of parameter K in Equation 4 works well within $K \in [10, 30]$, with better performance when $K > 20$ on all three large datasets Market, MARS and Duke. The jitter in accuracies with changing K in this range stays within $\sim\pm 2\%$.

Since CUHK03 is a relatively small dataset, both DaF [39] and k-reciprocal [48] report results on CUHK03 by using different parameters values for their methods than used for the other datasets. While we used the same ECN parameters of t=3, and q=8 on CUHK03, we obtained higher performance with the parameter K=10 instead of K=25 (as used on all other datasets) for the rank-list distance in Equation 4. The reported results in table 2 on CHUK03 dataset are with K=10, however with K=25, we still get better performance than the most state-of-the-art methods with mAP of 28.4% and rank-1 of 26.0%.

Complexity analysis: The computational complexity of ECN is $\mathcal{O}(N^2 logN)$ (same as other re-ranking methods) but it executes fewer computation steps by avoiding re-computing the neighbors' lists for each image pair. In its variant with ECN (orig-dist) it offers close improvements without having to re-compute the rank lists based distance (hence even fewer steps). For example, on the large Duke dataset (re-ranking on 19,889 total images), computation times averaged over five runs are 124.6s for the related work k-reciprocal [48] while 115.3s and 73.2s for our ECN (rank-dist) and ECN (orig-dist) respectively.

5.3. State-of-the-art

In Table 3 we compare the performance of our approach with the published state-of-the-art on the three popular datasets (Market, Duke, and MARS). In the top section of the table we compare approaches without any re-ranking to our pose-sensitive embedding. The embedding achieves top accuracy on both MARS and Duke datasets. On the Market dataset our embedding performs slightly worse than the DPFL [3] approach which employs two or more multi-scale embeddings. Across all three datasets the increase in mAP achieved by including pose information on the base ResNet ranges from 7.4% to 11.7%. In the bottom section of Table 3 we include the best published methods with re-ranking. In combination with our proposed re-ranking scheme we set a new state-of-the-art on all three datasets by large margins. On Market we increase mAP by 11.4%, on Duke by 19.2%, and on MARS by 4.5%.

Real World Considerations: In real-world applications re-id methods needs to be scalable (large gallery sizes) and are used in combination with automatic person detectors which can generate errors, such as mis-aligned detections or false positives. To investigate the scalability of our proposed PSE model, we evaluate on the Market+500k dataset to judge its robustness in real world deployment with very large galleries. The Market+500k dataset extends the Market dataset by including up to 500,000 distractor persons images. The relative change in mAP and rank-1 accuracy of our PSE model in comparison to other state-of-the-art approaches is depicted in Table 4. While our embedding outperforms the published state-of-the-art without any distractors, the drop in accuracy observed when adding distractors is also notably less steep than that of other approaches. At 500,000 distractors our PSE's mAP has dropped by 12.5% while related approaches dropped by more than 14%, similarly PSE drops in rank-1 accuracy by $\sim$7% while the related approaches drop by $\sim$10%. This shows the quality of our PSE model for this more realistic setting.

In order to test our PSE embedding under detector errors, we train and evaluate its performance on the PRW dataset [45]. Using the DPM detections provided with the dataset

Method			Market-1501		Duke		MARS	
			mAP	R-1	mAP	R-1	mAP	R-1
P2S[49]		CVPR17	44.3	70.7	-	-	-	-
Spindle[41]		CVPR17	-	76.9	-	-	-	-
Consistent Aware[21]		CVPR17	55.6	80.9	-	-	-	-
GAN[47]		ICCV17	56.2	78.1	47.1	67.7	-	-
Latent Parts [16]		CVPR17	57.5	80.3	-	-	56.1	71.8
ResNet+OIM [35]		CVPR17	-	82.1	-	68.1	-	-
ACRN[29]		CVPR17-W	62.6	83.6	52.0	72.6	-	-
SVD [33]		ICCV17	62.1	82.3	56.8	76.7	-	-
Part Aligned [42]		ICCV17	63.4	81.0	-	-	-	-
PDC [32]		ICCV17	63.4	84.1	-	-	-	-
JLML [20]		IJCAI17	65.5	85.1	-	-	-	-
DPFL [3]		ICCV17-W	**72.6**	**88.6**	60.6	79.2	-	-
Forest [50]		CVPR17	-	-	-	-	50.7	70.6
DGM+IDE [38]		ICCV17	-	-	-	-	46.8	65.2
Our	ResNet-50 Baseline		59.8	82.6	50.3	71.5	49.5	64.5
Our	**PSE**		69.0	87.7	**62.0**	**79.8**	**56.9**	**72.1**
Smoothed Manif. [2]		CVPR17	68.8	82.2	-	-	-	-
IDE (R)+XQDA+k-reciprocal [48]		CVPR17	61.9	75.1	-	-	68.5	73.9
IDE (R)+KISSME+k-reciprocal [48]		CVPR17	63.6	77.1	-	-	67.3	72.3
DaF [39]		BMVC17	72.4	82.3	-	-	-	-
Our	PSE+ k-reciprocal [48]		83.5	90.2	78.9	84.4	70.7	74.9
Our	PSE+ rank-dist (Eq. 4)		80.5	89.6	74.5	82.8	67.7	74.9
Our	PSE+ ECN (orig-dist)		80.5	**90.4**	75.7	84.5	68.6	75.5
Our	**PSE+ ECN (rank-dist)**		**84.0**	90.3	**79.8**	**85.2**	**71.8**	**76.7**

Table 3. Comparison of our approach with the published state-of-the-art. The top section of the table compares our embedding with state-of-the-art approaches that do not use re-ranking. The lower section compares our combination of embedding and re-ranking to other state-of-the-art methods that use re-ranking.

Method		mAP by #Distractors				R-1 by #Distractors			
		0	100k	200k	500k	0	100k	200k	500k
I+V†[46]		59.9	52.3	49.1	45.2	79.5	73.8	71.5	68.3
APR†* [22]		62.8	56.5	53.6	49.8	84.0	79.9	78.2	75.4
TriNet†§ [11]		**69.1**	61.9	58.7	53.6	84.9	79.7	77.9	74.7
Our	ResNet-50 Baseline	59.8	54.6	51.8	47.5	82.6	77.7	75.7	73.2
Our	Views Only	66.9	61.5	58.9	54.8	**88.2**	**84.4**	**83.2**	**81.2**
Our	Pose Only	63.0	57.7	54.9	50.6	83.6	80.0	77.9	75.1
Our	PSE	69.0	**63.4**	**60.8**	**56.5**	87.7	84.1	82.6	80.8

Table 4. Results of the PSE embedding on the Market-1501+500k distractors dataset († = unpublished works, $*$ = additional attribute ground truth, § = x10 test-time augmentation).

Detector	Method	#detections=3			#detections=5			#detections=10		
		mAP	R-1	R-20	mAP	R-1	R-20	mAP	R-1	R-20
DPM	IDE_{det} [45]	17.2	45.9	77.9	18.8	45.9	77.4	19.2	45.7	76.0
DPM-Alex	IDE_{det} [45]	20.2	48.2	78.1	20.3	47.4	77.1	19.9	47.2	76.4
DPM-Alex	IDE_{det}+CWS [45]	20.0	48.2	78.8	20.5	48.3	78.8	20.5	48.3	78.8
IAN (ResNet-101) [34]		23.0	61.9	-	-	-	-	-	-	-
DPM	Baseline	25.4	59.0	83.9	27.5	59.1	83.9	28.3	58.1	83.3
DPM	View only	28.5	63.4	87.3	30.8	63.1	86.8	31.4	62.0	86.1
DPM	Pose only	26.2	59.1	84.6	28.4	58.6	84.4	29.1	58.1	83.4
DPM	PSE	**29.3**	**65.1**	**88.3**	**31.7**	**65.0**	**88.2**	**32.4**	**64.5**	**87.5**

Table 5. Results of PSE on PRW dataset (robustness against false detections): Considering 3, 5 and 10 detections per image.

we observe similar trends as on Market or Duke. Both types of pose information improve notably over the baseline and achieve a further increase in accuracy when combined in the PSE embedding. The performance is stable when considering more detections per image (hence increasing false positives) as shown in Table 5. The PSE embedding achieves state-of-the-art accuracy, outperforming related approaches by at least 6.3% in mAP (when an average of 3 detections per image are considered). The results confirm the intuition that pose information is a helpful cue in identifying and handling mis-aligned and false-positive person detections.

6. Conclusion

We have presented two related but independent contributions for person re-id and retrieval applications. We showed that both the fine and coarse body pose cues are important for re-id and proposed a new pose-sensitive CNN embedding which incorporates these. The PSE model currently relies on an external pose predictor, it would be useful to fully integrate this into the model. The re-ranking method is unsupervised and can be used for general image and video retrieval applications. Both our person re-id model and re-ranking method set new state-of-the-art on a number of challenging datasets independently and in concert with each other.

Acknowledgement: The research described in this work is funded in part by the BMBF grant No. 13N14029.

References

[1] S. Bai and X. Bai. Sparse contextual activation for efficient visual re-ranking. *IEEE Transactions on Image Processing*, 25(3):1056–1069, 2016.

[2] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCV workshop on cross domain human identification*, 2017.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 2, page 6, 2011.

[6] Y.-J. Cho and K.-J. Yoon. Improving person re-identification via pose-aware multi-shot matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2016.

[7] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[9] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *ICCV IEEE International Conference on Computer Vision*, pages 1305–1313, 2015.

[10] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[11] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[12] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[13] R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on computers*, 100(11):1025–1034, 1973.

[14] H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Computer Vision and Pattern Recognition,CVPR*, pages 1–8. IEEE, 2007.

[15] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen. Person re-identification with content and context re-ranking. *Multimedia Tools and Applications*, 74(17):6989–7014, 2015.

[16] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

[18] W. Li, Y. Wu, M. Mukunoki, and M. Minoh. Common-near-neighbor analysis for person re-identification. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1621–1624. IEEE, 2012.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[20] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, 2017.

[21] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[23] A. J. Ma and P. Li. Query based adaptive re-ranking for person re-identification. In *Asian Conference on Computer Vision*, pages 397–412. Springer, 2014.

[24] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 777–784. IEEE, 2011.

[25] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, and H. Qi. Person re-identification using visual attention. *arXiv preprint arXiv:1707.07336*, 2017.

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[27] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017.

[28] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie. Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison. In *ICCV, IEEE International Conference on*, pages 2494–2501. IEEE, 2011.

[29] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1435–1443. IEEE, 2017.

[30] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3013–3020. IEEE, 2012.

[31] Springer. *MARS: A Video Benchmark for Large-Scale Person Re-identification*, 2016.

[32] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision ICCV*, pages 3960–3969, 2017.

[33] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[34] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng. IAN: the individual aggregation network for person search. *CoRR*, abs/1705.05552, 2017.

[35] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. CVPR*, 2017.

[36] M. Ye, J. Chen, Q. Leng, C. Liang, Z. Wang, and K. Sun. Coupled-view based ranking optimization for person re-identification. In *International Conference on Multimedia Modeling*, pages 105–117. Springer, 2015.

[37] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[38] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[39] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. In *BMVC*, 2017.

[40] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *arXiv preprint arXiv:1706.00384*, 2017.

[41] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. 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.

[42] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. *ICCV*, 2017.

[43] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

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

[45] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. *arXiv preprint arXiv:1604.02531*, 2016.

[46] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*, 2016.

[47] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[48] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. pages 1318–1327, 2017.

[49] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[50] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.