Deep Spatial Pyramid for Person Re-identification

Sławomir Bąk Peter Carr
Disney Research
Pittsburgh, PA, USA, 15213
{slawomir.bak,peter.carr}@disneyresearch.com

Abstract

Re-identification refers to the task of finding the same subject across a network of surveillance cameras. This task must deal with appearance changes caused by variations in illumination, a person's pose, camera viewing angle and background clutter. State-of-the-art approaches usually focus either on feature modeling – designing image descriptors that are robust to changes in imaging conditions, or dissimilarity functions – learning effective metrics to compare images from different cameras. Typically, with novel deep architectures both approaches can be merged into a single end-to-end training, but to become effective, this requires annotating thousands of subjects in each camera pair. Unlike standard CNN-based approaches, we introduce a spatial pyramid-like structure to the image and learn CNNs for image sub-regions at different scales. When training a CNN using only image sub-regions, we force the model to recognize not only the person's identity but also the spatial location of the sub-region. This results in highly effective feature representations, which when combined with Mahalanobis-like metric learning significantly outperform state-of-the-art approaches.

1. Introduction

Person re-identification (re-id) refers to the task of finding the same subject across a network of non-overlapping surveillance cameras. The visual appearance model is fundamental for solving the re-id problem but its sensitivity to imaging conditions, *e.g*. variations in illumination, changing camera viewpoints, different person's poses; make the problem very challenging.

Recent studies have shown [4, 6, 21, 22, 30, 33] that *metric learning* approaches often achieve the best performance in re-identification. Despite the huge progress in *deep learning*, state-of-the-art re-identification techniques are still usually based on handcrafted image features combined with Mahalanobis-like metric learning [4, 21, 22]. Insufficient data in re-identification datasets (the small number of subjects and images) makes training Convolutional Neural Networks (CNNs) from scratch very difficult for person re-identification. Training CNNs on larger datasets (*e.g*. ImageNet) and fine-tuning to re-id usually does not provide effective representation due to a domain mismatch (*e.g*., the difference in the image content and quality).

Recently, Xiao *et al*. [30] have shown that CNN models can also effectively be applied to person re-identification by first merging multiple re-identification datasets into a single dataset and then training jointly the CNN through challenging multi-class identification task *i.e*., classifying a training image into one of T identities (where T is the total number of subjects across all training datasets). To increase the performance for a given camera pair, a domain-guided dropout strategy was proposed to fine-tune the model to specific imagining conditions.

In this paper, we also employ the strategy for merging re-identification datasets into a single training dataset. However, instead of training CNNs for classifying person identities using whole images, we propose to train CNNs using only rectangular sub-regions of images (*i.e*., patches) and force the neural network to recognize both the person's identity and the patch location. This makes the task more difficult and more iterations are usually required for convergence of the neural network, but in actual fact yields better deep feature representations for patches. Moreover, feeding CNNs with patches increases the number of classes predicted during training, thus improving the generalization capabilities of the learned features [26].

Inspired by spatial-pyramid-like matching [17], we train different CNN models for different sizes of patches. Instead of applying spatial-pyramid to the pooling layer [13, 29], we apply it directly to the input image (the first layer). To combine image regions at different spatial levels of the pyramid, rather than using spatial pyramid kernels we let Mahalanobis-like metric learning discover the optimal weighting strategy. Our contributions are the following:

978-1-5386-2939-0/17/$31.00

- We propose to introduce a spatial pyramid-like structure to the image and to learn deep feature representations for patches at different scales. We denote it as the *deep spatial pyramid* features.

- Deep features extracted at different scales are further combined by learning metrics locally for patches. As the employed metric learning is based on a likelihood-ratio test, we pool the patch dissimilarities by average to form the final dissimilarity measure.

- We conduct extensive experiments on two benchmark datasets – VIPeR [11] and CUHK01 [18]. The results illustrate that by combining deep spatial pyramid features and Mahalanobis-like metric learning, we achieve new state-of-the-art performance, outperforming existing approaches by large margins.

2. Related work

Re-identification techniques can be divided into two groups: *feature modeling* [3, 7, 9] – designs descriptors (usually handcrafted) which are robust to changes in imaging conditions, and *metric learning* [1, 8, 16, 19, 20, 32, 37] – searches for effective distance functions to compare features from different cameras.

The latter are usually more effective as they can adapt to specific lighting conditions across cameras. Many different machine learning algorithms have been considered for learning a robust similarity function. Gray *et al*. employed Adaboost for feature selection and weighting [12], Prosser *et al*. defined the person re-identification as a ranking problem and used an ensemble of RankSVMs [25]. Recently features learned from deep convolution neural networks have also been investigated [1, 6, 19, 28, 30, 36].

However, despite advances in *deep learning*, state-of-the-art re-identification performance still belongs to handcrafted image features combined with Mahalanobis-like metric learning [4, 22, 33]. It might be due to the fact that insufficient data in re-identification datasets (the small number of subjects and images) makes training Convolutional Neural Networks (CNNs) from scratch very difficult for person re-identification.

The most common choice of Mahalanobis-like metric learning remains KISS metric learning [16]. The KISS metric uses a statistical inference based on a likelihood-ratio test of two Gaussian distributions modeling positive and negative pairwise differences between features. Owing to its effectiveness and efficiency, the KISS metric is a popular baseline that has been extended to linear [21, 24] and non-linear [23, 31] subspace embeddings. Most of these approaches learn a Mahalanobis distance function for handcrafted feature vectors extracted from full bounding box images.

In this paper, we propose to learn deep features for patches with Convolutional Neural Networks (CNNs) through challenging multi-class patch identification task. The deep patch features are trained at different scales forming the spatial pyramid-like structure. The idea of matching features at different scales (pyramid matching) was originally proposed in [10] and then it was further extended to 2D images in [17]. First, a sequence of grids at resolutions $0, \ldots, l$ is constructed and then the number of matches that occur at each level of resolution are combined using weighted sums. The matches found at finer resolutions are usually weighted more than matches found at coarser resolutions. The weights are usually handcrafted, where our approach lets metric learning to discover the optimal weighting strategy. Recent studies have already introduced the spatial-pyramid idea to the pooling layer [13, 29], but our approach significantly differs from these techniques. We incorporate the pyramid at the input layer, modifying the training strategy. This results in highly effective deep feature representations that together with metric learning yields excellent re-identification accuracy.

3. Method

The proposed pipeline for learning similarity measures between images consists of three stages: (1) deep spatial pyramid feature learning (Sec. 3.1), (2) deep feature extraction (Sec. 3.2) and (3) metric learning (Sec. 3.3). We refer to our approach as *Deep Spatial Pyramid* (**DSP**).

3.1. Deep Spatial Pyramid Feature Learning

Similar to [30], we use multiple datasets to train CNNs and we adopt the CNN model from [30] as the main component of our framework (3 convolutional layers, 3 BN-Inception layers [15, 27] and two fully connected layers). This model learns a set of high-level feature representation through a challenging multi-class identification task, *i.e*., classifying a training image into one of T identities. As the generalization capabilities of the learned features increase with the number of classes predicted during training [26], we need T to be relatively large (*e.g*. several thousand).

Instead of training a single CNN model for determining the person's identity using whole images, we propose to train three CNN models (see Fig. 1). This includes CNNs that are trained only using sub-regions of images (*e.g*. patches) to determine both the person's identity and the patch location. For training CNNs at level 1 and 2, we propose the following strategy. First, each image is divided into a set of *non-overlapping* patches of size $(\frac{\mathrm{h}}{4} \times \mathrm{w})$ for level 1 and $(\frac{\mathrm{h}}{4} \times \frac{\mathrm{w}}{2})$ for level 2, where h and w correspond to image height and width, respectively. Level 1 consists of horizontal stripes to learn features that are viewpoint invariant, and level 2 is introduced to provide finer details

within the stripes. Each patch (although comes from the same image but from the different location) gets assigned a new identity. As a result, the CNN models are forced to determine not only the person's identity but also the patch location. Given a training dataset with images of T identities, the task becomes to classify patches into $C_0 = T$ identities at level 0, $C_1 = 4T$ identities at level 1 and $C_2 = 8T$ identities at level 2. When each CNN is trained to classify the large number of identities and configured to keep the dimension of the last hidden layer relatively low (*i.e.*, for simplicity we set the number of dimensions for fc7 to 256 for all three CNNs), it forms compact and highly robust feature representations for re-identification.

3.2. Deep Feature Extraction

Let us assume that the trained CNNs can be used as deep feature extractors, *i.e.*, fc7-0 extracts deep features from the whole images, fc7-1 extracts deep features from stripes of size $(\frac{h}{4} \times w)$ and fc7-2 extracts deep features from patches of size $(\frac{h}{4} \times \frac{w}{2})$ (see Fig. 1).

We divide each image into a dense set of *overlapping* rectangular patches to extract fc7 features (notice the difference with the training phase where we use *non-overlapping* patches). At level 1 we use a horizontal stride $\frac{h}{8}$ which in our configuration results in 13 patches ($l_1 = 13$) and at level 2 we use a stride $(\frac{h}{8} \times \frac{w}{4})$ which results in 39 patches ($l_2 = 39$). We found that using features from overlapping patches significantly improves the recognition accuracy (see Sec. 4.3).

Image i is then represented by a set of deep features $\mathcal{X}_i = \{\mathbf{x}_i^0, \mathbf{x}_i^1, \ldots, \mathbf{x}_i^{l_1}, \mathbf{x}_i^{l_1+1}, \ldots, \mathbf{x}_i^{l_1+l_2}\}$, where $\mathbf{x}_i^0$ corresponds to feature fc7-0 (level 0) extracted from the whole image; each feature from $\mathbf{x}_i^1$ to $\mathbf{x}_i^{l_1}$ is fc7-1 computed by feeding CNN at level 1 with overlapping stripes of size $(\frac{h}{4} \times w)$; and each feature from $\mathbf{x}_i^{l_1+1}$ to $\mathbf{x}_i^{l_1+l_2}$ is fc7-2 computed using CNN at level 2 from overlapping patches of size $(\frac{h}{4} \times \frac{w}{2})$; superscript l is an iterator that determines the feature level and the location.

3.3. Metric Learning

One could concatenate all features from $\mathcal{X}_i$ into a single high-dimensional feature vector and then follow a standard two stage processing for metric learning [16], *i.e.*, first apply dimensionality reduction (*e.g.* PCA) and then perform metric learning on the reduced subspace of differences between feature vectors. As our experiments illustrate (see Sec. 4.5), instead of learning a metric for the global concatenated feature, it is better to learn metrics on a level of patches (*i.e.*, for $\mathbf{x}_i^l$) and then integrate all metrics by computing the total dissimilarity (Eq. 3). Similar phenomenon has been found in [2].

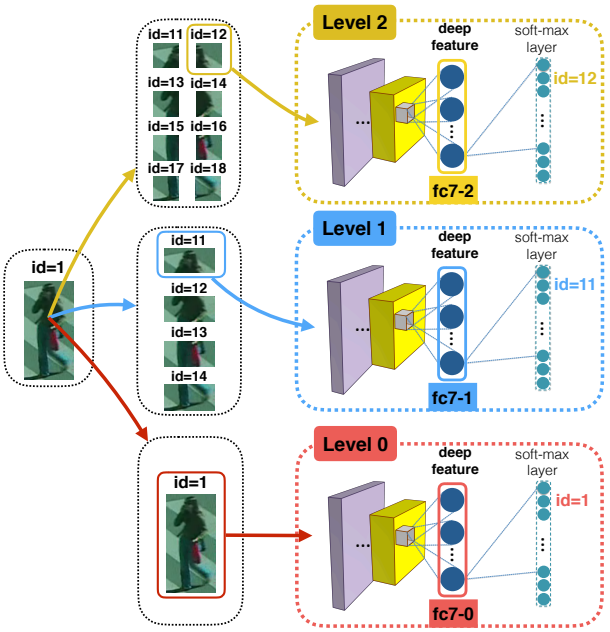


Figure 1: **Deep Spatial Pyramid feature learning**. We train 3 CNNs through multi-class identification task. At levels 1 & 2 each image is divided into a set of non-overlapping rectangular sub-regions. The identity of each sub-region is extended by its location. As a result, the CNNs are forced to recognize not only the person identity but also the sub-region location. Learned CNN features fc7-0, fc7-1 and fc7-2 are further integrated with Metric Learning to produce the final dissimilarity function.

3.3.1 LSSL Metric Learning

Yang *et al.* [32] introduced Large Scale Similarity Learning (LSSL) that can be seen as an extension of well known KISS metric learning [16]. Following [32], we define the distance between two features $\mathbf{x}_i^l$ and $\mathbf{x}_j^l$ extracted from two images i and j at level and location l as

$$\Phi^{(l)}(\mathbf{x}_i^l, \mathbf{x}_j^l) = (\mathbf{x}_i^l - \mathbf{x}_j^l)^T \mathbf{M}_d^l (\mathbf{x}_i^l - \mathbf{x}_j^l) - \quad (1)$$

$$\lambda (\mathbf{x}_i^l + \mathbf{x}_j^l)^T \mathbf{M}_c^l (\mathbf{x}_i^l + \mathbf{x}_j^l). \quad (2)$$

Matrices $\mathbf{M}_d^l$ and $\mathbf{M}_c^l$ can be efficiently learned by assuming Gaussian structure for both the pairwise difference space ($\mathbf{d}_{ij}^l = \mathbf{x}_i^l - \mathbf{x}_j^l$) to learn $\mathbf{M}_d^l$ and for the pairwise commonness space ($\mathbf{c}_{ij}^l = \mathbf{x}_i^l + \mathbf{x}_j^l$) to learn $\mathbf{M}_c^l$.

Typically [16], the space of pairwise differences $\mathbf{d}_{ij}^l = \mathbf{x}_i^l - \mathbf{x}_j^l$ is divided into positive pairwise set $\mathbf{d}_{ij}^{+(l)}$ when i and j contain the same person and negative pairwise set $\mathbf{d}_{ij}^{-(l)}$ otherwise. Learning metric $\mathbf{M}_d^l$ then involves computing two covariance matrices: $\Sigma_d^{+(l)}$ for positive pairwise differences ($\Sigma_d^{+(l)} = (\mathbf{d}_{ij}^{+(l)})(\mathbf{d}_{ij}^{+(l)})^T$) and $\Sigma_d^{-(l)}$ for nega-

tive pairwise differences ($\Sigma_d^{-(l)} = (\mathbf{d}_{ij}^{-(l)})(\mathbf{d}_{ij}^{-(l)})^T$). From the log-likelihood ratio, the Mahalanobis metric becomes $\mathbf{M}_d^l = (\Sigma_d^{+(l)})^{-1} - (\Sigma_d^{-(l)})^{-1}$. Analogously, $\mathbf{M}_c^l$ can be learned by replacing the pairwise difference space $\mathbf{d}_{ij}^l = \mathbf{x}_i^l - \mathbf{x}_j^l$ with the pairwise commonness space $\mathbf{c}_{ij}^l = \mathbf{x}_i^l + \mathbf{x}_j^l$ and following the same procedure. Compared to the standard KISS metric learning [16], Yang *et al*. [32] introduced the commonness term (Eq. 2), which makes the dissimilarity measure more effective. Additionally, [32] shows that based on a pair-constrained Gaussian assumption, covariance for pairs containing different people ($\Sigma_d^{-(l)}$ and $\Sigma_c^{-(l)}$) can be directly deduced from image pairs containing the same person (for details see [32]). Parameter λ is used to balance between difference and commonness of feature vectors. Similarly to [32], we set $\lambda = 1.5$ in all experiments.

3.4. Final Dissimilarity

Typically, spatial pyramid matching uses the weighted sum to combine the dissimilarities that occur at different levels. Matches found at finer levels are usually weighted more than matches found at coarser levels. In our approach, as relevant features are emphasized and irrelevant ones are ignored during the metric learning procedure, we found that any further weighting of $\Phi^{(l)}$'s does not have significant impact on the performance. As a result, our total dissimilarity measure between two images i and j becomes

$$\mathcal{D}(i,j) = \frac{1}{1+l_1+l_2} \sum_{l=0}^{l_1+l_2} \Phi^{(l)}(\mathbf{x}_i^l, \mathbf{x}_j^l). \qquad (3)$$

4. Experiments

We carry out experiments on two challenging datasets: **VIPeR** [11] and **CUHK01** [18]. To learn deep spatial pyramid features we additionally use **CUHK03** [19] and **PRID2011** [14] datasets. We report re-identification performance employing the CMC curve [11] and its rank-1 accuracy. The CMC curve provides the probability of finding the correct match in the top r ranks.

4.1. Datasets and Evaluation Protocols

VIPeR [11] is one of the most popular person re-identification datasets. It contains 632 image pairs of pedestrians captured by two outdoor cameras. VIPeR images contain large variations in lighting conditions, background, viewpoint, and image quality. We follow the common evaluation protocol for this database: randomly dividing 632 image pairs into 316 image pairs for training and 316 image pairs for testing. We repeat this procedure 10 times and compute the average CMC curves for obtaining reliable statistics.

CUHK01 [18] contains 971 persons captured with two cameras. The first camera captures the side view of pedestrians and the second camera captures the frontal view or the back view. We follow the common evaluation setting: the persons are split into 485 for training and 486 for testing. We repeat this procedure 10 times for computing averaged CMC curves. As 2 images for each person per camera are provided, we evaluate both the single-shot and the multi-shot setting.

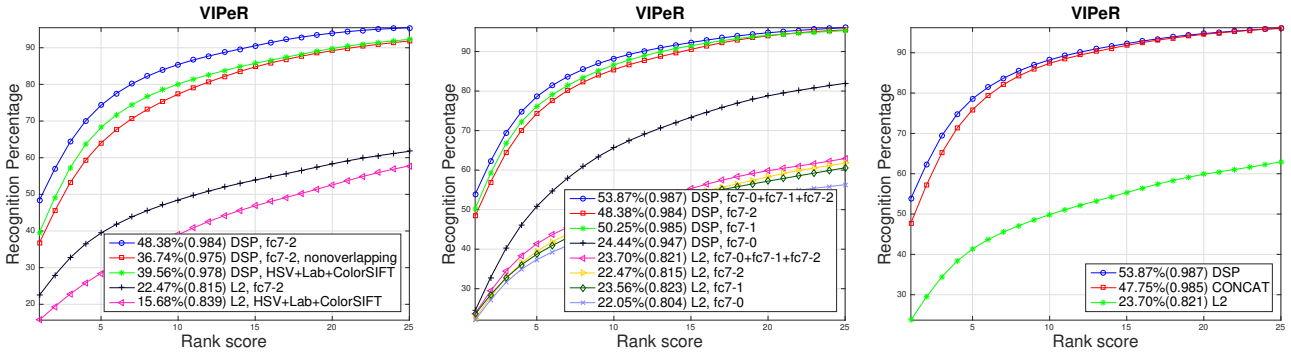
CUHK03 [19] and **PRID2011** [14] datasets are used for learning our deep spatial pyramid features. **CUHK03** is one of the largest published person re-identification datasets. It contains 1467 identities appearing in two camera views, so it fits very well for learning the CNN model [30]. **PRID2011** contains 200 individuals appearing in two cameras and additionally it contains 185 identities that appear in the first camera but do not reappear in the second one, and 549 identities that appear only in the second camera, in total 934 identities. Merging both datasets, we obtain $T = 1467 + 934 = 2401$ identities. When training at level 0, the CNN is learned to classify $T_0 = 2401$ identities; at level 1 we force the CNN to classify $T_1 = 4 \times 2401 = 9604$ identities and when training at level 2 the CNN has to distinguish $T_2 = 8 \times 2401 = 19208$ identities. The dimensionality of the last hidden layer in each CNN is kept to be low – 256. The architecture and training parameters are kept the same as in [30]. Unlike [30], we do not perform any fine-tuning on test datasets. The trained CNNs are used as feature extractors on unseen datasets and then metric learning is performed.

4.2. Image Settings

We fixed the evaluation settings across both datasets. All images are scaled to be 160×64 pixels. At level 1 the patch size is 40×64 pixels with 20 pixels horizontal stride. This results in $l_1 = 13$ fc7-1 features. At level 2 the patch size is 40×32 pixels with 20×16 stride. This results in $l_2 = 39$ fc7-2 features. Before applying metric learning the original dimensionality of fc7 features (256) is reduced by PCA to 68 dimensions (found by cross-validation).

4.3. Image Representation

A common practice in person re-identification is to use handcrafted features (such as combination of color histograms and texture descriptors) extracted from a person image in a dense grid fashion [2, 4, 21]. We compare our deep patch representation with a combination of color histograms – Lab, HSV (each 30 bins per channel) and color SIFT descriptor (SIFT computed for each Lab channel) [2]. The dimensionality of concatenated HSV, Lab and color SIFT is 564 ($30 \times 3 + 30 \times 3 + 128 \times 3 = 564$). We reduced this dimensionality to 35 (found by cross-validation). In accordance with layout proposed for fc7-2, we extract the



(a) (b) (c)

Figure 2: Performance comparison on **VIPeR** dataset. Rank-1 identification rates as well as *nAUC* values provided in brackets are shown in the legend next to the method name: (a) comparison of fc7 features *vs*. handcrafted features; (b) comparison of fc7 features on different levels of the spatial pyramid; (c) performance of concatenated fc7 features (CONCAT).

handcrafted descriptor from a dense grid with overlapping rectangular patches. From Fig. 2(a) it is apparent that deep patch features extracted from fc7-2 are more discriminative, significantly outperforming the handcrafted representation before (ℓ_2-norm) and after applying metric learning.

Additionally, we evaluated the performance of fc7-2 features extracted from non-overlapping patches. The results indicate that using overlapping deep patch features yields significantly better performance compared with non-overlapping patches (compare the blue curve with the red one in Fig. 2(a).

4.4. Deep Spatial Pyramid

In this section we evaluate the performance of fc7 features extracted at different levels of the pyramid. A comparison in Fig. 2(b) illustrates how the performance changes after applying metric learning. We can notice that metric learning is much more beneficial when applied to fc7 features extracted from patches (*e.g*. fc7-1 and fc7-2) than to fc7-0 features extracted from whole images (compare the black curve with the green and the red one). The best performing layer is fc7-1 and the best re-identification accuracy is achieved by combining all levels of the spatial pyramid.

4.5. Metrics for Patches

A common approach is to concatenate all features into a single high-dimensional feature vector, apply PCA and learn a single metric for comparing images. In [2] it has been shown that this might not the optimal approach and it is better to learn metrics on the level of patches. We concatenate fc7 features extracted from all levels into a single feature vector and reduce the dimensionality to 110 components, maximizing the rank-1 accuracy (CONCAT). Fig. 2(c) indicates that indeed operating on patches yields better re-identification accuracy.

METHOD	VIPeR	CUHK01, M=1	CUHK01, M=2
DSP	**53.9**	**72.0**	**79.2**
Dropout [30]	38.6	-	66.6
Null Space [33]	42.2	-	64.9
Null Space (fusion) [33]	51.1	-	**69.0**
Triplet Loss [6]	47.8	53.7	-
Gaussian+XQDA [22]	49.7	**57.8**	67.3
Specific SVM [34]	42.6	-	65.9
SCSP [4]	**53.5**	-	-
DPML [2]	41.4	35.8	37.5
DeepNN [1]	34.8	47.5	-
LOMO+XQDA [21]	40.0	-	63.2
Mirror [5]	42.9	40.4	-
Ensembles [23]	45.9	53.4	-
MidLevel [35]	29.1	-	34.3

Table 1: Performance comparison on **VIPeR** and **CUHK01**; CMC rank-1 accuracies are reported. The best scores are shown in red. The second best scores are highlighted in blue. Our approach significantly outperforms the best state of the art approaches.

4.6. Comparison with Other Methods

Table 1 illustrates the performance comparison of our **DSP** method with state-of-the-art approaches across 2 datasets. For CUHK01 we report accuracies for both the single-shot setting (M=1) and the multi-shot setting (M=2). Our method outperforms all state-of-the-art techniques on all datasets. The maximum improvement is achieved on the CUHK01 dataset. We improve the state-of-the-art rank-1 accuracy for M=1 (57.8) by 14.2% (72.0%) and for M=2 (69.0) by 10.2% (79.2%). Compared with Dropout [30] (from which we adopted the CNN architecture), our method achieves a gain of 15.3% for VIPeR dataset and a gain of 12.6% for CUHK01 dataset.

5. Conclusion

Standard Mahalanobis metric learning approaches rely on handcrafted image features and do not benefit from new deep learning architectures. Alternatively, learning CNNs from scratch is very difficult due to insufficient data in re-identification datasets. In this work we presented the novel deep spatial pyramid framework that learns very effective deep features for patches at different scales. Integrating these features with metric learning leads to new state-of-the-art performance on re-identification datasets.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] S. Bak and P. Carr. Person re-identification using deformable patch metric learning. In *WACV*, 2016.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010.

[4] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016.

[5] Y.-C. Chen, W.-S. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *IJCAI*, 2015.

[6] 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 *CVPR*, 2016.

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

[8] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2010.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[10] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[11] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *PETS*, 2007.

[12] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

[14] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[16] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[20] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[21] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[22] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[23] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

[24] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[25] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[26] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[28] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[29] X. S. Wei, B. B. Gao, and J. Wu. Deep spatial pyramid ensemble for cultural event recognition. In *ICCVW*, 2015.

[30] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[31] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[32] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *AAAI*, 2016.

[33] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[34] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, 2016.

[35] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.

[36] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[37] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.