Aggregating Deep Pyramidal Representations for Person Re-Identification

Niki Martinel, Gian Luca Foresti, Christian Micheloni
Machine Learning and Perception Lab / Artificial Vision and Real-Time Systems Lab
University of Udine, Udine, Italy
{niki.martinel, gianluca.foresti, christian.micheloni}@uniud.it

Abstract

Learning discriminative, view-invariant and multi-scale representations of person appearance with different semantic levels is of paramount importance for person Re-Identification (Re-ID). A surge of effort has been spent by the community to learn deep Re-ID models capturing a holistic single semantic level feature representation. To improve the achieved results, additional visual attributes and body part-driven models have been considered. However, these require extensive human annotation labor or demand additional computational efforts. We argue that a pyramid-inspired method capturing multi-scale information may overcome such requirements. Precisely, multi-scale stripes that represent visual information of a person can be used by a novel architecture factorizing them into latent discriminative factors at multiple semantic levels. A multi-task loss is combined with a curriculum learning strategy to learn a discriminative and invariant person representation which is exploited for triplet-similarity learning. Results on three benchmark Re-ID datasets demonstrate that better performance than existing methods are achieved (e.g., more than 90% *accuracy on the Duke-MTMC dataset).*

1. Introduction

Person re-identification (Re-ID) is usually the task of associating a person acquired by disjoint cameras at different time instants. The problem has recently gained increasing attention [52] due to its open challenges like changes in viewing angle, background clutter, and occlusions.

To address these issues, existing approaches seek either the best feature representations (e.g., [58, 32]), learn optimal matching metrics (e.g., [38, 29, 40]), or investigate deep learning-based methods combining the two aforementioned solutions (e.g., [37, 10, 15, 45]). While feature representation and metric learning-based methods have obtained reasonable performance on benchmark datasets (e.g., [9, 12]), the deep learning-based solutions currently dominate this community, with convincing superiority against competitors.

Deeply-learned representations carry highly discriminative information, especially when this is obtained from body parts. This is substantiated by the results achieved with part-informed deep features that continuously raise the state-of-the-art Re-ID benchmarks (e.g., [49, 46, 27]). Approaches like [47, 54, 33] exploit external cues, e.g., leverage on human pose estimation solutions (e.g., [56, 22, 5, 14]). Other methods directly process the input without considering semantic part cues. Existing solutions apply keypoint-based body parts division strategies [55], learn an attention mechanism [34], or consider a uniform partitioning scheme [49]. These do not require part labeling and yet perform on par with part cues-based methods.

Motivation. It is a matter of fact that semantic body partitions offer stable cues for a good alignment and subsequent feature extraction. However, obtaining such partitions requires optimal body part detections which inevitably introduce additional computational efforts. While this issue can be mitigated by incorporating attention-like mechanisms within a single architecture (e.g., [34]), existing approaches do not have considered that body partitions may have different importance when analyzed at various scales. Indeed, by nature, images contain objects of many sizes as well as their representing features. Due to this, single scale analysis may miss relevant information at other scales.

We hypothesize that these challenges can be addressed by leveraging on pyramid methods [1]. The image pyramid technique offers an efficient framework that mirrors the multiple scales of processing in the human visual system. Thus, with this paper, *we speculate on the importance of pyramid representations for capturing image relevancy at different levels of detail (both semantically and visually).* With this goal, we introduce a novel deep architecture – named PyrNet– which aims to grasp and leverage on pyramidal information to better tackle the Re-ID problem.

Contributions. Concretely, our contribution is a novel siamese architecture which: (i) is capable of capturing different semantic concepts belonging to the input by extracting information at different levels of detail (i.e., at different

network depths); (ii) aggregates the minutiae captured at various levels to exploit multiple Re-ID decisions within a single model; (iii) adopt a mixed learning strategy that combines identity and similarity learning.
These objectives are achieved as follows.

i) The striped pyramidal block is introduced in a deep architecture at different depths (Section 3.3). The proposed pyramidal pooling captures different minutiae of an image by processing it at various scales, while the horizontal stripes pooling strategy carries information about the relative displacements of image features.

ii) We propose to learn separate classifiers for each striped pyramidal representation, then aggregate them (Section 3.4). This allows us to fuse the Re-ID decisions taken by looking at features with different levels of detail and semantic meanings.

iii) The mixed learning strategy (Section 3.5) allows to gradually shift from the "simple" identification task to the "more difficult" similarity learning problem. We first obtain a robust person representation (identification task) that is gently modified such that it should be similar for an akin person, dissimilar otherwise (similarity learning task).

To substantiate our contributions, we have conducted extensive experiments on three Re-ID benchmark datasets. Our solution achieves better performance than existing methods, while introducing negligible computational complexity.

2. Related Work

The Re-ID community is currently very active [52]. A brief overview follows.

Hand-Crafted Visual Features. Works belonging to this group address the Re-ID problem by designing discriminative appearance feature descriptors. Multiple local and global feature [35] were combined with patch matching strategies [68], saliency learning [38], joint attributes [24] and camera network-oriented schemes [6]. Among all the methods in this category, to date, the most widely used appearance descriptors are the Gaussian of Gaussian (GOG) [39], the Local Maximal Occurrence (LOMO) [29] and the Weighted Histogram of Overlapping Stripes (WHOS) [32].

Optimal Matching Metrics. Approaches grouped in the second family learn an optimal non-Euclidean dissimilarity measure. Specifically, metric learning approaches were introduced y leveraging on positive semi-definite conditions [30] , exploiting the null-space [64] or acceleration techniques for fast optimization [36]. While most of the existing methods capture the global structure of the dissimilarity space, local solutions [28, 41, 65] were also proposed. Following the success of both approaches, methods combining them in ensembles [72, 40] were introduced. Different solutions yielding similarity measures were also investigated by proposing to learn listwise [9] and pairwise [72] similarities.

Deep Learning. Currently, the best Re-ID performance on benchmark datasets is obtained by deep learning-based solutions (e.g., [2, 51, 44]). The success of such approaches is commonly driven by the exploitation of body part features. Following the impressive progress of human pose estimation [56, 22, 5, 14], methods exploiting the output of such frameworks were proposed [69, 47, 54, 33]. However, the conceptual gap between the pose estimation and the Re-ID problem does not guarantee that the detected body parts are optimal for the Re-ID task. In light of such considerations, body partitioning estimators have been abandoned in favor of methods either considering fixed body parts [49] or attention-inspired mechanisms [61, 34, 67, 43]. More specifically, in [49] authors introduced a network simultaneously looking at image stripes which are then separately considered through a part-based classifier. Similarly, in [61], highly active locations of different feature maps were exploited to identify regions of interest later used for part loss computation. The attention mechanism [60] was considered in [34, 67] to let the network model decide which body regions are more relevant for Re-ID.

In this paper, we propose to learn and exploit a human body part and multi-scale-based representation. Our approach neither hinges on human pose estimation frameworks nor on attention-inspired mechanisms which require labeled part mask/box data or additional computationally demanding network branches (e.g., [67, 66, 43]). With respect to all such methods, [34] and [11] are the closest to our approach. Specifically, in [34] an attention mechanism is used to generate multi-level features which are concatenated with the output of an Inception-v2 architecture [50] to form the feature representation. In [11], last ResNet convolutional layer feature maps are split into horizontal stripes which are separately considered for identity classification.

Though sharing the idea of considering features extracted at different semantic levels, there are significant differences with our method. Specifically, our siamese architecture introduces (i) a pyramidal block that captures multi-scale image features at different depths of the architecture and directly consider such information to compute a loss related to a specific semantic concept; (ii) an aggregated objective function that leverages on separate losses computed with respect to features extracted at different levels of detail; (iii) a joint identification and similarity learning strategy.

3. PyrNet

Our goal is to take a couple of person images and determine their similarity. Towards this end, we introduce a

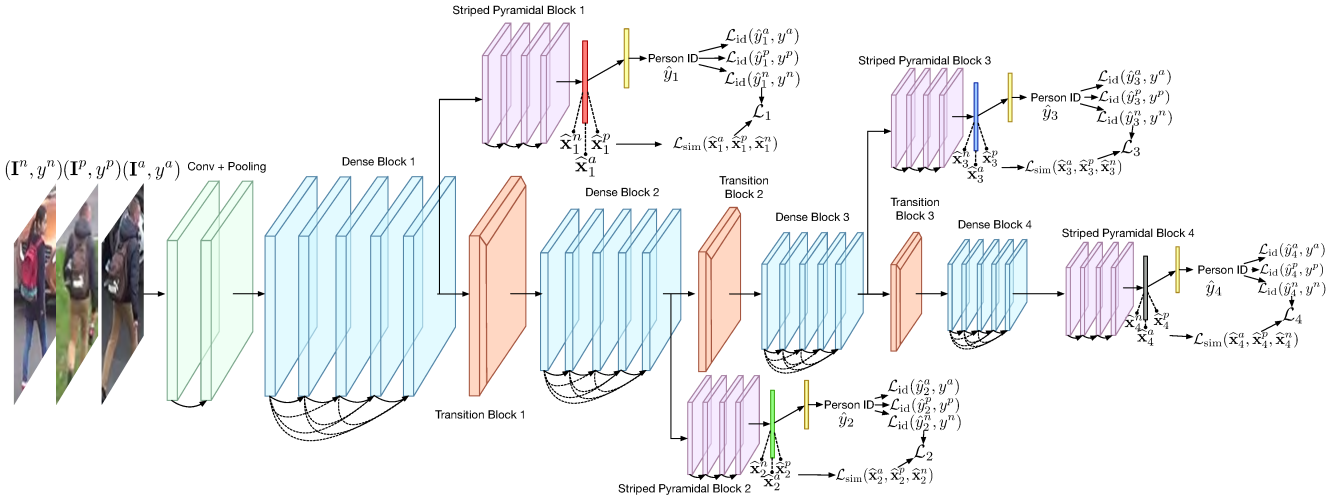

Figure 1: Proposed PyrNet architecture. The siamese network processes a batch of three tuples at a time, each consisting of an image and the corresponding label. Two contain data of a same person (i.e., $(\mathbf{I}^a, y^a)$ and $(\mathbf{I}^p, y^p)$), while the third one (i.e., $(\mathbf{I}^n, y^n)$) carries data acquired for a different individual. Multi-scale and semantically different features are obtained through the striped pyramidal blocks introduced at various depths. Then each of these is separately processed (i) to predict the image identity, and (ii) to bring "closer" the representations obtained from images of the same person and push "further" the representations generated for different individuals.

novel deep architecture shown in Figure 1.

3.1. Notation and Definitions

Let $\mathcal{I} = \{\mathbf{I}\}_{i=1}^{n}$ with $\mathbf{I} \in \mathbb{R}^{W \times H}$ be a set of training images acquired by a camera network. Also let $(\mathbf{I}^a, y^a)$, $(\mathbf{I}^p, y^p)$, and $(\mathbf{I}^n, y^n)$ denote the *anchor*, *positive*, and *negative* tuples with y representing the identity of a person. The three tuples are selected such that $y^a = y^p$ and $y^a \neq y^n$. For each image multiple visual feature representations are computed to capture different image details. Given a level of detail denoted as $l \in \mathbb{N}$ and the feature extraction function $f_l : \mathbb{R}^{W \times H} \mapsto \mathbb{R}^d$ parameterized by a set of trainable parameters $\mathcal{W}_l$, the feature representation of $\mathbf{I}$ at the l-th level of detail is $\widehat{\mathbf{x}}_l = f_l(\mathbf{I}; \mathcal{W}_l)$.

3.2. Backbone Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) can be substantially deeper, more accurate, and efficient to train if they contain shorter connections between layers close to the input and those close to the output. Leveraging on such outcomes, the proposed PyrNet builds upon the recent success of the DenseNet architecture [20]. While other similar solutions (e.g., ResNet [17]) may have been considered, we have chosen the DenseNet one since it has several compelling advantages: it alleviates the vanishing-gradient problem, substantially reduces the number of trainable parameters, and improves efficiency by increasing variation in the input of subsequent layers.

The DenseNet architecture is mainly composed by a set of alternating *dense blocks* and *transition blocks*.

A dense block encloses a set of stacked composite functions consisting of batch normalization (BN), rectified linear unit (ReLU) and a 3×3 convolution (Conv). A composite function processes the feature maps generated by all the preceding composite functions that are part of the same block.

A transition block acts as a down-sampling layer that changes the size of the input feature maps. Specifically, a transition layers consist of a BN layer followed by 1×1 Conv and 2×2 average pooling.

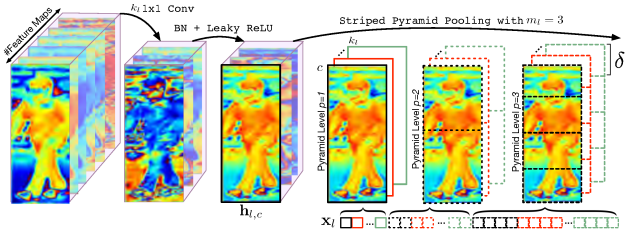

Figure 2: Striped pyramidal block composition with $m_l = 3$ pyramid levels. The number of input feature maps is initially reduced by means of a 1×1 Conv layer. Then the result of subsequent BN and ReLU operations is processed by a striped pyramid pooling layer. This pools the features of overlapping feature map stripes generated by considering different image pyramid levels. Resulting pooled features are then concatenated to generate the striped pyramidal pooled vector $\mathbf{x}_l$.

3.3. Striped Pyramidal Block

The image pyramid approach offers a flexible multi-resolution framework that mirrors the multiple scales of processing in the human visual system [1]. This captures object and features at different sizes.

We propose to use such a multi-scale feature property in our PyrNet architecture. This is achieved by the Striped Pyramidal Block (SPB) which analyzes a given feature map at different scales (see Figure 2). More specifically, the features maps generated by the l-th dense block of the backbone architecture are fed into a `Conv` layer with k_l 1×1 kernels, which is followed by `BN` and `Leaky ReLU` (with a negative slope of 0.1) layers. The c-th feature map resulting from such a step, denoted as $\mathbf{h}_{l,c}$, is then processed by the striped pyramid pooling function parametrized by the number of pyramid levels m_l. This computes the vector

$$\mathbf{x}_{l,c} = \left[\mathcal{P}\left(\mathbf{h}_{l,c}(:, j : j + \delta)\right) | j \in \{0, \delta, \cdots, (s-1)\delta\}\right]^T \tag{1}$$

where $s = 2^{p-1}$ is the number of vertical stripes into which the image is divided at level $p \in \{1, \cdots, m_l\}$ of the pyramid. Each stripe with height $\delta = \lfloor \frac{H}{s} \rceil$ is fed to the stripe pooling function $\mathcal{P}(\cdot) : \mathbb{R}^{W \times \delta} \mapsto \mathbb{R}$ (e.g., the max/average pooling operator). The feature vector obtained by applying such an operator to all the k_l input feature maps yields to the *striped pyramidal pooled vector* $\mathbf{x}_l \in \mathbb{R}^{d_l}$ where $d_l = k_l(2^{m_l} - 1)$. Such an operator captures different features at multiple levels of detail. The horizontal stripes strategy grasps information about the relative vertical displacements of image features which is crucial in Re-ID (we no dot expect the legs features to be above the torso ones).

Since the striped pyramid pooling function has no trainable parameters, the whole block introduces a limited set of parameters to be learned. This boils down to the number of 1×1 convolutions and the `BN` mean and standard deviation (i.e., to $k_l^{-1}k_l + 2k_l$ parameters in total, where k_l^{-1} denotes the number of features maps received by the SPB).

To improve network efficiency and to compress noisy data as much as possible while preserving information about what the SPB output represents, a bottleneck block is introduced after each SPB. This consists of a fully connected (`FC`) layer with 512 neurons followed by `BN`, `Leaky ReLU` (with a negative slope of 0.1) and `Dropout` with 0.5 probability. The output of such a block is the *bottleneck feature vector* $\widehat{\mathbf{x}}_l \in \mathbb{R}^{512}$.

3.4. Visual Concepts

CNNs emit features that are semantically tied to the considered level of depth (i.e., first layers respond to colors, edges and corners, while deeper layers are activated by more complicated and abstracted features [63]). Due to this, as shown in [3], different layers have different importance in retrieving objects sharing a specific characteristic. Thus the output of the backbone architecture at its l-th level might contain a set of features representing semantically different concepts that can have different importance for Re-ID.

To incorporate such intuitions within the proposed architecture, we added an SPB at the output of each dense block of the backbone architecture. Then, separately exploit the features generated by each of such blocks to compute specialized identity and similarity losses. This performs Re-ID using different visual concepts.

3.5. Mixed Learning Strategy

Recent works demonstrated that robust representations can be learned by fine-tuning an existing architecture using an identity loss [71, 62, 49]. Others showed that similar results can be obtained by learning a similarity measure through Siamese architectures looking at image pairs or triplets [2, 18, 15]. In this paper, we combine both schemes into a single joint identity-similarity loss.

Let consider the bottleneck features computed with l levels of detail for the anchor, positive and negative images of a triplet. These are $\widehat{\mathbf{x}}_l{}^a$, $\widehat{\mathbf{x}}_l{}^p$, and $\widehat{\mathbf{x}}_l{}^n$, respectively.

Identity Loss. Each of such feature vectors is then fed to an `FC` layer predicting the identity label $\hat{y}_l$. This is considered together with the ground truth label y to compute the cross entropy loss $\mathcal{L}_{\text{id}}(\hat{y}_l, y)$.

Similarity Loss. The margin ranking loss $\mathcal{L}_{\text{sim}}(\widehat{\mathbf{x}}_l{}^a, \widehat{\mathbf{x}}_l{}^p, \widehat{\mathbf{x}}_l{}^n)$ is computed as

$$\max\left(\|\widehat{\mathbf{x}}_l{}^a - \widehat{\mathbf{x}}_l{}^p\| - \|\widehat{\mathbf{x}}_l{}^a - \widehat{\mathbf{x}}_l{}^n\| + \alpha, 0\right). \tag{2}$$

This loss ensures that the representation of the positive sample is closer to the anchor sample than that of the negative one, by at least a margin α.

Mixed Loss. The two aforementioned losses are then combined to obtain the mixed loss

$$\mathcal{L}_l = \lambda \mathcal{L}_{\text{sim}}(\widehat{\mathbf{x}}_l{}^a, \widehat{\mathbf{x}}_l{}^p, \widehat{\mathbf{x}}_l{}^n) + (1-\lambda)/3 \sum_{\pi \in \{a,p,n\}} \mathcal{L}_{\text{id}}(\hat{y}_l^\pi, y^\pi) \tag{3}$$

with π indicating the anchor, positive and negative element of the triplet and $\lambda \in [0, 1]$ controlling the trade-off between the identity and the similarity losses. The sum of the mixed losses computed for all the levels of detail (i.e., $\mathcal{L} = \sum_l \mathcal{L}_l$) is considered for network optimization.

4. Experimental Results

4.1. Datasets

We evaluated our approach on three publicly available benchmark datasets, namely CUHK03[26, 75], Market-1501 [70], and Duke-MTMC [42] (see Figure 3 for few sample images). We report on the average performance using the Cumulative Matching Characteristic (CMC) and the mean Average Precision (mAP) indicators.

CUHK03-NP[1]. The CUHK03 dataset [26] contains 14,096 images of 1,467 different identities. Each person is captured from two cameras in the CUHK campus. The dataset provides both manually annotated and and DPM-detected

[1] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

(a) (b) (c)

Figure 3: Images of 15 different persons (columns) acquired by disjoint cameras (rows) from the (a) CUHK03, (b) Market-1501 and (c) DUKE-MTMC-reID datasets.

bounding boxes. We have run the experiments under both scenarios with the protocol proposed in [75] providing fixed train/test splits with 767 and 700 disjoint identities.

Market-1501[2]. The Market-1501 dataset [70] contains 32,668 images of 1,501 persons taken from 6 disjoint cameras. Multiple images of a same person have been obtained by means of a state-of-the-art detector. The provided train/test partitions contain 750 and 751 person identities. The training set comes with 12,936 images, while the gallery and the query sets have 19,732 and 3,368 images.

Duke-MTMC[3]. The Duke-MTMC-reID (Duke) dataset contains manually annotated bounding boxes generated by 8 cameras [42]. Out of a total of 1,812 people, 1,404 occur in more than one camera. These form the training and test sets consisting of 702 persons each. The training set includes 16,522 images. The test set has 17,661 gallery and 2,228 query images, respectively.

4.2. Implementation Details[4]

We have taken the ImageNet pretrained DenseNet architecture and added an SPB at the output of each dense block –indicated through $l \in L = \{1, \cdots, 4\}$. To have striped pyramid pooled vectors having a similar number of features within each SPB, we set $m_l = 8 - l$ and $k_l = 2^{l+|L|}$ (see Section 4.3.1 for more details). The sum of Euclidean distances between the bottleneck feature vectors computed at different levels of detail for probe and gallery images is considered to compute the results.

Hard-Triplet Mining. Mining hard triplets is crucial for learning [16, 57]. We follow a similar strategy to [13]. At each epoch, we randomly sample 5,000 images and extract the corresponding bottleneck feature vectors to compute all pairwise dissimilarities. Then, for each of the 5,000 anchors, we randomly pick a positive sample among the 3 ones that have the largest dissimilarity and a negative sample among the 10 ones that have the smallest dissimilarity. This is a simple and effective strategy which brings limited computational overhead (the whole process requires less than 30 seconds with the considered configuration).

Mixed Learning Strategy. We argue that, as humans, a set of notions can be learned more easily when each concept to be grasped is presented by increasing degree of complexity. Thus, we adopt a curriculum learning strategy [4]. We initially focus on the person identification task before shifting to the more challenging similarity learning problem. This is achieved by increasing λ with a step of $1/\#\text{epochs}$.

Data Augmentation. Horizontal flipping is applied with 50% chance. Photometric distortions [19] and the AlexNet-style color augmentation [17] are applied to a 192×384 random crop of the 204×408 original image. Finally, the random erasing augmentation [76] is applied with an initial 50% probability. This linearly increments at each epoch and reaches a 95% probability at the end of training.

Optimization. Training was performed for 100 epochs via stochastic gradient descent with mini-batches containing 16 samples. The initial learning rate has been set to 0.1, reduced by a factor of 10 every 40 epochs. The learning rate of the DenseNet layers has been set to be $10\times$ smaller than others. Momentum has been set to 0.9 and a weight decay penalty of 0.0005 had been applied to all layers.

4.3. Performance Analysis

4.3.1 Ablation Study

Backbone Architecture. The original DenseNet architecture has 4 different variants with 121, 161, 169, and 201 layers [20]. In Table 1, the performance achieved using these models. The best ranking results as well as the highest mAP are obtained by the 201 layers variant, thus it has been considered to run all the following experiments.

Visual Concepts. The proposed architecture leverages on

[2] http://www.liangzheng.org/Project/project_reid.html
[3] https://github.com/layumi/DukeMTMC-reID_evaluation
[4] Code available at https://github.com/iN1k1

Table 1: Performance comparison on the Duke dataset using different backbone architecture depths. Evaluation time is computed for a single probe image. Best result is in red, second best in blue.

Depth	Rank-1	Rank-5	mAP	Train/Eval Time [s]
121	81.73	91.29	68.18	10907 / 0.047
161	**84.37**	**92.01**	**70.13**	19310 / 0.059
169	83.17	**91.70**	69.30	20127 / 0.065
201	**84.69**	**92.01**	**70.69**	22387 / 0.068

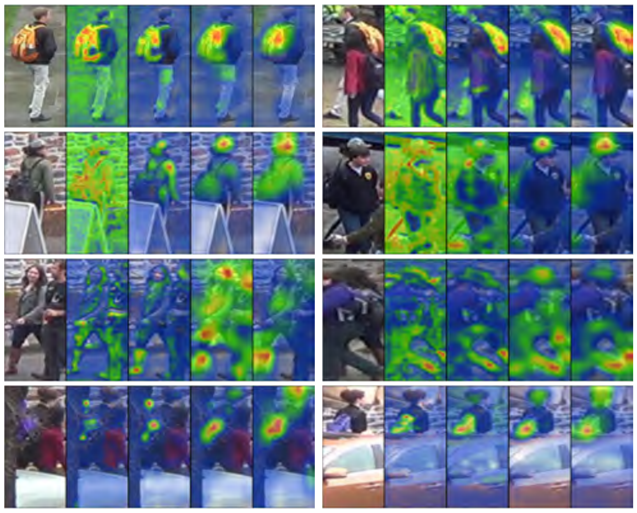

Figure 4: Visual concepts and corresponding implicit attention computed for the 4 different levels of details considered by the proposed architecture. Each row show a different person, while each column depicts the input image overlaid with color-coded visual attention heatmaps for levels 1 to 4 (from left to right). Visual attention goes from blue (low importance) to red (high importance).

different visual representations obtained from the output of each SPB. To understand how important the representation at a specific level of detail is, we computed the results in Table 2. These show the performance of PyrNet using different combinations of the striped pyramid pooled vector (i.e., $\mathbf{x}$) and the bottleneck feature vector (i.e., $\widehat{\mathbf{x}}$).

Results show that, the features generated through the output of the 3$^{\text{rd}}$ and 4$^{\text{th}}$ dense blocks yields to the highest performance when either separately or jointly considered. Overall, the best mAP performance (74.02%) is obtained by considering only the bottleneck feature vectors computed from such blocks. This might indicate that features representing high levels of abstractions are more meaningful than those related to low level semantics. This is substantiated by the performance obtained considering the features generated by low-level dense blocks (i.e., $l \in \{1, 2\}$). Indeed, in such cases, the corresponding mAPs do not reach the 60% which is always surpassed by considering a single high level semantic representation (i.e., $l \in \{3, 4\}$). In light of such outcomes, only the bottleneck feature vectors $\widehat{\mathbf{x}}_3$ and $\widehat{\mathbf{x}}_4$ are considered in the following experiments.

Table 2: Rank-1/mAP performance achieved by our solution using either the striped pyramid pooled vector ($\mathbf{x}$) or the bottleneck feature vector ($\widehat{\mathbf{x}}$) representation at different levels of detail (l). When jointly considered ($\mathbf{x} + \widehat{\mathbf{x}}$), the sum of Euclidean distances is considered. Best mAP result is in red, second best in blue.

$l \downarrow$	$\mathbf{x}_l$	$\widehat{\mathbf{x}}_l$	$\mathbf{x}_l + \widehat{\mathbf{x}}_l$
1	49.28/28.58	63.42/43.43	61.76/41.57
2	71.10/49.42	77.56/59.36	77.60/58.95
3	84.43/67.97	86.71/72.87	86.58/72.80
4	**85.32/69.59**	86.45/72.56	86.62/72.81
1+2	64.86/42.35	74.60/56.07	73.38/54.30
1+3	77.78/59.34	83.44/68.16	82.41/67.19
2+3	81.51/64.31	85.01/71.70	84.83/71.01
1+4	79.89/62.34	82.94/68.48	83.26/68.19
2+4	83.12/66.64	85.50/71.97	85.32/71.65
3+4	**85.14/69.77**	**87.07/74.02**	**87.48/73.81**
1+2+3	76.93/58.07	82.81/67.82	82.05/66.60
1+2+4	78.68/60.60	82.85/68.28	82.50/67.49
1+3+4	82.23/65.91	85.82/72.21	85.23/71.49
2+3+4	84.16/68.06	**86.40/73.83**	**86.31/73.20**
1+2+3+4	81.06/64.35	85.23/71.60	84.69/70.69

To better understand what are the visual concepts that are learned at the different levels of the architecture, we have computed the visual results shown in Figure 4. These show the implicit attention that is payed by the PyrNet at the 4 different levels of detail through a color-coded heatmap. First levels heavily drive the focus of subsequent ones. The first two rows show that a large portion of the image is considered to be relevant for the first levels. Such a portion is significantly reduced at deeper levels, yet all agree on the more relevant image regions. A different trend is shown for the last two samples where first levels focus on weak visual concepts (object parts, e.g., a few purple backpack spots hidden under the tree branches in the last row). These are expanded to a more complete semantical representation at deeper levels (e.g., the whole purple backpack).

Striped Pyramidal Block. The proposed SPB is controlled by the number of 1×1 convolutions (k_l) and the number of pyramid levels (m_l). In Figure 5, we show the mAP performance obtained by using a fixed number of 1×1 convolutions for all levels of detail (i.e., $k_l \in \{32, 64, 128, 256\}$) as well as on the results achieved using a different number of feature maps for each level of detail (i.e., $k_l = 2^{l+|L|}$). Results show that performance improves when the number of feature maps increases, even though it is influenced by the number of pyramid levels. The best overall result is obtained with a dynamic number of pyramid levels (i.e., with $m_l = 8 - l$) and a variable number of feature maps. We hypothesize that this is due to the spatial resolution of the pooled regions as well as to the the dimensionality of each striped pyramidal pooled vector. Indeed, with deeper layers, the spatial resolution of the feature maps is reduced, hence with a large number of pyramid levels we pool image stripes comprising only few vertical pixels. This produces high-dimensional vectors that depend on the vertical location of the visual concepts within the image.

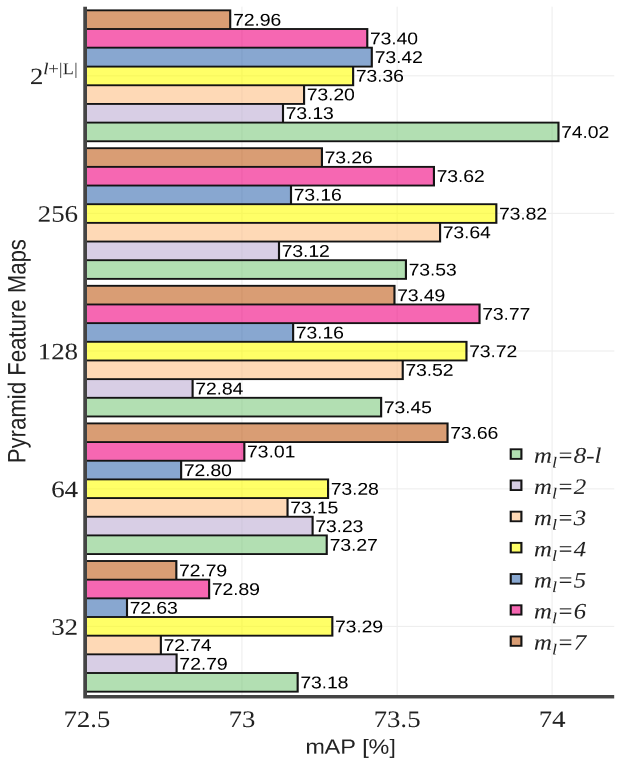

Figure 5: mAP performance on the Duke dataset using different number of 1×1 convolutions (k_l) and pyramid levels (m_l) within the SPB block.

When using a number of feature maps and pyramid levels that are dependent on a level of detail, the SPB block achieves three advantages: (i) it generates a proper number of feature maps that within DenseNet architecture are expected to increase with network depth; (ii) it keeps an appropriate dimension of the stripes; (iii) the pooling operator thus works on significant information.

Mixed Learning Strategy. We propose to train PyrNet through a mixed identity–similarity loss. To understand how the balance between the two different task affects the achieved performance, we have computed the results shown in Figure 6. These show that learning a specific identity representation by assigning less weight (e.g., $\lambda \in \{0, \cdots, 0.4\}$) to the similarity task is very important. This is even more evident when only the similarity task is considered ($\lambda = 1$). In such a case, the achieved mAP is just 17.92%. On the other hand, the best performance is obtained when λ is dynamically modified by linearly increasing its value till it equals 1 at the end of training. This generates a gap of about 4% with the best results obtained through a static value of λ (i.e., $\lambda = 0.3$). This substantiate the importance of having a suitable curriculum learning strategy for Re-ID.

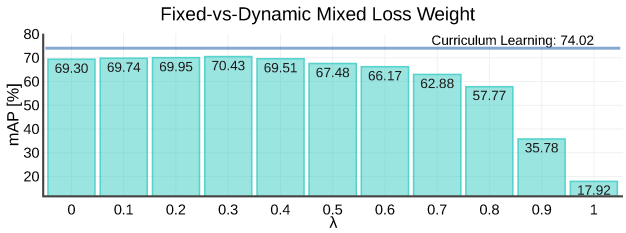

Figure 6: mAP performance on the Duke dataset using different values for the identity–similarity loss fusion weight. The horizontal solid line on the top represents the mAP obtained by dynamically increasing the value of λ over training.

4.3.2 State-of-the-art Comparisons

In Table 3, we report on the performance comparisons with existing methods on the three considered datasets.

CUHK03-NP. The comparison with the existing literature shows that the proposed method significantly outperforms current state-of-the-art solutions on this dataset (under both the *detected* and *labeled* settings). Specifically, PyrNet has more than 15% and about 20% gaps with the previously best performing method [7] when the rank-1 and the mAP metrics are considered, respectively. The gain is even more evident when a re-ranking [75] step is exploited. In such a case, our method is the only one above 80% rank-1 and mAP for the *labeled* scenario.

Market-1501. In the following, we report on the performance achieved by the proposed method considering both the single-query and the multiple-query protocols [70]. Results show that similar performance to the state-of-the-art are achieved under both scenarios. Specifically, the highest rank-1 and rank-5 performance (95.2% and 98.8%) are obtained the considering the multiple-query setup. With the single-query setup, we reach the second place on the leader board. The best performance are obtained by [23], which however is trained with an aggregation of 10 different Re-ID benchmarks resulting in a total of about 111,000 images with higher spatial resolution. This demonstrates that the proposed architecture is able to learn meaningful feature representations as well as a similarity measure with limited data. In addition, when re-ranking [75] is considered, our method has the best performance under both scenarios.

Duke-MTMC. Results demonstrate that the proposed approach achieves the best performance on such a dataset. In particular, it significantly outperforms more complex recent works that leverage on additional clues [7], attention-like mechanisms [59], and pose/view-aware solutions [33, 43]. As for the Market-1501, the closest approach to our work is [23], which however pays about 1% in terms of both rank-1/5 and mAP performance with respect to PyrNet. With the re-ranking step [75], our method is the only one that achieves more than 90% accuracy at rank 1. It also reaches the highest mAP (87.7%) with a gap of about 3% with [23].

Table 3: State-of-the-art comparisons on the three considered datasets. Results for CUHK03 are shown considering *detected/labeled* bounding boxes. Results for Market-1501 have been computed considering the *single-query/multiple-query* protocol. Best result is in red, second best in blue.

Method	CUHK03		Market-1501			Duke-MTMC			Publication
	Rank-1	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	
SpindleNet	–	–	76.9/–	91.5/–	–/–	–	–	–	CVPR2017 [66]
Part-Aligned	–	–	81.0/–	92.0/–	63.4/–	–	–	–	ICCV2017 [67]
HydraPlus-Net	–	–	76.9/–	91.3/–	–/–	–	–	–	ICCV2017 [34]
MSCAN	–	–	80.3/86.8	–/–	57.5/66.7	–	–	–	arXiv2017 [25]
Verif-Identif	–	–	79.5/85.8	–	59.9/70.3	68.9	–	49.3	TOMM2018 [74]
LSRO	–	–	84.0/88.4	–	66.1/79.1	67.7	–	47.1	ICCV2017 [73]
SVDNet	41.5/40.9	37.3/37.8	82.3/–	92.3/–	62.1/–	76.7	86.4	56.8	ICCV2017 [48]
DPFL	40.7/43.0	37.0/40.5	88.9/–	92.3/–	73.1/–	73.2	–	60.6	ICCV2017 [10]
APR	–	–	84.3/–	93.2/–	64.7/–	70.7	–	59.1	arXiv2017 [31]
PAN	36.3/36.9	34.0/35.0	82.81/–	93.5/–	63.3/–	71.6	83.9	51.5	arXiv2017 [73]
TriNet	–	–	84.9/90.5	94.2/96.3	69.1/76.4	–	–	–	arXiv2017 [18]
RDR	–	–	92.2/**94.7**	**97.9**/**98.6**	81.2/**87.3**	85.2	**93.9**	72.8	arXiv2018 [2]
PSE	–	–	87.7/–	94.5/–	69.0/–	79.8	89.7	62.0	CVPR2018 [43]
HA-CNN	41.7/44.4	38.6/41.0	91.2/93.8	–/–	75.5/82.8	80.5	–	63.8	CVPR2018 [27]
Pose-transfer (D, Tri)	41.6/45.1	38.7/42.0	87.6/–	–/–	68.9/–	78.5	–	56.9	CVPR2018 [33]
AACN	–	–	85.9/89.8	–/–	66.9/75.1	76.8	–	59.2	CVPR2018 [59]
MLFN	**52.8**/**54.7**	**47.8**/49.2	90.0/92.3	–	74.3/82.4	81.0	–	62.8	CVPR2018 [7]
DuATM	–	–	91.4/–	97.1/–	76.6/–	81.8	90.2	68.6	CVPR2018 [45]
DKP	–	–	90.1/–	96.7/–	75.3/–	80.3	89.5	63.2	CVPR2018 [44]
AOCS	47.1/–	43.3/–	86.5/91.3	–	70.4/78.3	79.2	–	62.1	CVPR2018 [21]
GCSL	–	–	93.5/–	–	81.6/–	84.9	–	69.5	CVPR2018 [8]
BraidNet-CS+SRL	–	–	83.7/–	–	69.5/–	76.4	–	59.5	CVPR2018 [53]
MGCAM	46.7/50.1	46.9/**50.2**	83.8/–	–	74.3/–	–	–	–	CVPR2018 [46]
SPReID	–	–	**93.7**/–	97.6/–	**83.4**/–	**85.9**	92.9	**73.3**	CVPR2018 [23]
PyrNet	**68.0**/**71.6**	**63.8**/**68.3**	**93.6**/**95.2**	**98.2**/**98.8**	**81.7**/**86.7**	**87.1**	**94.1**	**74.0**	Proposed
AACN+ReRank [75]	–	–	88.7/92.2	–/–	83.0/87.3	–	–	–	CVPR2018 [59]
RDR+ReRank [75]	–	–	**93.0**/**94.2**	95.9/**96.9**	90.0/**91.2**	89.4	93.6	85.6	arXiv2018 [2]
AOCS+ReRank [75]	54.6/–	56.1/–	88.7/92.5	–	83.3/88.6	84.1	–	78.2	CVPR2018 [21]
SPReID+ReRank [75]	–	–	**94.6**/–	**96.8**/–	**90.9**/–	88.9	93.3	85.0	CVPR2018 [23]
PyrNet +ReRank [75]	**77.1**/**80.8**	**78.7**/**82.7**	**94.6**/**96.1**	**96.9**/**97.9**	**91.4**/**94.0**	**90.3**	**94.3**	**87.7**	Proposed

4.3.3 Discussion and Limitations

The conducted ablation and the achieved performance on the three considered benchmarks demonstrate that: (i) Focusing on different semantically meaningful information is very useful for achieving good results. (ii) The proposed multi-scale pyramid representation can capture relevant image details at a low computational cost with a limited number trainable parameters. (iii) Progressively focusing on a more difficult task while training though curriculum learning significantly increase Re-ID performance. Results show that, such strategies yield to significant performance improvements over the state-of-the-art methods.

The conducted ablation study demonstrated that most of the relevant information comes from the deeper layers of the considered DenseNet backbone. However, visual attention insights show that lower layers are also fundamental to correctly drive the selection of the regions to be considered. Thus, a careful selection of backbone dense blocks to be considered for Re-ID is very important. We hypothesize that such a limitation can be mitigated by including an attention mechanism [60] that selects the most appropriate features out of the ones produced by all SPBs together.

5. Conclusion

In this paper we have proposed a novel neural architecture to address the person Re-ID problem. First, to try mimicking the multi-scale processing conducted within the human vision system, we have introduced an SPB. This leverages on the pyramid method to capture the person appearance information with different levels of detail. Then, to capture the importance of features generated at different depths (hence with different semantic meanings), an SPB is added at the output of each dense block belonging to the backbone DenseNet architecture. The identity and similarity losses computed by each SPB are jointly considered in a mixed learning strategy. Significant performance improvement is obtained by increasing the complexity of the learning strategy over time, through curriculum learning. Results conducted on three datasets show the achievement of, compared to the considered backbone and state-of-the-art methods, better performances at a lower computational cost.

Acknowledgment

This research was partially supported by the PRESNET MoD project CIG68827500FB. The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. Pyramid methods in image processing. 1984.

[2] J. Almazan, B. Gajic, N. Murray, and D. Larlus. Re-ID done right: towards good practices for person re-identification. *arXiv preprint*, 2018.

[3] M. Aubry and B. C. Russell. Understanding deep features with computer-generated imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2875–2883, 2015.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum Learning. In *International Conference on Machine Learning*, 2009.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *International Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[6] A. Chakraborty, A. Das, and A. Roy-Chowdhury. Network Consistent Data Association. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2015.

[7] X. Chang, T. M. Hospedales, and T. Xiang. Multi-Level Factorisation Net for Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[8] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group Consistent Similarity Learning via Deep CRF for Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018.

[9] J. Chen, Z. Zhang, and Y. Wang. Relevance Metric Learning for Person Re-Identification by Exploiting Listwise Similarities. *IEEE Transactions on Image Processing*, 7149(c):1–1, 2015.

[10] Y. Chen, X. Zhu, and S. Gong. Person Re-Identification by Deep Learning Multi-Scale Representations. In *International Conference on Computer Vision*, 2017.

[11] Y. Fu, Y. Wei, Y. Zhou, and H. Shi. Horizontal Pyramid Matching for Person Re-identification. *arXiv*, 2018.

[12] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni. Discriminant Context Information Analysis for Post-Ranking Person Re-Identification. *IEEE Transactions on Image Processing*, 26(4):1650–1665, apr 2017.

[13] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-End Learning of Deep Visual Representations for Image Retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[14] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. In *International Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[15] Y. Guo and N.-M. Cheung. Efficient and Deep Person Re-Identification using Multi-Level Similarity. In *International Conference on Computer Vision and Pattern Recognition*, 2018.

[16] B. Harwood, V. K. B. G, G. Carneiro, I. Reid, and T. Drummond. Smart Mining for Deep Metric Learning. In *International Conference on Computer Vision*, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 770—-778, 2016.

[18] A. Hermans, L. Beyer, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint*, 2017.

[19] A. G. Howard. Some Improvements on Deep Convolutional Neural Network Based Image Classification. *ArXiv e-prints*, 1312.5402, 2013.

[20] G. Huang, L. V. D. Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *International Conference on Computer Vision and Pattern Recognition*, 2017.

[21] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially Occluded Samples for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018.

[22] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conferece on Computer Vision*, volume 9910 LNCS, pages 34–50, 2016.

[23] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah. Human Semantic Parsing for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, 2018.

[24] R. Layne, T. Hospedales, and S. Gong. Re-id: Hunting Attributes in the Wild. In *Proceedings of the British Machine Vision Conference 2014*, pages 1.1–1.12. British Machine Vision Association, 2014.

[25] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification. *arXiv preprint*, 2017.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 152–159. Ieee, jun 2014.

[27] W. Li, X. Zhu, and S. Gong. Harmonious Attention Network for Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[28] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning Locally-Adaptive Decision Functions for Person Verification. In *International Conference on Computer Vision and Pattern Recognition*, pages 3610–3617. IEEE, jun 2013.

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In *International Conference on Computer Vision and Pattern Recognition*, 2015.

[30] S. Liao and S. Z. Li. Efficient PSD Constrained Asymmetric Metric Learning for Person Re-identification. In *International Conference on Computer Vision*, pages 3685–3693, 2015.

[31] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving Person Re-identification by Attribute and Identity Learning. *arXiv preprint*, mar 2017.

[32] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person Re-Identification by Iterative Re-Weighted Sparse Ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, aug 2015.

[33] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose Transferrable Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[34] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. In *International Conference on Computer Vision*, pages 350–359, 2017.

[35] B. Ma, Y. Su, and F. Jurie. Covariance Descriptor based on Bio-inspired Features for Person Re-identification and Face Verification. *Image and Vision Computing*, 32:379–390, apr 2014.

[36] N. Martinel. Accelerated low-rank sparse metric learning for person re-identification. *Pattern Recognition Letters*, 112:234–240, 2018.

[37] N. Martinel, M. Dunnhofer, G. L. Foresti, and C. Micheloni. Person Re-Identification via Unsupervised Transfer of Learned Visual Representations. In *International Conference on Distributed Smart Cameras*, pages 1–6, Stanford, CA, USA, 2017.

[38] N. Martinel, C. Micheloni, and G. L. Foresti. Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning. *IEEE Transactions on Image Processing*, 24(12):5645–5658, dec 2015.

[39] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical Gaussian Descriptor for Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.

[40] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel. Learning to rank in person re-identification with metric ensembles. In *International Conference on Computer Vision and Pattern Recognition*, 2015.

[41] S. Pedagadi, J. Orwell, and S. Velastin. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[42] 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 Conferece on Computer Vision*, 2016.

[43] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking. In *International Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[44] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. End-to-End Deep Kronecker-Product Matching for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, 2018.

[45] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, 2018.

[46] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided Contrastive Attention Model for Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[47] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-Driven Deep Convolutional Model for Person Re-identification. *IEEE International Conference on Computer Vision*, 2017-Octob:3980–3989, 2017.

[48] Y. Sun, L. Zheng, W. Deng, and S. Wang. SVDNet for Pedestrian Retrieval. In *International Conference on Computer Vision*, 2017.

[49] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). In *International Conference on Computer Vision and Pattern Recognition*, 2018.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision, 2016.

[51] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang. Eliminating Background-bias for Robust Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 5794–5803, 2018.

[52] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics. *ACM Computing Surveys*, 46(2):1–37, nov 2013.

[53] Y. Wang, Z. Chen, F. Wu, and G. Wang. Person Re-identification with Cascaded Pairwise Convolutions. In *International Conference on Computer Vision and Pattern Recognition*, pages 1470–1478, 2018.

[54] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval, 2017.

[55] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. In *ACM Multimedia*, 2017.

[56] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 4724–4732, 2016.

[57] C.-y. Wu, U. T. Austin, A. J. Smola, and U. T. Austin. Sampling Matters in Deep Embedding Learning. In *International Conference on Computer Vision*, pages 2840–2848, 2017.

[58] Z. Wu, Y. Li, and R. J. Radke. Viewpoint Invariant Human Re-Identification in Camera Networks Using Pose Priors and Subject-Discriminative Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1095–1108, may 2015.

[59] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-Aware Compositional Network for Person Re-identification.

In *International Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018.

[60] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv*, 2015.

[61] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep Representation Learning with Part Loss for Person Re-Identification. *ArXiv e-prints*, pages 1–9, 2017.

[62] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep Representation Learning with Part Loss for Person Re-Identification. *arXiv preprint*, 2017.

[63] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *European Conferece on Computer Vision*, 2012.

[64] L. Zhang, T. Xiang, and S. Gong. Learning a Discriminative Null Space for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[65] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-Specific SVM Learning for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2016.

[66] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, and S. Yi. Spindle Net : Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. In *International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1077–1085, 2017.

[67] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-Learned Part-Aligned Representations for Person Re-Identification. In *International Conference on Computer Vision*, pages 3219–3228, 2017.

[68] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[69] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose Invariant Embedding for Deep Person Re-identification. *arXiv preprint*, 2017.

[70] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable Person Re-identification : A Benchmark. In *International Conference on Computer Vision*, 2015.

[71] L. Zheng, Y. Yang, and A. G. Hauptmann. Person Re-identification : Past, Present and Future. *arXiv preprint*, pages 1–20, 2016.

[72] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by Relative Distance Comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, jun 2013.

[73] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian Alignment Network for Large-scale Person Re-identification. *ArXiv e-prints*, 2017.

[74] Z. Zheng, L. Zheng, and Y. Yang. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):1–20, dec 2018.

[75] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking Person Re-identification with k-reciprocal Encoding. In *International Conference on Computer Vision and Pattern Recognition*, 2017.

[76] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random Erasing Data Augmentation. 2017.