Re-ID done right: towards good practices for person re-identification

Jon Almazán[1] Bojana Gajić[2*] Naila Murray[1] Diane Larlus[1]
[1]Computer Vision Group
NAVER LABS Europe
firstname.lastname@naverlabs.com
[2]Computer Vision Center
Dept. de Ciències de la Computació, UAB
bgajic@cvc.uab.es

Abstract

Training a deep architecture using a ranking loss has become standard for the person re-identification task. Increasingly, these deep architectures include additional components that leverage part detections, attribute predictions, pose estimators and other auxiliary information, in order to more effectively localize and align discriminative image regions. In this paper we adopt a different approach and carefully design each component of a simple deep architecture and, critically, the strategy for training it effectively for person re-identification. We extensively evaluate each design choice, leading to a list of good practices for person re-identification. By following these practices, our approach outperforms the state of the art, including more complex methods with auxiliary components, by large margins on four benchmark datasets. We also provide a qualitative analysis of our trained representation which indicates that, while compact, it is able to capture information from localized and discriminative regions, in a manner akin to an implicit attention mechanism.

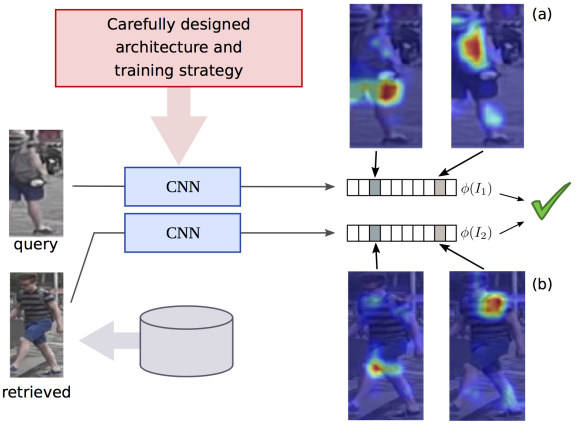

Figure 1. By careful design of our deep architecture and training strategy (Section 3), our approach builds global representations that capture the subtle details required for person re-identification by training the embedding dimensions to respond strongly to discriminative regions/concepts such as the backpack or the hem of the shorts. Heatmaps indicate image regions that strongly activate different dimensions of the embedding.

1. Introduction

Person re-identification (re-ID) is the task of correctly identifying individuals across different images captured under varying conditions, such as different cameras within a network. This task is of high practical value in a wide range of applications including surveillance or content-based image retrieval. Different from classification, there is no overlap between the persons seen at train time and at test time.

Heavily studied for more than two decades [2, 19], most works that address this problem have sought to propose either a suitable image representation, often with hand-crafted rules, or a suitable image similarity metric. Following the great success of deep learning in a large number of computer vision tasks, including image classification [16], object detection [33], and semantic segmentation [8], a dominant paradigm in person re-ID has emerged, where methods use or fine-tune successful deep architectures for this retrieval task [5, 17, 38]. This paradigm leads to compact global image representations well-suited for person re-identification. However, within this general framework there remain many design choices, in particular those related to network architectures, training data, and model training, that have a large impact on the effectiveness of the final person re-ID model. In this paper, we focus on identifying which of these design choices matter.

One potential limitation of using global representations designed for image classification is the absence of any explicit mechanism to tackle the misalignment inherent to human pose variations and person detection errors. Consequently, many recent works in the literature have explored strategies to alleviate this problem by explicitly aligning body parts between images [37, 49], for example by using pre-trained part or human joint detectors, or by enriching the training set with auxiliary data such as attributes [38].

*Work done during an internship at NAVER LABS Europe.

In this work, we adopt a different approach that combines a simple deep network with an appropriate training strategy, and whose design choices were both carefully validated on several datasets. The result is a simple yet powerful architecture that produces global image representations that, when compared using a dot-product, outperform state-of-the-art person re-identification methods by large margins, including more sophisticated methods that rely on attention models, extra annotations, or explicit alignment.

Our contribution is threefold. First, we identify a set of key practices to adopt, both for representing images efficiently and for training such representations, when developing person re-ID models (Section 3). Many of these principles have been adopted in isolation in various related works. However, we show that when applied jointly, significant performance improvements result. We carefully evaluate different modeling and learning choices that impact performance. A key conclusion is that curriculum learning is critical for successfully training the image representation and several of our principles reflect this.

Second, our method significantly improves over previous published results on four standard benchmark datasets for person re-identification (Section 4.3). For instance, we show an absolute improvement of 8.1% mAP in the Market-1501 dataset compared with the current state of the art.

Third, we provide a qualitative analysis of the information captured by the visual embedding produced by our architecture. Our analysis illustrates, in particular, the effectiveness of the model in localizing image regions that are critical for re-ID without the need for explicit attention or alignment mechanisms (Section 4.4). We also show how individual dimensions of the embedding selectively respond to localized semantic regions producing a high similarity between pairs of images from the same person.

We believe that our approach, which is easy to reproduce[1], can serve as a baseline of choice for future improvements in this field of research.

2. Related Work

A vast literature addresses the person re-identification problem (the reader may refer to [19] for a recent survey). Traditionally, works on person re-ID sought to improve similarity scores between images [20, 29, 28], usually through metric learning. These methods typically used color-based histograms as input feature vectors [27, 20, 29, 45, 28] due to their discriminative power particularly with respect to clothing, and also to their small memory footprint. Recent research on person re-identification has mostly focused on end-to-end training of deep architectures. This research has taken two main directions: improving generic deep image representations using sophisticated learning objectives appropriate for person re-identification, or designing task-specific image representations.

Task-specific learning objectives. This line of research most often involves proposing loss functions suitable for the re-ID task, and in particular for learning effective similarity metrics. [59] proposes a metric to learn similarities between an image and a set of images, as opposed to learning similarities between pairs of images as is typical. [58] proposes a method to locally modify, in an online manner at test time using only negative examples, a global similarity metric that was trained offline. [39] added an orthogonality constraint on the final fully-connected layer of a deep network in order to improve the discriminability of the learned features. [60] proposes to train a re-ID model using as a similarity metric a hybrid of the Euclidean, Cosine and Mahalanobis distances. [47] learns an embedding that aims to project images of the same person into the same point in the embedding space. [1] proposes to learn a method to modify image embeddings such that the learned similarity metric between images is smooth in the underlying image manifold. [46] proposes to learn an image embedding for re-ID by training a network to predict both person IDs and attribute labels.

Most recent works use cross-entropy or softmax loss functions for training their person re-identification models. Others treat person re-ID not as a recognition but rather as a ranking problem, and use losses that are more appropriate for ranking. For example, the contrastive loss [41] and the triplet loss or variants thereof [10, 38, 17, 5] have been used to train Siamese architectures. [10] proposes a scheme to limit the size of triplet batches while still obtaining informative samples, while [5] proposes a quadruplet loss, which adds to the triplet loss a term that enforces a margin constraint on the distance between image pairs that are unrelated. [17] shows that, with appropriate training settings, the triplet loss can outperform more complicated objective functions. In this work, we propose several good practices that can be viewed as encouraging curriculum learning (*c.f.* section 3.3) that, when combined with the standard triplet loss, lead to large improvements over previous methods which have used varieties of the triplet loss.

Task-specific representations. Many works in this line have focused on addressing the alignment problem via use of part detectors, pose estimation, or attention models. Spatial transformer networks have been used to globally align images [55] and to localize salient image regions for finding correspondences [32]. In a similar vein, [50, 24, 25] use multiple parallel sub-branches which learn, in an unsupervised manner, to consistently attend to different human body parts. [37] uses a pre-trained pose estimation network to provide explicit part localization, while a similar approach [49] integrates a pose estimation network into their deep re-ID model. [52] uses joint localization to create a new image that contains only the body parts. Rather than

[1] To aid reproducibility we will release trained models and the evaluation code upon acceptance.

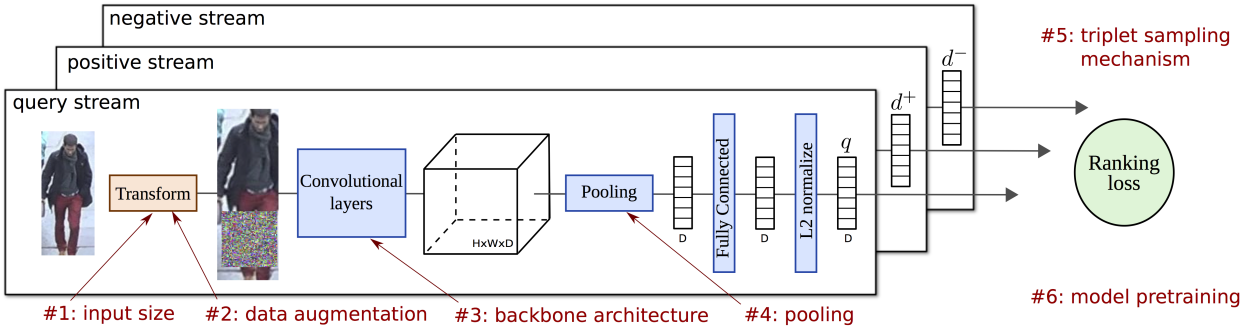

Figure 2. **Summary of our training approach**. Image triplets are sampled and fed to a three stream Siamese architecture, trained with a ranking loss. Weights of the model are shared across streams. Each stream encompasses an image transformation, convolutional layers, a pooling step, a fully connected layer, and an ℓ_2-normalization, all these steps being differentiable. In red we show the steps that require a careful design and that we extensively discuss and evaluate in our paper.

localize parts, [22] represents images with fixed grids and learns cell correspondences across camera views. Several works have proposed multi-scale architectures with mechanisms for automatic scale selection [30] or scale fusion [6]. [21] combines a multi-scale architecture with unsupervised body part localization using spatial transformer networks. In Section 4, we compare to such works and show that our learned representation can address alignment and scale variations without using additional scale, human parsing, or attention models.

Other relevant areas of research in re-ID are data scarcity, re-ranking, and end-to-end re-ID. [56] uses GANs to synthesize crops of pedestrians which were used to train a deep re-ID network in a semi-supervised manner. [57] applies k-reciprocal nearest neighbor reranking to the re-ID problem. [23, 43] both tackle end-to-end re-ID by incorporating person detection into their proposed pipelines.

3. Learning a global representation for re-ID

We now describe the design of our deep architecture and our strategy for effectively training it for person re-ID.

3.1. Architecture design

The architecture of our image representation model in most ways resembles that of standard deep image recognition models. However, it incorporates several important modifications that proved beneficial for image retrieval tasks [12, 31]. The model contains a backbone convolutional network, pre-trained for image classification, which is used to extract local activation features from input images of an arbitrary size and aspect ratio. These local features are then max-pooled into a single vector, fed to a fully-connected layer and ℓ_2-normalized, producing a compact vector whose dimension is independent of the image size. Figure 2 illustrates these different components and identifies the design choices (#1 to #4) that we evaluate in the experimental section (Section 4.2).

Different backbone convolutional neural networks, such as ResNet [17], ResNeXt [44], Inception [40] and Densenet [18] can be used interchangeably in our architecture. In Section 4.2, we present results using several flavors of ResNet [17], and show the influence of the number of convolutional layers on the accuracy of our trained model.

3.2. Architecture training

A key aspect of the previously described representation is that all the operations are differentiable. Therefore, all the network weights (*i.e.* from both convolutional and fully-connected layers) can be learned in an end-to-end manner.

Three-stream Siamese architecture. To train our representation end-to-end we use a three-stream Siamese architecture in which the weights are shared between all streams. This learning approach has been successfully used for person re-identification [10, 38, 17] as well as for different retrieval tasks [12, 31]. Since the weights of the convolutional layers and the fully-connected layer are independent of the size of the input image, this Siamese architecture can process images of any size and aspect ratio. The three-stream architecture takes image triplets as input, where each triplet contains a query image I_q, a positive image I^+ (*i.e.* an image of the same person as in the query image), and a negative image I^- (*i.e.* an image of a different person). Each stream produces a compact representation for each image in the triplet, leading to the descriptors q, d^+ and d^- respectively. We then define the ranking triplet loss as

$$L(I_q, I^+, I^-) = \max(0, m + q^T d^- - q^T d^+), \quad (1)$$

where m is a scalar that controls the margin. This loss ensures that the embedding of the positive image I^+ is closer to the query image embedding I_q than that of the negative image I^-, by at least a margin m.

We now discuss key practices for improved training of our model.

Image size. Typically, training images are processed in

Good practices for person re-ID

- Pre-training for identity classification
- Sufficiently large image resolution
- State-of-the-art base architecture
- Hard triplet mining
- Dataset augmentation with difficult examples

Figure 3. Summary of good practices for building a powerful representation for person re-identification.

batches and therefore resized to a fixed input size, which leads to distortions. We argue that images should be upscaled to increase the input image size, and that they should not be distorted. To this end, we process triplets sequentially, allowing a different input size for each image and allowing the use of high resolutions images even in the most memory hungry architectures (e.g. ResNet-152 or Densenet). To account for the reduced batch size, we accumulate the gradients of the loss with respect to the parameters of the network for every triplet, and only update the network once we achieve the desired effective batch size.

Pretraining. We found it crucial to use pre-trained models with our architecture. First, we follow standard practice and use networks pre-trained on ImageNet [9]. To achieve the highest performance, it was also quite important to perform an additional pre-training step by fine-tuning the model on the training set using a classification loss, *i.e*. to train the model for person identification. We discuss this further in Section 3.3 and in the ablative study in Section 4.2.

Data augmentation.

To augment the dataset we adopt an image "cut-out" strategy, which consists of adding random noise to random-sized regions of the image. We progressively increase the maximum size of these regions during training, progressively producing more difficult examples. This strategy improves the results because it serves two purposes: it is a data augmentation scheme that directly targets robustness to occlusion and it allows for model regularization by acting as a "drop-out" mechanism at the image level. As a result, this strategy avoids the over-fitting inherent to the small size of the training set and significantly improves the results. We also considered standard augmentation strategies such as image flipping and cropping [36] but found no added improvement, as we show in Section 4.2.

Hard Triplet Mining. Finally, mining *hard* triplets is crucial for learning. As already argued in [17, 15, 42], when applied naively, training with a triplet loss can lead to underwhelming results. Here we follow the hard triplet mining strategy introduced in [13]. First, we extract the features for a set of N randomly selected examples using the current model and compute the loss of all possible triplets. Then, to select triplets, we randomly select an image as a query and randomly pick a triplet for that query from among the 25 triplets with the largest loss. To accelerate the process, we only extract a new set of random examples after the model has been updated k times with the desired batch size b. This is a simple and effective strategy which yields good model convergence and final accuracy, although other hard triplet mining strategies [15, 42] could also be considered.

3.3. Curriculum learning for re-ID

Similarly to humans, who learn a set of concepts more easily when the concepts to be learned are presented by increasing degree of complexity, it has been shown that curriculum learning has a positive impact on the speed and quality of the convergence of deep neural networks [3]. We adopt this learning strategy in our approach. In particular, three of our design principles described in this section aim to progressively increase the difficulty of the task being learned by our model. First, our hard-negative mining strategy samples triplets that increase in difficulty as learning continues. Second, our pre-training strategy first trains our model to solve the task of person ID classification (which requires the model to first recognize individuals within a closed set of possible IDs) before training it for the more challenging task of re-identifying persons. Third we observed that when training with cut-out, we achieve best results when the percentage of the image that is replaced by noise progressively increases. We believe that this general training principle is crucial to our results (reported in Section 4.3).

Figure 3 summarizes the good practices that we propose for both designing and training a deep architecture for person re-identification.

4. Experiments

4.1. Experimental details

Datasets. We consider four datasets for evaluation.

The **Market-1501** dataset [53] (Market) is a standard person re-ID benchmark with images from 6 cameras of different resolutions. DPM detections [11] were annotated as containing one of the 1,501 identities, among which 751 are used for training and 750 for testing. The training set contains 12,936 images with 3,368 query images. The gallery set is composed of images from the 750 test identities and of distractor images, 19,732 images in total. There are two possible evaluation scenarios for this database, one using a single query image and one with multiple query images.

The **MARS** dataset [51] is an extension of Market that targets the retrieval of gallery tracklets (*i.e*. sequences of images) rather than individual images. It contains 1,261 identities, divided into a training (631 IDs) and a test (630 IDs) set. The total number of images is 1,067,516, among which 518,000 are used for training and the remainder for testing.

The **DukeMTMC-reID** dataset [56] (Duke) was created

flip	crop	cut-out	Market	Duke
-	-	-	75.9	69.6
✓	-	-	77.2	69.7
-	✓	-	76.8	69.4
-	-	✓	**81.2**	**72.9**
✓	✓	✓	**81.2**	**72.9**

Table 1. **Impact of different data augmentation strategies**. We report mean average precision (mAP) on Market and Duke.

Largest dimension	Market	Duke
256 pixels	78.2	69.2
416 pixels	81.2	72.9
640 pixels	81.2	73.1

Table 2. **Impact of the input image size**. We report mean average precision (mAP) on Market and Duke.

by manually annotating pedestrian bounding boxes every 120 frames of the videos from 8 cameras of the original DukeMTMC dataset [34]. It contains 16,522 images of 702 identities in the training set, and 702 identities, 2,228 query and 17,661 gallery images in the test set.

The **Person Search** dataset [43] (PS) differs from the previous three as it was created from images collected by hand-held cameras and frames from movies and TV dramas. It can therefore be used to evaluate person re-identification in a setting that doesn't involve a known camera network. It contains 18,184 images of 8,432 identities, among which 5,532 identities and 11,206 images are used for training, and 2,900 identities and 6,978 images are used for testing.

Evaluation. We follow standard procedure for all datasets and report the mean average precision over all queries (mAP) and the cumulative matching curve (CMC) at rank-1 and rank-5 using the evaluation codes provided.

Training details. As mentioned in Section 3.1, for the convolutional part of our network we evaluate different flavors of ResNet [16], concretely ResNet-50, ResNet-101 and ResNet-152 (we study their impact in the following section). For all of them, we start with the publicly available pre-trained model on ImageNet, and fine-tune the weights of the convolutional layers for person identification in the training set of the specific dataset. To do this, we follow standard practice and extract random-sized crops and then resize them to 224×224 pixels. We train with stochastic gradient descent (SGD) with momentum of 0.9, weight decay of $5 \cdot 10^{-5}$, a batch size of 128, and an initial learning rate of 10^{-2}, which we decrease to 10^{-4}. We use the weights of this pre-trained network for the convolutional layers of our architecture and we randomly initialize the fully-connected layer, whose output we set to 2,048 dimensions. We then train the ranking network using our Siamese architecture with input images of variable size, while fixing the largest side to M pixels (whose influence we also study in the following section). We use again SGD with a batch size of 64 and an initial learning rate of 10^{-3}, which we decrease using a logarithmic rate that halves the learning rate every 512 iterations. We observe in all our experiments that the model converges after approximately 4,096 iterations. For the hard triplet mining we set the number of random examples to $N = 5,000$ and the number of updates to $k = 16$. We set the margin of the triplet loss to $m = 0.1$. Exactly the same training settings were used across all four datasets.

		Market	Duke
a) pooling strategy	average	80.1	71.4
	max	**81.2**	**72.9**
b) backbone architecture	ResNet-50	76.3	67.6
	ResNet-101	81.2	72.9
	ResNet-152	**81.4**	**74.0**
c) pretraining for class.	no	77.1	71.1
	yes	**81.2**	**72.9**

Table 3. Top (a): influence of the **pooling strategy**. Middle (b): results for different **backbone architectures**. Bottom (c): influence of **pretraining the network for classification** before considering the triplet loss. We report mAP for Market and Duke.

4.2. Ablative study

In this section we evaluate key design choices in our architecture and training strategy that relate to the good practices we propose in Figure 3.

Image transformation. We first focus on data augmentation (#2 in Figure 2). As discussed in Section 3, we apply different transformations to the images at training time, namely flips, crops and cut-outs. Here we study how each transformation impacts the final results, reported in Table 1. We observe that cut-out has a very strong impact on the performance and renders the other two data augmentation schemes superfluous. We believe that this is because cut-out makes our representation much more robust to occlusion, and also avoids over-fitting on such little training data.

Second, we consider the impact of the size of the input image (#1). Images from the Market dataset have a fixed size of 256×128, while images from Duke have a variable size, with 256×128 pixels on average. In our experiments, we rescale images so that the largest image dimension is either 256, 416, or 640 pixels, without distorting the aspect ratio. We report results in Table 2 and observe that using a sufficiently large resolution is key to achieving the best performance. Increasing the resolution from 256 to 416 improves mAP by 3%, while increasing it further to 640 pixels shows negligible improvement. We set the input size to 416 pixels for the rest of this paper.

Pooling. Table 3 (a) compares two pooling strategies (#4) over the feature map produced by the convolutional layers. As we see that max pooling performs better than average

	Type	Market-1501 SQ			Market-1501 MQ			MARS			Duke-reID			PS
		mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP
MG [41]	-	39.6	65.9	-	48.4	76.0	-	-	-	-	-	-	-	-
CRAFT [7]	-	45.5	71.8	-	54.3	79.7	-	-	-	-	-	-	-	-
SpindleNet [49]	P	-	76.9	91.5	-	-	-	-	-	-	-	-	-	-
Zheng *et al*. [51]	-	-	-	-	-	-	-	49.3	68.3	82.6	-	-	-	-
Part-Aligned [50]	A	63.4	81.0	92.0	-	-	-	-	-	-	-	-	-	-
OL-MANS [58]	-	-	60.7	-	-	66.8	-	-	-	-	-	-	-	-
HydraPlus-Net [25]	A	-	76.9	91.3	-	-	-	-	-	-	-	-	-	-
MSCAN [21]	P	57.5	80.3	-	66.7	86.8	-	66.4	**83.0**	**93.7**	-	-	-	-
OIM [43]	-	-	82.1	-	-	-	-	-	-	-	-	68.1	-	**77.9**
PDC [37]	P	63.4	84.1	92.7	-	-	-	-	-	-	-	-	-	-
Verif-Identif. [54]	-	59.9	79.5	-	70.3	85.8	-	-	-	-	49.3	68.9	-	-
LSRO [56]	-	66.1	84.0	-	76.1	88.4	-	-	-	-	47.1	67.7	-	-
SVDNet [39]	-	62.1	82.3	92.3	-	-	-	-	-	-	56.8	76.7	**86.4**	-
SSM [1]	-	68.8	82.2	-	76.2	88.2	-	-	-	-	-	-	-	-
DPFL [6]	-	**73.1**	**88.9**	92.3	-	-	-	-	-	-	**60.6**	**79.2**	-	-
DML[48]*	-	68.8	87.7	-	**77.1**	**91.7**	-	-	-	-	-	-	-	-
APR [46]*	At	64.7	84.3	93.2	-	-	-	-	-	-	51.9	70.7	-	-
PAN [55]*	A	63.3	82.8	93.5	-	-	-	-	-	-	51.5	71.6	83.9	-
PBF [52]*	P	56.0	79.3	90.8	-	-	-	-	-	-	-	-	-	-
TriNet [17]*	-	69.1	84.9	**94.2**	76.4	90.5	**96.3**	**67.7**	79.8	91.4	-	-	-	-
Ours	-	**81.2**	**92.2**	**97.9**	**87.3**	**94.7**	**98.6**	**79.7**	**85.8**	**96.5**	**72.8**	**85.2**	**93.9**	**92.6**
Improvement	-	+8.1	+3.3	+3.7	+10.2	+3.0	+2.3	+12.0	+2.8	+2.8	+12.2	+6.0	+7.5	+14.7
Re-ranking [57]	-	63.6	77.1	-	-	-	-	68.4	73.9	-	-	-	-	-
TriNet (re-rank) [17]*	-	81.1	86.7	93.4	87.2	91.8	95.8	77.4	81.2	90.8	-	-	-	-
Ours (re-rank)	-	**90.0**	**93.0**	**95.9**	**91.2**	**94.2**	**96.9**	**85.7**	**87.2**	**94.9**	**85.6**	**89.4**	**93.6**	-

Table 4. Comparison with state of the art methods on the Market-1501, MARS, Duke-reID and Person Search datasets. The "Type" column indicates methods that include the following additional components: a part-based representation (P) with extra annotations, an attention mechanism (A), or attribute annotations at train time (At). Bold numbers show the current state of the art, while red numbers correspond to the best number overall. * indicates methods published only in *arXiv*.

pooling on both datasets, we use it for the rest of this paper.

Backbone architecture. Table 3 (b) compares different architectures for the convolutional backbone of our network (#3). Results show that using ResNet-101 significantly improves the results compared with using ResNet-50 (about +5 mAP for both datasets). The more memory hungry ResNet-152 only marginally improves the results.

Fine-tuning for classification. Table 3 (c) shows the importance of fine-tuning the convolutional layers for the identity classification task before using the ranking loss to adjust the weights of the whole network (#6). As discussed in Section 3.3, training the model on tasks of increasing difficulty is highly beneficial.

4.3. Comparison with the state of the art

Table 4 compares our approach to the state of the art. Our method consistently outperforms all methods by large margins on all 4 re-ID datasets and all metrics. In particular, we achieve a mAP of 81.2% on Market, an 8.1% absolute improvement compared with the best published results [6]. We also outperform [17] by 12.0% mAP on MARS. On the Duke dataset, we achieve a mAP of 72.8%, outperforming the previous best reported mAP [6] by 12.2%. It is also important to note that our approach using ResNet-50, reported in Table 3 b), still outperforms prior art by a significant margin, showing that all of our design choices play a crucial role, not only the backbone architecture. We also report the performance of our method with standard re-ranking[2] and we again see large improvements with respect to prior art that uses re-ranking, across all datasets and metrics. For example, for Market, we achieve a mAP of 90%, 8.9% above the best previously-reported mAP from [17].

Looking closely at the approaches that report results on these datasets, we first note that our approach outperforms all recent methods that also use a variant of the triplet loss and hard triplet mining [17, 50]. As we show in this section, combining these key principles with the others mentioned in Figure 3 is crucial for effective training of our image repre-

[2]We expand both the query and the dataset by averaging the representation of the first 5 and 10 closest neighbors, respectively.

sentation for Re-ID. It is also worth emphasizing that our approach even outperforms recent works that propose complex models for aligning images based on attributes [46] or body parts via pose estimation [52], part detection [21, 49] or attention modules [50], most of which require extra resources such as annotations or pre-trained detectors. As we discuss in the next section, our model is able to discriminate body regions without such additional architectural modules.

We also report results for the Person Search dataset in last column of Table 4. This dataset differs from traditional re-ID datasets in that the different views of each person do not correspond to different cameras in a network. Nevertheless, our approach performs quite well in this different scenario, achieving a mAP of 92.6%, which is a 14.7% absolute improvement over the previous best reported result [43]. This shows the generality of our approach.

4.4. Qualitative analysis

In this section we perform a detailed analysis of our trained model's performance and induction biases.

Re-identification examples. In Figure 4, we show good results (top) and failure cases (bottom) for several query images from the Market dataset. We see that our method is able to correctly re-identify persons despite pose changes or strong scale variations. We observe that failure cases are mostly due to confusions between two people that are extremely difficult to differentiate even for a human annotator, or to unusual settings (for instance the person holding a backpack in front of him as in e.).

Localized responses and clothing landmark detection. In Section 3, we argued that, using our proposed approach, we obtain an embedding that captures invariance properties useful for re-ID. To qualitatively analyze this invariance, we use Grad-Cam [35], a method for highlighting the discriminative regions that CNN-based models activate to predict visual concepts. This is done by using the gradients of these concepts flowing into the final convolutional layer. Similar to [14], given two images, we select the 5 dimensions that contribute the most to the dot-product similarity between their representations. Then, for each image, we propagate the gradients of these 5 dimensions individually, and visualize their activations in the last convolutional layer of our architecture. In Figure 5, we show several image pairs and their respective activations for the top 5 dimensions.

We first note that each of these output dimensions are activated by fairly *localized image regions* and that the dimensions often reinforce one-another in that image pairs are often activated by the same region. This suggests that the similarity score is strongly influenced by localized image content. Interestingly, these localized regions tend to contain body regions that can inform on the type of clothing being worn. Examples in the figure include focus on the hem of a pair of shorts, the collar of a shirt, and the edge of a sleeve. Therefore, rather than focusing on aligning human body joints, the model appears to make decisions based on *attributes of clothing* such as the length of a pair of pants or of a shirt's sleeves. This type of information has been leveraged explicitly for retrieval using the idea of "fashion landmarks", as described in [26]. Finally, we observe that some of the paired responses go *beyond appearance similarity* and respond to each other at a more abstract and semantic level. For instance, in the top right pair the strong response of the first dimension to the bag in the first image seems to pair with the response to the strap of the bag in the second image, the bag itself being occluded (see also the backpack response of Figure 1 as an other example).

Figure 4. For several queries from Market, we show the first 10 retrieved images together with the mAP and the number of relevant images (in brackets) of that query. Green (resp. red) outlines images that are relevant (resp. non-relevant) to the query.

Implicit attention. We now qualitatively examine which parts of the images are highly influential, independently of the images they are matched with. To do so, given an image and its embedding, we select the first 50 dimensions with the strongest activations. We then propagate and accumulate the gradients of these dimensions, again using Grad-Cam [35], and visualize their activations in the last convolutional layer in our architecture. As a result, we obtain a visualization that highlights parts of the images that, *a priori*, will have the most impact on the final results. This can be seen as a visualization of the implicit attention mechanism that is at play in our learned embedding.

We show such *implicit attention masks* in Figure 6 across several images of the same person, for three different persons. We first observe that the model attends to regions known to drive attention in human vision, such as high-resolution text [4] (e.g. in rows 1 and 2). We also note that our model shows properties of contextual attention, particularly when image regions become occluded. For example,

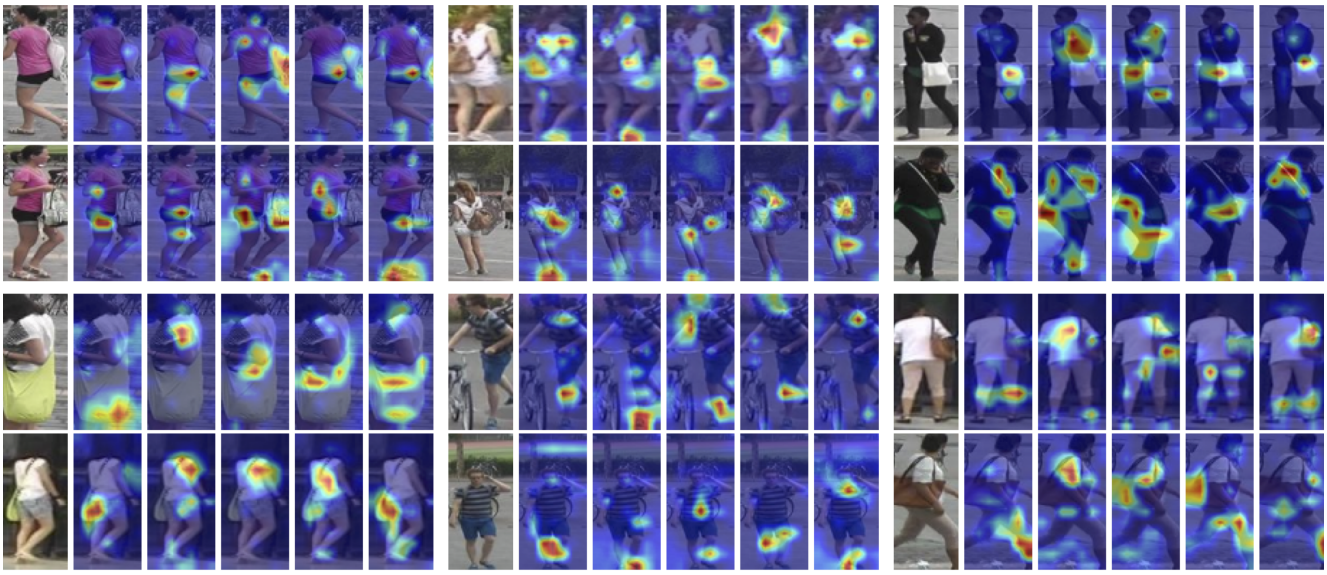

Figure 5. **Matching regions** For pairs of matching images, we show maps for the top 5 dimensions that contribute most to the similarity.

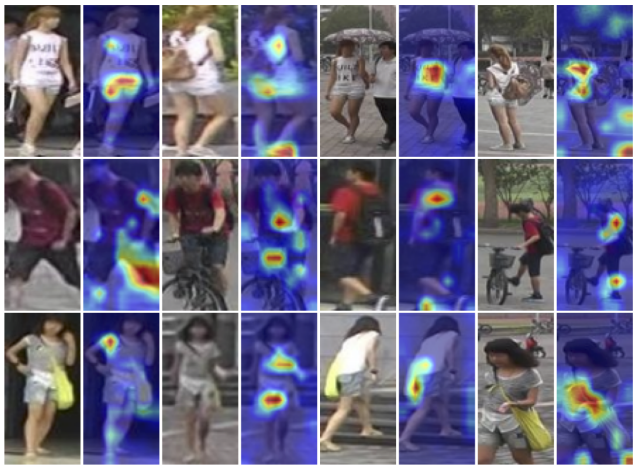

Figure 6. **Implicit attention** We highlight regions that correspond to the most highly-activated dimensions of the final descriptor. They focus on unique attributes, such as backpacks, bags, or shoes.

when the man in the second row faces the camera, text on his t-shirt and the hem of his pants are attended to. However, when his back or side is to the camera, the model focuses more intently on the straps of his backpack.

4.5. Re-ID in the presence of noise

To test the robustness of our model, we evaluate it in the presence of noise using Market+500K [53], an extension of the Market dataset that contains an additional set of 500K distractors. To generate these distractors, the authors first collected ground-truth bounding boxes for persons in the images. They then computed the IoU between each predicted bounding box and ground-truth bounding box for a given image. A detection was labeled a distractor if its IoU with all ground-truth annotations was lower than 20%.

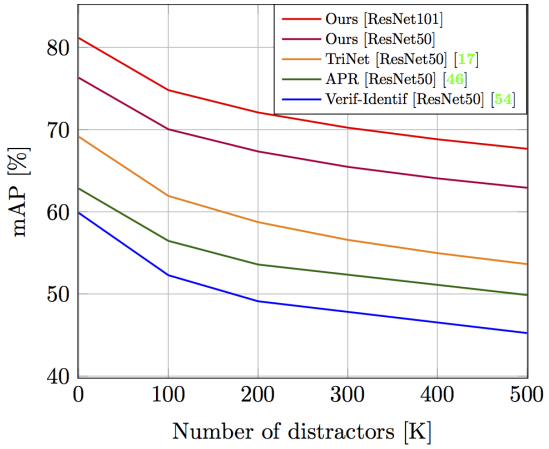

Figure 7. Performance comparison in the presence of distractors.

We evaluate our ResNet-50- and ResNet-100-based models, trained on Market, on this expanded dataset, while increasing the number of distractors from 0 to 500K. We selected distractors by randomly choosing them from the distractor set and adding them to the gallery set. Both models significantly outperform the current state-of-the-art results in the presence of this noise, as presented in Figure 7. Note that our best model, with 500K added distractors, performs on par with [17]'s performance with 0 added distractors.

5. Conclusions

In this paper, we have proposed a set of good practices for designing and training an efficient and effective image representation model for the task of person re-identification.

We showed through extensive experiments that our model outperforms all state-of-the-art approaches for this task by large margins, across four datasets and three metrics. We believe that our proposed approach can serve as a useful baseline for future contributions to the field.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *Proc. CVPR*, 2017. 2, 6

[2] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 2014. 1

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. ICML*, 2009. 4

[4] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 2009. 7

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *Proc. CVPR*, 2017. 1, 2

[6] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *Proc. ICCV Workshop*, 2017. 3, 6

[7] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *PAMI*, 2017. 6

[8] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proc. CVPR*, 2016. 1

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR*, 2009. 4

[10] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 2015. 2, 3

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 4

[12] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. ECCV*, 2016. 3

[13] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 2017. 4

[14] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proc. CVPR*, 2017. 7

[15] B. Harwood, V. Kumar B G, G. Carneiro, I. Reid, and T. Drummond. Smart mining for deep metric learning. In *Proc. ICCV*, 2017. 4

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1, 5

[17] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 1, 2, 3, 4, 6, 8

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, 2017. 3

[19] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv*, 2016. 1, 2

[20] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, 2012. 2

[21] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proc. CVPR*, 2017. 3, 6, 7

[22] W. Lin, Y. Shen, J. Yan, M. Xu, J. Wu, J. Wang, and K. Lu. Learning correspondence structures for person re-identification. *TIP*, 2017. 3

[23] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. In *Proc. ICCV*, 2017. 3

[24] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017. 2

[25] 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 *Proc. ICCV*, 2017. 2, 6

[26] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR*, 2016. 7

[27] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Proc. CVPR*, 2012. 2

[28] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proc. CVPR*, 2015. 2

[29] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proc. CVPR*, 2013. 2

[30] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *Proc. ICCV*, 2017. 3

[31] F. Radenovic, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016. 3

[32] T. Rahman, M. Rochan, and Y. Wang. Person re-identification by localizing discriminative regions. In *Proc. BMVC*, 2017. 2

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NIPS*, 2015. 1

[34] 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. 5

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. CVPR*, 2017. 7

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 4

[37] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proc. ICCV*, 2017. 1, 2, 6

[38] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *Proc. ECCV*, 2016. 1, 2, 3

[39] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *Proc. ICCV*, 2017. 2, 6

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proc. CVPR*, 2016. 3

[41] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Proc. ECCV*, 2016. 2, 6

[42] C.-T. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Sampling matters in deep embedding learning. In *Proc. ICCV*, 2017. 4

[43] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. CVPR*, 2017. 3, 5, 6, 7

[44] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*, 2017. 3

[45] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Proc. ECCV*, 2014. 2

[46] L. Yutian, Z. Liang, Z. Zhedong, W. Yu, and Y. Yi. Improving person re-identification by attribute and identity learning. *arXiv*, 2017. 2, 6, 7

[47] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proc. CVPR*, 2016. 2

[48] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *arXiv*, 2017. 6

[49] 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 *Proc. CVPR*, 2017. 1, 2, 6, 7

[50] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proc. ICCV*, 2017. 2, 6, 7

[51] 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 *Proc. ECCV*, 2016. 4, 6

[52] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv*, 2017. 2, 6, 7

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proc. ICCV*, 2015. 4, 8

[54] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *TOMM*, 2017. 6

[55] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv*, 2017. 2, 6

[56] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. ICCV*, 2017. 3, 4, 6

[57] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proc. CVPR*, 2017. 3, 6

[58] J. Zhou, P. Yu, W. Tang, and Y. Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *Proc. ICCV*, 2017. 2, 6

[59] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *Proc. CVPR*, 2017. 2

[60] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. X. Zheng. Deep hybrid similarity learning for person re-identification. *Trans. CSVT*, 2017. 2