Person Re-Identification by Deep Learning Attribute-Complementary Information

Arne Schumann
Fraunhofer IOSB
76131, Karlsruhe
arne.schumann@iosb.fraunhofer.de

Rainer Stiefelhagen
Karlsruhe Institute of Technology
76131, Karlsruhe
rainer.stiefelhagen@kit.edu

Abstract

Automatic person re-identification (re-id) across camera boundaries is a challenging problem. Approaches have to be robust against many factors which influence the visual appearance of a person but are not relevant to the person's identity. Examples for such factors are pose, camera angles, and lighting conditions. Person attributes are a semantic high level information which is invariant across many such influences and contain information which is often highly relevant to a person's identity. In this work we develop a re-id approach which leverages the information contained in automatically detected attributes. We train an attribute classifier on separate data and include its responses into the training process of our person re-id model which is based on convolutional neural networks (CNNs). This allows us to learn a person representation which contains information complementary to that contained within the attributes. Our approach is able to identify attributes which perform most reliably for re-id and focus on them accordingly. We demonstrate the performance improvement gained through use of the attribute information on multiple large-scale datasets and report insights into which attributes are most relevant for person re-id.

1. Introduction

The problem of re-identifying a person across cameras has gained increasing research interest in recent years [36]. It facilitates multi-camera person tracking and its many applications include surveillance of secure areas, public events, customer analysis, or enrichment of multimedia content, such as movies or TV series. The main challenge of person re-id approaches is to develop a person representation which is robust to influences, such as pose, camera characteristics and angle, or lighting, while remaining discriminative enough to be able to differentiate between different individuals. Many works on the subject focus on either learning a powerful feature representation of persons [2, 9, 4] or a distance measure which robustly matches person representations across camera views [10, 14, 18]. Moreover, in recent years, person attributes have been identified as an additional, powerful information to aid in the re-id task [15, 24, 11]. Attributes represent a high level semantic description of a person and are invariant to many influences, such as pose and camera angles. They often contain meaningful information which can be very localized in the image and is easily missed by approaches which focus on global appearance. A number of examples in which attribute information can help distinguish between persons whose visual appearance is otherwise very similar is depicted in Figure 1. Attributes can also be more easily communicated to human security personnel and enable a broader range of applications, such as purely textual search queries. In practical ap-

Figure 1. Examples of persons with very similar global appearance but different attributes. Attribute information can be a decisive clue to distinguish people in such cases.

plications an important challenge is how to best make use of automatically generated attribute classifications which contain a varying degree of accuracy.

In this work, we propose a deep learning approach which includes such attribute information into the learning process of a CNN. Image features based on CNNs have recently proven very successful for person re-id [17, 1, 3, 28]. However, the resulting features focus on global appearance and lack interpretable semantic content which attributes can provide. Our approach aims to leverage this complementary information in attributes and CNN features. An overview of the approach is depicted in Figure 2. We begin by training an attribute model on data separate from our target re-id datasets. The attribute predictions generated by this model is then included into our re-id approach through a triplet loss adapted specifically for this purpose. The loss allows the CNN to learn a person representation which is complementary to that of the attributes. We cannot assume the attribute predictions to be highly reliable due to possible biases on their training data. Furthermore, we cannot directly measure the accuracy of attribute predictions, because we generally do not have corresponding annotations on the target datasets. We overcome these problems by adapting our CNN architecture to automatically learn a weight for each attribute which controls the attribute's impact on the re-id task. The resulting weighting notably improves re-id accuracy compared to the direct use of unweighted attributes and allows us to draw conclusions on the reliability of our classifier with respect to individual attributes on the target datasets. We perform evaluations on multiple large-scale re-id datasets. We investigate the re-id accuracy of baselines which rely solely on either attributes or CNNs and show that our combined approach outperforms either one. Our approach achieves competitive or state-of-the-art performance on all evaluated datasets.

Our main contributions are summarized as follows: **(1)** We propose a novel approach which integrates semantic information of attributes into the learning process of a CNN trained for person re-id. **(2)** We show that attributes improve re-id accuracy on multiple large-scale public datasets and result in a strong performance of our combined model. **(3)** Our evaluations yield additional insights into which attributes are best suited to aid in person re-id.

The remainder of this work is organized as follows. In Section 2 we discuss related approaches on attribute classification, person re-id, and their combination. We describe our attribute classifier in Section 3. Our own attribute-based re-id approach is outlined in Section 4. The influences of attributes on the final re-id accuracy are evaluated in Section 5 and we summarize our findings in Section 6.

2. Related Work

We focus our discussion of related work on the most relevant recent approaches which rely on CNN features, attributes, or a combination thereof.

Many recent re-id approaches rely on CNNs and deep learning [30, 34, 38, 37]. [30] uses domain guided dropout to learn a person representation across multiple datasets. In [37] a feature embedding is learned with the help of an additional verification loss. Zheng et al. [34] train a pose invariant person representation by normalizing input images according to person pose predictions. In [38] generative adversarial networks are used to generate additional training samples.

Often, siamese or triplet loss networks are used to learn based on comparison between matching and mismatching person images [6, 17, 1, 3, 28, 29]. Li et al. [17] use a filter pairing architecture to match persons in a CNN. A special neighborhood matching layer was introduced by Ahmed et al. [1]. In [3] Cheng et al. use a triplet loss and a simple body part segmentation to learn a robust feature embedding. In [28] a gated siamese network is used to focus the comparison of person images on relevant regions. In a later work [29] LSTMs are used to guide the attention of the network.

Attributes have been used in person re-id for some time [15, 24, 11]. A few recent works focus specifically on combining attribute information with CNNs. Khamis et al. [13] use a triplet loss architecture for re-id in combination with an attribute loss and leverage multiple data sources. In [22] fine tune CNNs for attribute recognition and employ metric learning for subsequent person re-id. Recently, Lin et al. [20] used a combination of re-id and attribute classification losses to learn a joint representation for person re-id.

3. Multiview Attribute Detection

Person attributes can be grouped into two categories. Some attributes, such as *long pants* or *backpack*, are localized and thus only visible in certain regions of an image while others, such as *age* or *gender*, are global attributes which cannot be assigned to a single specific image region. Based on this observation we propose a CNN architecture for attribute classification which combines local and global image information. An overview is given in the left part of Figure 2.

We base our network architecture on a GoogleNet which is pretrained on ImageNet [27]. We use the GoogleNet to capture global information at image level. We connect the output of the inception_5b layer to a fully connected layer of size 600 which yields our global person representation.

In order to include local information relevant to specific attributes into the network we divide the conv1_7x7 layer of GoogleNet into three equal horizontal regions. Each of these regions represents a local view on part of the im-

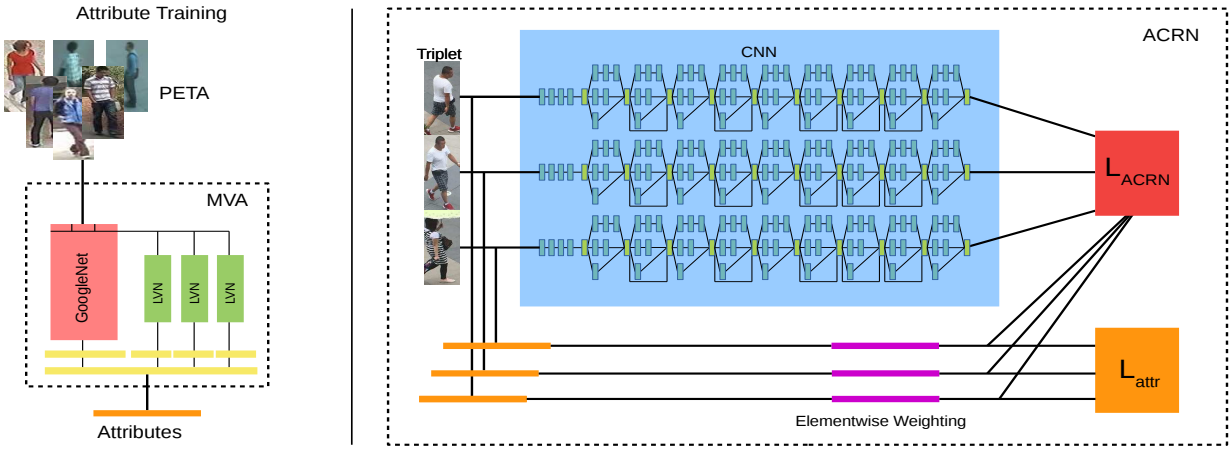

Figure 2. Overview of our approach. An attribute representation is pretrained (on PETA). Attributes are then used to support the learning of a CNN feature through a triplet loss L_{ACRN}. An additional triplet loss L_{attr} learns weights for attributes in order to reduce the influence of unreliable attributes on the re-id result.

age. For each view we train a small sub-network which can learn region-specific information. The local view networks (LVN) consist of two further convolutional layers followed by two residual blocks which each perform a pooling step. The final element of each LVN is a fully connected layer of size 200. Both, the GoogleNet and the LVNs, employ batch normalization [12].

The global and local information generated by the fine-tuned GoogleNet and the LVNs are combined through a final layer of size 2048. This combined representation of multi view attribute information (MVA) is then used for attribute classification. We employ a single multi-class cross-entropy loss instead of one loss for each attribute. However, such a multi-label classification approach suffers from imbalances in the training data. Some attributes are much more frequent than others and we cannot compensate by data sampling, because attributes co-occur and balancing the occurrence frequency of one attribute will change that of others. To address this we follow the approach of Yu et al. [31, 16] of weighting the attributes in the loss:

$$L_{wce} = \sum_{i=1}^{L} \frac{1}{2w_i} * p_i * log(q_i) + \frac{1}{2(1-w_i)}(1-p_i) * log(1-q_i) \quad (1)$$

where w_i is a weight specific to attribute i which reflects its relative frequency in the training data (i.e. its ratio of positive to negative labels), p_i is the attribute prediction and q_i is the corresponding label.

We train the network on the PETA dataset [5] collection. The GoogleNet is fine-tuned for attribute classification using an initial learning rate of 10^{-4} which is decreased stepwise by a factor of 0.1. The weights of the LVNs are initialized randomly and their learning rate is set to a higher value. For training we use a batchsize of 64.

4. Attribute-Complementary Re-id Net

In this section we detail the architecture and training process of our attribute-complementary re-id net (ACRN).

Based on recent developments of CNN architectures [26] we base our network on a combination of inception layers and residual connections. Our input image size is 160×64. We start by a basic feature extraction through 4 convolutional layers of size 3×3 and pooling. This is followed by a normal inception v3 layer to increase the channel count for subsequent residual connections. The network then consists of eight inception layers. Layers 2, 4 and 8 of those perform a pooling step. The others are bridged by residual connections. After the final pooling we attach a fully connected layer of size 512 which serves as our feature representation for re-id.

We use a triplet network architecture. Training samples are served to the network in sets of three: one anchor image, one match to the anchor (positive) and one mismatch (negative). Such architectures are trained with a classic triplet loss:

$$\begin{aligned} L_{triplet} &= \frac{1}{N}\sum_{i=1}^{N} d_i^{f^p} - d_i^{f^n} + m \\ d_i^{f^p} &= \|f_i^a - f_i^p\|_2^2 \\ d_i^{f^n} &= \|f_i^a - f_i^n\|_2^2 . \end{aligned} \quad (2)$$

The term $d_i^{f^p}$ denotes the feature distance from the anchor to the positive and $d_i^{f^n}$ the distance from the anchor to the negative. The loss encourages the feature distances between

images of different persons to be large and distances for matching persons to be small. The margin m controls the severity of the separation.

We incorporate the attribute information generated by our attribute net into the triplet training process at loss level. This is achieved by adding the attribute distances of the anchor sample to the positive and negative samples in an analogous manner:

$$\begin{aligned} L_{ACRN} =& \frac{1}{N}\sum_{i=1}^{N} d_i^{f^p} - d_i^{f^n} + m + \gamma(d_i^{att^p} - d_i^{att^n}) \\ d_i^{att^p} =& \|att_i^a - att_i^p\|_2^2 \\ d_i^{att^n} =& \|att_i^a - att_i^n\|_2^2 \end{aligned} \tag{3}$$

where $d_i^{att^p}$ and $d_i^{att^n}$ mark the distances of the samples in triplet i based on their attribute representations. The addition of this attribute information has no direct impact on the triplet losses' gradient formulas. Take, for example, the gradient for the positive sample in the triplet:

$$\begin{aligned} \frac{\partial L_i}{\partial f_i^p} =& \begin{cases} 2(f_i^a - f_i^p), & if\ d_i^{f^n} - d_i^{f^p} \leq \tilde{m} \\ 0, & otherwise \end{cases} \\ \tilde{m} =& m + \gamma(d_p^{a^i} - d_n^{a^i}). \end{aligned} \tag{4}$$

Instead, the attribute distance in combination with the margin m influence the cases in which the gradient is passed through the ACRN (i.e. when it is non-zero). If, for example, the attributes fail to produce a distance $d_i^{att^p} < d_i^{att^n}$, then the attribute part of the loss adds to the margin, the loss becomes more strict, and the gradient of the ACRN is more likely to be non-zero for this sample. Conversely, if the attribute information already achieves $d_i^{att^p} < d_i^{att^n}$, then the gradient is more likely to be zero. This allows the ACRN to focus on cases where the attribute information does not suffice for a successful re-id. In a sense the net learns to re-rank a basic ranking generated by the attributes. The approach is motivated by the assumption that this might be a simpler task than re-id without attribute information. We use a parameter γ to control the degree of influence of the attribute information on the ARCN loss. We empirically set γ to 0.5 which strikes a good balance between maintaining attribute information and not overpowering the base margin of the loss.

Because our attribute representation is learned on separate data, we can expect some of the attributes to be much less reliably predicted than others on our target re-id dataset. Inaccurate attributes will lead to more cases in which the attribute portion of our loss is incorrect. If those cases become too frequent, the ACRN net will have a more difficult task to solve than plain re-id. We thus introduce an additional layer for weighting the attributes on our target dataset. The weight layer simply performs an elementwise multiplication of our attributes with a learned weight. This weight layer is not connected to the input of the net and thus learns global weights for the whole dataset. We train these weights by an additional triplet loss L_{attr} which tries to separate the triplet samples based solely on weighted attribute information and ensures that attributes which are beneficial for re-id are weighted more strongly.

At test time we first use the attribute net to generate a set of attribute predictions. The ACRN is then used to compute our complementary deep re-id feature. Two person images can then be compared by simple addition of the ACRN feature distance and the weighted attribute distance.

We train our network on each target dataset an initial 10 epochs for person ID classification in order to establish a basic feature representation. We then switch to the ACRN setup and continue training using the triplet loss. Our initial learning rate is set to 0.01 and reduced by a factor of 0.8 after two epochs. We use a batchsize of 64 for the initial training and 40 for ACRN training.

5. Evaluation

We first evaluate the performance of our attribute classifier on its source dataset. Then we investigate the usefulness of these attribute predictions for re-id on a number of target datasets.

5.1. Attribute Net

Method	mA
ACN [25]	81.15%
DeepMAR [16]	82.89%
WPAL-GoogleNet-GMP [31]	**85.50%**
MVA (Ours)	84.61%

Table 1. Results of our approach for attribute recognition on PETA in mean accuracy.

We evaluate our attribute classification approach on the PETA pedestrian attribute dataset [5]. PETA is an attribute-labeled collection of pre-existing pedestrian datasets and contains 19,000 images. The images are annotated with 61 binary attributes and 4 multi-class attributes. For evaluation of our attribute net we follow the established protocol and use 11,600 images for training and validation as well as 7,600 images for testing. The images are randomly chosen. For comparison to related work we further focus on the 35 attributes which have a ratio of positive to negative labels of more than 5%. As evaluation metric we use the mean accuracy (mA) which is computed as the mean of the accuracy among positive samples and the accuracy among negative samples. We compare our approach to three recent works: ACN [25], DeepMAR [16], and WPAL-GoogleNet-GMP [31]. Results are shown in Table 1. Our attribute net has competitive accuracy but does not outperform the cur-

rent state-of-the-art [31]. In Table 2 we show some of our most and least reliably detected attributes on PETA.

Attribute	mA	Attribute	mA
AgeAbove61	92.8%	Hat	89.4%
Male	92.3%	Muffler	89.3%
Long Hair	90.5%	Short Sleeve	88.3%
Stripes	76.4%	Sandals	66.1%
Carry Other	76.1%	Sunglasses	65.3%
Logo	72.1%	V-neck	63.5%

Table 2. Attributes with the highest (top) and lowest (bottom) recognition rates of our approach on PETA.

5.2. Re-Id

We evaluate our ACRN approach to attribute based person re-identification on three large-scale public datasets.

CUHK3 The CUHK3 dataset [33] was recorded in a campus setting and consists of 5 pairs of camera views. The dataset contains 1,467 persons and more than 14,000 person images. CUHK3 does not contain any distractors (i.e. persons which appear only in the gallery). The bounding boxes are partly manually annotated and partly generated by the DPM detector [7]. We follow the provided evaluation protocol and set aside 100 persons for our test set while the remaining are used for training. We evaluate in a single-shot setting by randomly selecting one sample of each person as probe and another random sample from the opposite camera as gallery. Our CNN models are trained using the entire training set including the two camera pairs that are not present in the test data. At train time both labeled and detected data is used to train each model. We evaluate separately for both settings on the test set.

Market-1501 The Market-1501 dataset [35] provides 32,668 images of 1,501 persons. 751 persons are used for training. For testing a set of 3,368 query images is available. The gallery size of the Market-1501 dataset is 19,734 and contains 2,793 distractors.

DukeMTMC-reID The DukeMTMC-reID dataset [38] consists of persons cropped out of the DukeMTMC tracking dataset [23] which is recorded by 8 cameras. The dataset consists of 1,812 different persons of which 1,404 appear in more than one camera. 702 persons are set aside for the training set and the remaining 1,110 are used for testing. This results in a training set of 16,522 images, a probe set of 2,228 images and a gallery set of 17,661 images.

We evaluate our performance primarily based on mean average precision (mAP). mAP is computed as the mean AP over all queries in the test set. The AP of a query corresponds to the average of the precision scores at each rank where a correct result is returned. We further report Rank-1, -5, -10 and -20 accuracies to give an impression of the CMC. On Market-1501 and DukeMTMC-reID we use the provided evaluation code.

5.2.1 Baselines

We define four simple baseline methods to compare ACRN with.

ReIdCNN: This baseline trains a plain CNN model of the same architecture as ACRN but without any attribute information. We use the same training settings as for ACRN.

Attributes: For this baseline we evaluate the direct performance of the attribute scores generated by our multi view attribute net on the target re-id dataset. No learning on the target data is involved in this method.

Attributes-KISSME: This baseline applies KISSME metric learning [14] using our multi-view attribute predictions on the target data. This baseline shows to an extent the potential for re-id contained in the attribute predictions.

ReIdCNN+Attributes: In order to show the complementary nature of the information learned by ACRN we use this baseline which performs a simple score fusion between the ReIdCNN and Attributes baseline. Similar to ACRN we weight the attributes scores with 0.5 (the value of γ in ACRN).

Method	mAP	r1	r5	r10	r20
NullSpace [32]	-	54.70	84.75	94.80	95.20
Gated Siamese [28]	51.25	61.8	80.9	88.3	-
LSTM Siamese [29]	46.3	57.3	80.1	88.3	-
MLAPG [19]	-	57.96	87.09	94.74	98.00
PIE [34]	67.21	61.50	89.30	94.50	97.60
PIE+KISSME [34]	**71.32**	**67.10**	**92.20**	**96.60**	**98.10**
ReIdCNN	68.1	58.3	83.1	87.4	91.1
Attributes	10.3	12.5	28.9	37.6	41.1
Attributes-KISSME	21.1	34.3	45.3	53.2	59.1
ReIdCNN+Attributes	68.5	58.9	84.3	88.5	93.0
ACRN	70.2	62.63	89.69	94.72	97.12

Table 3. Results of our approach on the CUHK3 dataset (detected setting). We outperform most recent works with the exception of the Pose Invariant Alignment combined with metric learning.

The results of our baselines are given in Tables 3, 4, and 5 for CUHK3, Market-1501, and DukeMTMC-reID, respectively. On all three datasets similar trends can be observed. The ReIdCNN baseline performs strongly while pure attribute information can only achieve a very low person re-id accuracy. The main reasons for this are the comparatively low dimensionality of the attribute information and their presumably limited reliability due to varying performance of the attribute classifier. However, the application of KISSME metric learning to the attribute predictions on the target dataset shows that a higher potential for re-id is contained in the predictions. This indicates that certain attributes are helpful for re-id while others distort the result without metric learning. Finally, the combination of the

ReIdCNN with attribute information yields the best baseline performance but the result is dominated by the CNN and the overall improvement through attributes is very slight.

5.2.2 ACRN

When combining the attribute information with CNN features through our proposed ACRN network, another significant boost in re-id accuracy is achieved. Compared to the ReIdCNN baseline, the use of attribute information in ACRN can improve the resulting performance by 2.10% mAP on CUHK3 (detected), 3.76% mAP on Market-1501 and 2.54% mAP on DukeMTMC-reID. Compared to the ReIdCNN+Attributes baseline score fusion, the attribute information is much better used. Our ACRN net did not have to learn the information which is already contained in the attributes and thus could use more of its parameters for learning to compensate in failure cases. The ReIdCNN+Attributes baseline in contrary contains redundant information in the CNN which leads to reduced benefit of attributes.

We compare ACRN to a number of recent approaches. On CUHK3 (Table 3) we outperform most recent works, including Discriminative Nullspace [32], Gated Siamese Network [28], and the recent Pose Invariant Embedding (PIE) [34]. However, a combination of the PIE model based on ResNet50 with KISSME metric learning outperforms ACRN. We outperform PIE on the Market-1501 dataset (Table 4) as well as most other recent approaches. The Attribute Person-Recognition network (APR) outperforms our approach at top ranks. However, APR relies on additional training information in the form of attribute annotations (at person ID level) on the Market-1501 dataset itself. Finally, on the recent DukeMTMC-reID dataset we outperform all related approaches by a margin of 1.89% in Rank-1 accuracy. A qualitative impression of results by ACRN on our target datasets is given in Figure 3.

Method	mAP	r1	r5	r10	r20
Gated Siamese [28]	39.55	65.88	-	-	-
GAN [38]	55.95	79.33	-	-	-
PIE [34]	56.23	78.06	90.76	94.41	96.52
DeepTransfer [8]	**65.5**	83.7	-	-	-
APR [21]	64.67	**84.29**	**93.20**	95.19	**97.00**
ReIdCNN	58.84	79.51	90.40	92.21	96.5
Attributes	10.36	14.88	25.96	45.81	55.73
Attributes-KISSME	19.70	25.44	41.28	54.48	63.23
ReIdCNN+Attributes	59.99	81.45	91.36	93.78	96.81
ACRN	62.60	83.61	92.61	**95.34**	**97.00**

Table 4. Results of our approach on the Market-1501 dataset. Our performance is competitive and outperforms other recent works at higher ranks.

Method	mAP	r1	r5	r10	r20
LOMO+XQDA [28]	17.04	30.75	-	-	-
GAN [38]	47.13	67.69	-	-	-
APR [21]	51.88	70.69	-	-	-
ReIdNet	49.41	68.74	78.31	84.98	88.97
Attributes	7.23	11.34	24.42	32.31	41.24
Attributes+KISSME	12.83	21.97	42.28	50.00	60.53
ReIdCNN+Attributes	50.01	69.91	80.34	86.87	90.43
ACRN	**51.96**	**72.58**	**84.79**	**88.87**	**91.52**

Table 5. State-of-the-art results of our approach on the DukeMTMC-reID dataset.

5.2.3 Attributes in Re-Id

In Table 6 we show the attributes that are most and least strongly weighted by our approach on all three datasets. There is a clear correlation between the attribute weighting and their original accuracy on the source dataset. We observe that many of the highly rated attributes are common across re-id datasets. Furthermore, many of them focus on description of the upper body which is usually the largest portion of the image. Unsurprisingly, the attributes rated lowest by ACRN include those that occur rarely and are very specific (e.g. messenger bag). Unfortunately, it is exactly such rarely occurring attributes which have the potential to be the most distinguishing for person re-id.

CUHK3	**Market-1501**	**DukeMTMC-reID**
Male	Backpack	Jacket
Long Hair	Skirt	Casual Upper
Jacket	Male	Trousers
Backpack	Long Hair	Male
Sandals	Hat	Sandals
V-neck	V-Neck	Messenger Bag

Table 6. Attributes with the highest (top) and lowest (bottom) weights received by ACRN. Upper body attributes receive higher weights.

It can generally be observed that attributes which are visible only in very small portions of a person image are determined to be of less help for re-id by ACRN, even if they had an originally high detection accuracy on the PETA dataset (e.g. hat with 89.4%). The small amount of information in the image which the attribute classifier needs to focus on appears to get lost in the domain gap between source and target datasets.

6. Conclusion

We have presented a person re-id approach which includes automatically predicted attribute information into the training process of a CNN. This allowed the CNN to focus on learning information for person re-id which is complementary to the information contained in the attribute pre-

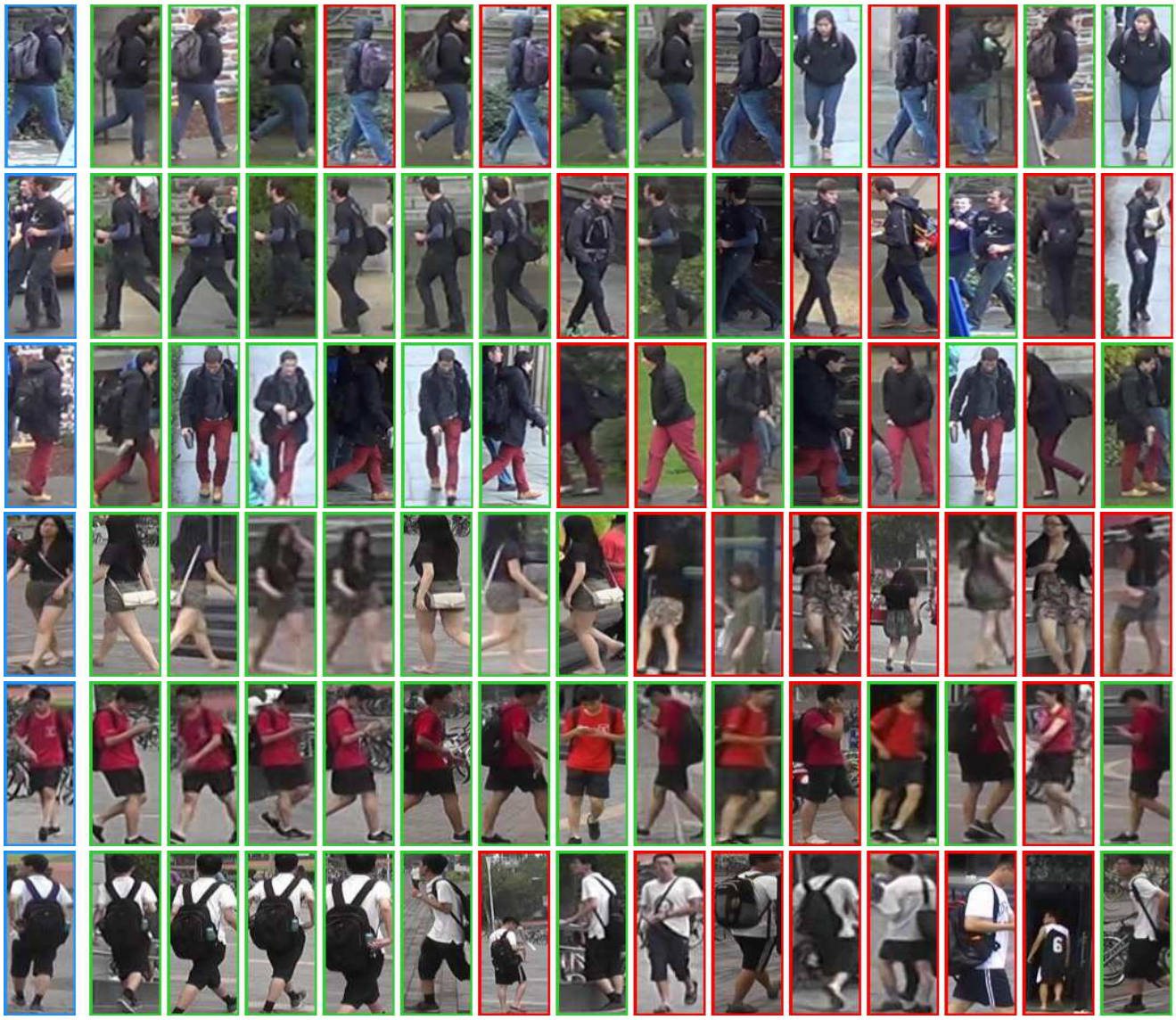

Figure 3. Qualitative results of our ACRN approach for challenging queries on the DukeMTMC-reID (top three rows) and Market-1501 (bottom three rows) datasets. The query images are highlighted in blue and the top 14 results are shown. Correct matches are highlighted green and false matches in red. Note that false results are often semantically and visually similar to the query.

dictions. Our experiments show that the approach outperforms both sole attributes or CNNs and even naïve fusion of the two. Our combined approach achieves competitive or state-of-the-art results on three public datasets.

For future work we intend to focus on making better use of the discriminative potential contained in attributes which are rare and thus often suffer from a low recognition accuracy.

Acknowledgement This work was partially supported by the German Federal Ministry of Education and Research (BMBF) under grant no. 13N14029.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013.

[3] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

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

[5] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM, 2014.

[6] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[8] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Computer Vision–ECCV 2008*, pages 262–275, 2008.

[10] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, pages 780–793. Springer, 2012.

[11] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1062–1070, 2015.

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

[13] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis. Joint learning for attribute-consistent person re-identification. In *European Conference on Computer Vision*, pages 134–146. Springer, 2014.

[14] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[15] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *Bmvc*, volume 2, page 8, 2012.

[16] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 111–115. IEEE, 2015.

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

[18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[19] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

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

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

[22] E. T. Matsukawa and E. Suzuki. Person re-identification using cnn features learned from combination of attributes. ICPR, 2016.

[23] 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.

[24] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3739–3747, 2015.

[25] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[28] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[29] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

[30] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*, 2016.

[32] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[33] W. L. R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification.

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

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

[36] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

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

[38] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017.