AANet: Attribute Attention Network for Person Re-Identifications

Chiat-Pin Tay[1,2], Sharmili Roy[2], and Kim-Hui Yap[1,2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore

Abstract

This paper proposes Attribute Attention Network (AANet), a new architecture that integrates person attributes and attribute attention maps into a classification framework to solve the person re-identification (re-ID) problem. Many person re-ID models typically employ semantic cues such as body parts or human pose to improve the re-ID performance. Attribute information, however, is often not utilized. The proposed AANet leverages on a baseline model that uses body parts and integrates the key attribute information in an unified learning framework. The AANet consists of a global person ID task, a part detection task and a crucial attribute detection task. By estimating the class responses of individual attributes and combining them to form the attribute attention map (AAM), a very strong discriminatory representation is constructed. The proposed AANet outperforms the best state-of-the-art method [20] using ResNet-50 by 3.36% in mAP and 3.12% in Rank-1 accuracy on DukeMTMC-reID dataset. On Market1501 dataset, AANet achieves 92.38% mAP and 95.10% Rank-1 accuracy with re-ranking, outperforming [11], another state of the art method using ResNet-152, by 1.42% in mAP and 0.47% in Rank-1 accuracy. In addition, AANet can perform person attribute prediction (e.g., gender, hair length, clothing length etc.), and localize the attributes in the query image.

1. Introduction

Given a query image, person re-ID aims to retrieve images of a queried person from a collection of network-camera images. The retrieval is typically attempted from a collection of images taken within a short time interval with respect to the queried image. This supports the underlying assumption that the query person's appearance and clothing attributes remain unchanged across the query and the collection images. Person re-ID is a challenging problem due to many factors such as partial/total occlusion of the subject, pose variation, ambient light changes, low image resolution, etc. Recent deep learning based re-ID solutions have demonstrated good retrieval performance.

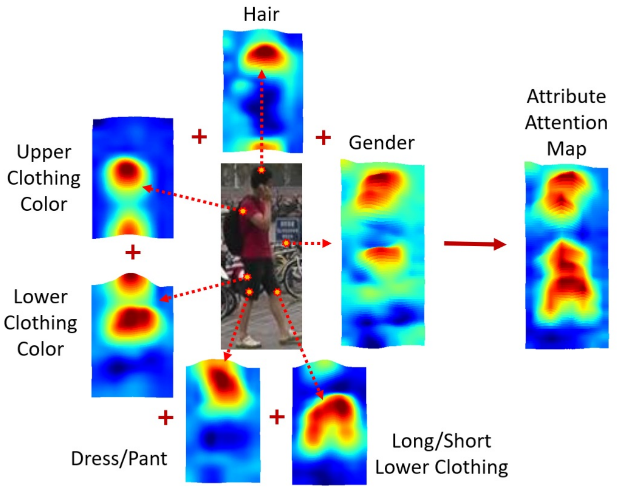

Figure 1. Class-aware heat maps are extracted and combined to form a discriminatory Attribute Attention Map (AAM) at the image level. The six heat maps shown here correspond to the six attributes such as hair, upper clothing color, lower clothing color etc. Best viewed in color.

The approaches used to solve the person re-ID problem can be broadly divided into two categories. The first category comprises of metric learning methods that attempt to learn an embedding space which brings images belonging to a unique person close together and those belonging to different persons far away. Various approaches such as triplet and quadruplet losses have been employed to learn such embedding spaces [10, 2].

The second category of methods poses the re-ID problem in a classification set-up. Such methods learn by using Softmax normalization and computing cross-entropy loss, based on person identity as ground truth, for back-propagation. Research has shown that by integrating semantic information such as body parts, human pose etc, the classifica-

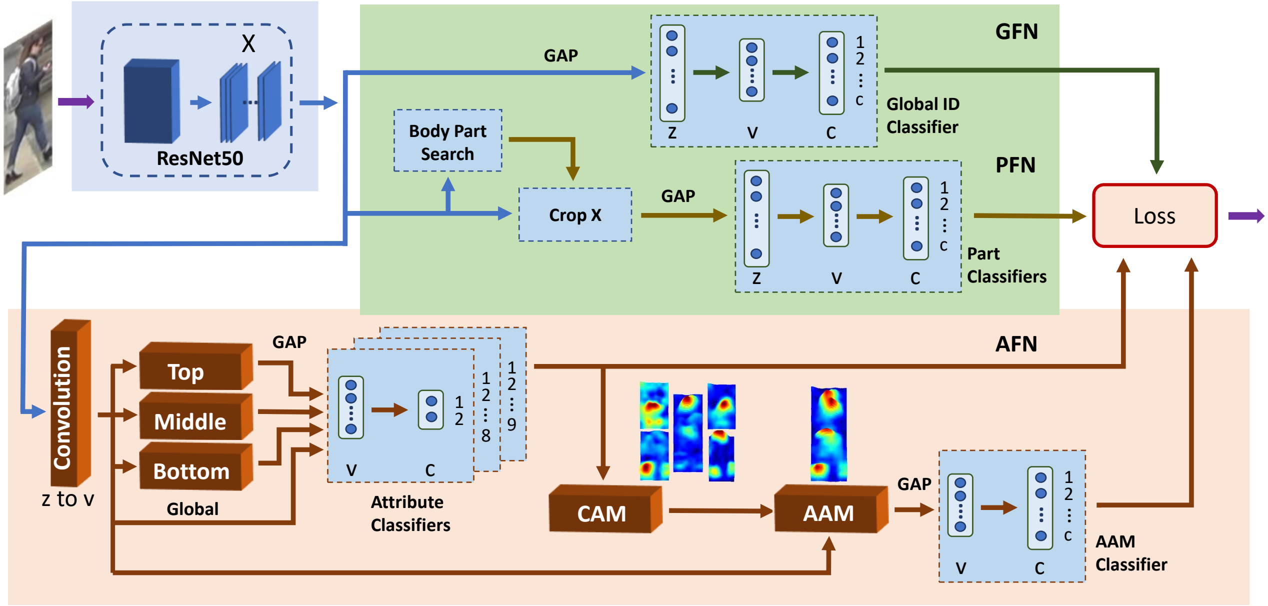

Figure 2. Overview of AANet. The backbone network, which is based on the ResNet-50 architecture, outputs the feature map X. The feature map X is forwarded to three tasks, namely the Global Feature Network(GFN), Part Feature Network (PFN) and Attribute Feature Network (AFN). The output of these three tasks are combined using homoscedastic uncertainty learning to predict the person identification. Best viewed in color.

tion and recognition accuracy can be significantly improved [18, 22, 24]. Person attributes, such as clothing color, hair, presence/absence of backpack, etc. are, however, not used by the current state-of-the-art re-ID methods. Since, a typical re-ID model assumes that the physical appearance of the person of interest would not significantly change between query image and the search images, physical appearance becomes a key information that can be mined to achieve higher re-ID performances. Such information is not utilized by current research and the state of the art re-ID methods.

In this work, we propose to utilize the person attribute information into the classification framework. The resulting framework, called the Attribute Attention Network (AANet), brings together identity classification, body part detection and person attribute into an unified framework that jointly learns a highly discriminatory feature space. The resulting network outperforms existing state of the art methods in multiple benchmark datasets.

Figure 2 gives an overview of the proposed architecture. The proposed framework consists of three subnetworks. The first network, called the Global Feature Network (GFN), performs global identity (ID) classification based on the input query image. The second network, called the Part Feature Network (PFN), focuses on body part detection. The third network is the Attribute Feature Network (AFN), which extracts class-aware regions from the persons attributes to generate Attribute Attention Map (AAM). This is shown in Figure 1. The three networks perform classification using person ID and attribute labels We use homoscedastic uncertainty learning to optimize the weights of the three sub-tasks for final loss calculations.

Since AANet performs person attribute classification as part of network learning, it also output attribute predictions for each query and gallery images. This enables attribute matching of the gallery images, with or without retrieval by query image.

Our key contributions can be summarized as follows:

1. We provide a new network architecture that integrates attribute features with identity and body part classification in a unified learning framework.

2. We outperform the existing best state-of-the-art re-ID method on multiple benchmark datasets and propose the new state of the art solution for person re-ID.

The rest of the paper is organized as follows. Section 2 provides an overview of the related works. In section 3 we describe the proposed AANet framework. Experimental results are provided in section 4 and 5. The paper is concluded in section 6.

2. Related works

In recent years, deep learning was used to solve various challenging computer vision tasks [14, 9, 10, 6, 7]. In this section, we provide an overview of the recent re-ID deep

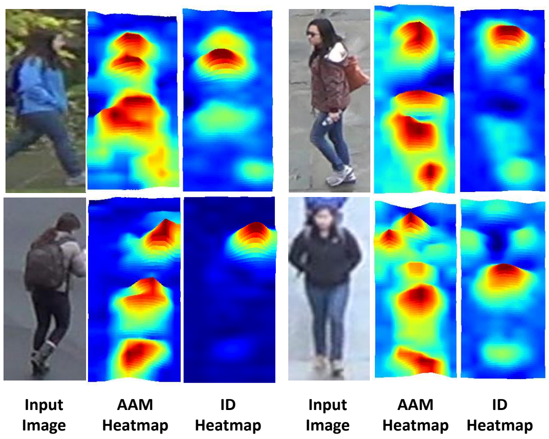

Figure 3. Comparison between proposed AAM and class activation ID heatmap generated in GFN. The AAM captures person attributes, therefore the activated areas lie mostly within the pedestrian body. The ID heatmap is influenced by the training dataset, is less dense and may include background as part of its feature map. Best viewed in color.

learning based methods that achieve close to state-of-the-art performance.

Deep learning based re-ID solutions are often posed as an identity classification problem. Authors in [23] used multi-domain datasets to achieve high re-ID performance in a classification set-up. Wang *et al.* [21] proposed to formulate the re-ID problem as a joint learning framework that learns feature representations using not only the query image but also query and gallery image pairs. Many solutions have used additional semantic cues such as human pose or body parts to further improve the classification performance. Su *et al.* [18] proposed a Pose-driven Deep Convolutional (PDC) model to learn improved feature extraction and matching models from end-to-end. Wei *et al.* [22] also adopted the human pose estimation, or key point detection approach, in his Global-Local-Alignment Descriptor (GLAD) algorithm. The local body parts are detected and learned together with the global image by the four-stream CNN model, which yields a discriminatory and robust representation. Yao *et al.* [24] proposed the Part Loss Networks (PL-Net) to automatically detect human parts and cross train them with the main identity task. Zhao *et al.* [25] follows the concept of attention model and uses a part map detector to extract multiple body regions in order to compute their corresponding representations. The model is learned through triplet loss function. Sun *et al.* [20] proposed a strong Part-based Convolutional Baseline method, with Refined Part Pooling method to re-align parts for high accuracy performance. Kalayeh *et al.* [11] used multiple datasets, deep backbone architecture, large training images, and human semantic parsing to achieve good accuracy results. Similarly, Jon *et al.* [1] proposed using deep backbone architecture and large input image, but with classification as first pass learning, followed by metric learning for accuracy fine-tuning.

Integrating semantic information such as body parts and pose estimation have shown significant improvement in re-ID performance. Since a person's attributes do not change significantly between the query image and the gallery images, we believe that attributes form a key information that can significantly impact person re-ID performance. This, however, has not been utilized in the current re-ID methods. In view of this, we propose to integrate physical attributes to the identity classification framework.

3. Proposed Attribute Attention Network (AANet)

The proposed AANet is a multitask network with three sub-networks, namely the GFN, PFN and AFN (Figure 2). The sub-network, GFN, performs global image-level ID classification. The PFN detects and extracts localized body parts before the classification task. The AFN uses person attributes for the classification task and generates the Attribute Activation Map (AAM) that plays a crucial role in identity classification. Some examples of AAM are shown in Figure 3. In the figure, the generated AAMs provide more discriminatory features than the ID heatmap. As a result, when GFN, PFN and AFN learn together, our AANet becomes more generic and better at predicting person ID. The various components of AANet are described in details in the following sections.

The backbone network of AANet is based on ResNet architecture (Figure 2) since ResNet is known to perform well in re-ID problems. We removed the fully connected layer of the backbone network so that AANet's sub-networks can be integrated. There are four classifiers within AANet. They are the Global ID Classifier, Part Classifiers, Attribute Classifiers and AAM Classifier. The Global ID and Part classifiers belong to GFN and PFN respectively. The Attribute and AAM Classifiers belong to AFN. All four classifiers have rather similar network design. All of them utilize global average pooling to reduce over-fitting, and there is a 3 layers (Z, V and C) architecture to increase network depth for better feature learning. The classifiers learn using Softmax normalization and Cross-entropy loss.

3.1. Global Feature Network

This network performs the identity (ID) classification using the query image (Figure 2). The convolutional feature map $X \in R^{Z \times H \times W}$ extracted by the backbone network is provided as input to a global average pooling (GAP) layer. This is followed by a 1x1 convolution layer that brings the dimensionality down to V. BatchNorm and Relu are then applied to V before linear transformation to C, which is used

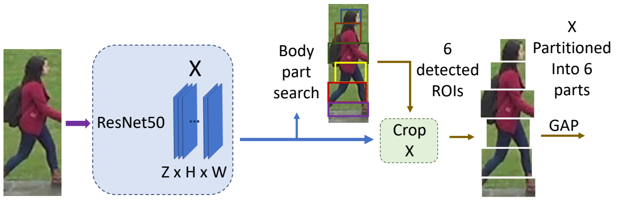

Figure 4. The PFN divides the feature map $X \in R^{Z \times H \times W}$ into six ROIs using peak activation detection and pooling. Features from these 6 ROIs are further used for identity classification. Best viewed in color.

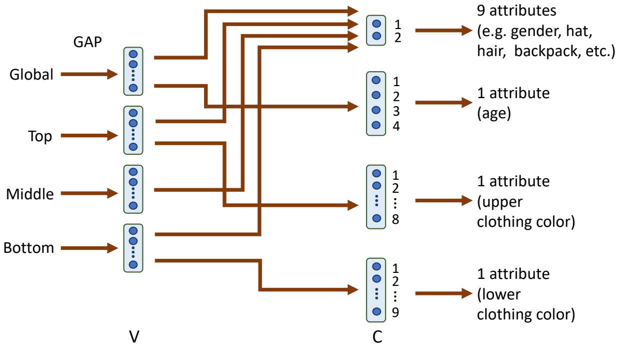

Figure 5. 12 attributes are generated from global, top, middle and bottom vectors on Market1501 dataset.

by Softmax function. Cross-entropy loss is calculated on Softmax output for learning using back-propagation.

3.2. Part Feature Network

This network performs ID classification on body parts using the same person ID labels used in GFN. The architecture is shown in Figure 4. The body part detector partitions the convolutional feature map X into six horizontal parts and estimate the corresponding regions of interest (ROIs). This is done by identifying peak activation regions in X. Let (h_z, w_z) denote the peak activation location in each feature map z of X where $z \in \{1, \ldots, Z\}$.

$$(h_z, w_z) = \arg\max_z X_z(h, w) \tag{1}$$

where $X_z(h, w)$ is the activation value at location (h, w) on the z'th feature channel of X. These locations are then clustered into 6 bins based on their vertical positions. These 6 bins constitute the 6 ROIs/ parts. The feature map X is now divided into 6 parts using these ROIs. Figure 4 shows this process. Once the 6 parts are computed, the subsequent processing, which is shown in Figure 2 is performed similarly as in GFN.

3.3. Attribute Feature Network

The AFN captures the key attribute information in the AANet architecture (Figure 2). The AFN consists of two sub-tasks (i) attribute classification and (ii) attribute attention map (AAM) generation. The first sub-task performs classification on individual person attributes. The second sub-task leverages on the output of first sub-task and generates class activation map (CAM) [28] for each attribute. CAM is a technique to localize the discriminatory image regions even though the network is trained on image-level labels only. Thus CAM fits well for AANet use. The CAMs generated from selected attribute classes are combined to form a feature map that is forwarded to the AAM Classifier for learning. We describe these two sub-tasks in the following paragraphs in detail.

(i) Attribute classification The first sub-task of AFN is to perform attribute classification. There are 10 and 12 annotated attributes on DuketMTMC-reID and Market1501 respectively. The first layer of AFN is a 1x1 convolution that downsized the channel depth of feature map X from Z to V. Next, we partition the feature maps into three different sets, namely the Top, Middle and Bottom feature maps, each responsible for extracting features from their localized regions. Part-based modeling is known to reduce background clutter and improve classification accuracy. The different parts focus on different attributes. The Top feature maps, for example, are used for capturing features such as hat, hair, sleeves and upper clothing color etc. Features from the lower half of the body are ignored in the Top feature map. As shown in Figure 5, the outputs of these feature maps, together with global feature map, are average pooled to generate 4 feature vectors at layer V. These 4 vectors are the input to the fully connected layer C. On Market1501, there are 4 classifiers at layer C, each generating their own attribute predictions.

(ii) Attribute Attention Map The Attribute Attention Map (AAM) is the input to the Attribute Classifier (Figure 2, which performs person ID classification. AAM combines class sensitive activation regions from individual attributes. These individual class-sensitive activation regions are extracted using CAM from each person attribute. As explained before, CAM uses GAP, with little tweak, to generate discriminatory image regions. Thus, CAM's output reveals image regions representing the attribute. Figure 6 shows some example of class sensitive activation regions and the combined AAM. For qualitative comparison, the second column in the figure shows the activation map generated by the global identification task (GFN). The subsequent columns show the class specific activation regions of various attributes such as gender, hair, sleeve, upper clothing color etc. The sixth column, for example, depicts the class specific activation region for upper clothing color. We can observe that the activation region corresponds to the upper clothing region in the input query image.

Out of the 12 available attributes, the gender, hair, upper and lower clothing colors, lower clothing type and length

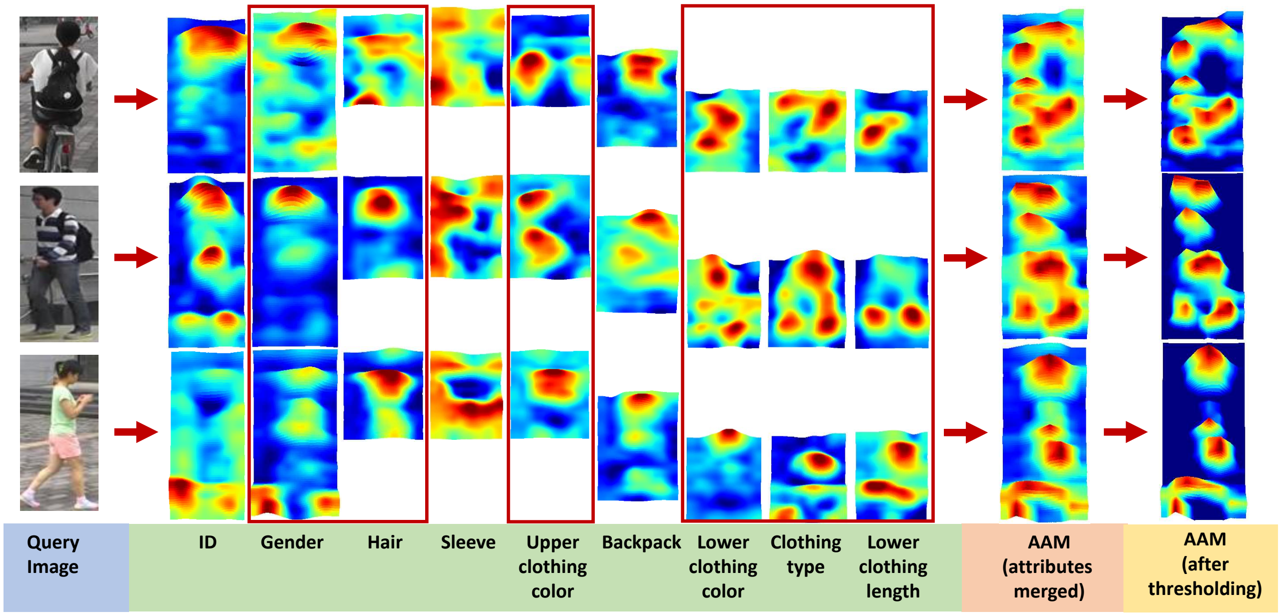

Figure 6. On Market1501 dataset, a total of 12 attention maps are used. Eight most important attributes are shown. Only visual cues (in red boxes) are selected for AAM generation. Global attention map obtained from the GFN is shown here for qualitative comparison with other attributes. Backpack, handbag, bag and hat attributes, despite being important visual cues, do not appear in all images, and are therefore dropped. Attention map for sleeves captures too much background information and thus is unsuitable for AAM. Best viewed in color.

are good choices for AAM generation. The AAM generation process involves merging the individual class specific activation regions by maximum operation and performing an adaptive thresholding. The thresholding process removes some background regions that sometimes appear within the class specific activation region. An example of this can be observed in the second row of Figure 6 where the Lower clothing activation map contains some background region but on thresholding the region is removed from the generated AAM. When qualitatively comparing the class activation map generated by the global feature network on the same query image, we can see that the AAM was more specific in localizing regions with distinct attribute information. The Attribute Classifier (Figure 2) takes the AAM and perform ID classification, and shares the learning experience with the GFN and PFN.

3.4. Loss calculation

The proposed AANet is formulated as a multitask network. The multitask loss function of the AANet is defined as follows:

$$\mathcal{L}_{total}(x, W, \lambda) = \sum_{i=0}^{T} \lambda_i \mathcal{L}_i(x, W) \tag{2}$$

Where x is a set of training images, W is the weights on input x. T is the total number of task loss $\mathcal{L}_i$. λ_i are the task loss weighting factors, and it plays an important role in optimizing the accuracy performance of AANet. If we assign equal weighting to all λ_i, the retrieval accuracy will not be optimal. In our work, we used homoscedastic uncertainty learning [13, 8, 12] to obtain the task loss weighting. We define the Bayesian probabilistic model classification likelihood output as

$$p(y|f^W(x), \sigma) = Softmax(\frac{1}{\sigma^2} f^W(x)) \tag{3}$$

Where $f^W(x)$ is the output of the neural network and is scaled by σ^2. σ is the observation noise. The log likelihood for this output is given by

$$\begin{aligned} log(p(y = c|f^W(x), \sigma)) &= \frac{1}{\sigma^2} f_c^W(x) \\ -log(\sum_{i=0}^{C} exp(&\frac{1}{\sigma^2} f_i^W(x))) \end{aligned} \tag{4}$$

Where C is the number of classes for the classification task. The task loss $\mathcal{L}(x, W, \sigma)$ can be formulated as $-log(p(y = c|f^W(x), \sigma))$. What we need is the cross-entropy loss of the non-scaled y, which if defined as $\mathcal{L}(x, W) = -log$ Softmax(y, f^W(x)) [13], the loss function can be simplified to

$$\mathcal{L}(x, W, \sigma) \approx \frac{1}{\sigma^2} \mathcal{L}(x, W) + log\, \sigma \tag{5}$$

By applying the above loss function to our AANet, the final AANet loss function is now given as

$$\mathcal{L}(x, W, \sigma_g, \sigma_p, \sigma_a, \sigma_{aa}) \approx \frac{1}{\sigma_g^2}\mathcal{L}_g(x, W) + \frac{1}{\sigma_p^2}\mathcal{L}_p(x, W) + \frac{1}{\sigma_a^2}\mathcal{L}_a(x, W) + \frac{1}{\sigma_{aa}^2}\mathcal{L}_{aa}(x, W) + log\, \sigma_g \sigma_p \sigma_a \sigma_{aa} \tag{6}$$

Where $\mathcal{L}_g$, $\mathcal{L}_p$, $\mathcal{L}_a$ and $\mathcal{L}_{aa}$ represent global, part, attribute and attribute attention loss respectively. σ_g, σ_p, σ_a and σ_{aa} represent observation noises for global, part, attribute and attribute attention tasks respectively, and are inversely proportional to λ_i

3.5. Implementation

We implemented AANet with ResNet-50 and ResNet-152 as backbone networks, and pre-trained them with the ImageNet [4] dataset. Training images are enlarged to 384 x 128, with only random flip as the data augmentation method. Batch size is set to 32 for ResNet-50, and 24 for ResNet-152. Using Stochastic Gradient Descent (SGD) as the optimizer, we train the network for 40 epoch. Learning rate starts at 0.1 for the newly added layers, and 0.01 for the pretrained ResNet parameters, and follows the staircase schedule at 20 epoch with a 0.1 reduction factor for all parameters. In all the three sub-networks, the value of Z is 2048 and that of V is 256. The value of C depends on the dataset under evaluation. For DukeMTMC-reID [16], C is 702 and for Market1501 [26] it is 751.

During testing, we concatenate the outputs of the V layer from all the classifiers, namely, the global identity classifier, the body part classifier, the attribute classifier and the AAM classifier (Figure 2) to form the representation of the query image. For ranking, we use the l_2 norm between these descriptors of the query and the gallery images.

4. Experimental Results

In the following experiments, we used the DukeMTMC-reID [16] and Market1501 [26] datasets to conduct our training and testing. DukeMTMC-reID is a subset of DukeMTMC dataset. The images are cropped from videos taken from 8 cameras. The dataset consists of 16,522 training images and 17,661 gallery images, with 702 identities for both training and testing. 408 distractor IDs are also included in the dataset. There are a total of 23 attributes annotated by Lin *et al.* [15]. We use all attributes, but with modification to the clothing color attributes. We merged all 8 upper clothing color attributes and 7 lower clothing color attribute into a single upper clothing attribute and a single lower clothing attribute respectively.

For Market1501, there are a total of 32,668 images for both training and testing. There are 751 identities allocated for training and 750 identities for testing. Lin *et al.* [15] also annotated this dataset, but with 27 person attributes. We use the same clothing color strategy as in DukeMTMC-reID, and use all attributes for training our model.

	DukeMTMC-reID	
Methods	mAP	Rank-1
FMN [5]	56.9	74.5
SVDNet [19]	56.8	76.7
DPFL [3]	60.6	79.2
KPM (Res-50) [17]	63.2	80.3
PCB (Res-50)[20]	69.2	83.3
Proposed AANet-50	**72.56**	**86.42**
GP-reID (Res-101) [1]	72.80	85.20
SPReID (Res-152) [11]	73.34	85.95
Proposed AANet-152	**74.29**	**87.65**
GP-reID (Res-101)[1] + RR	85.60	89.40
SPReID (Res-152)[11] + RR	84.99	88.96
Proposed AANet-152 + RR	**86.87**	**90.36**

Table 1. Performance comparison with other state-of-the-art methods using DukeMTMC-reID dataset. AANet-50 denotes AANet trained using ResNet-50. AANet-152 denotes AANet trained using ResNet-152. RR denotes Re-Ranking[27] .

4.1. Comparison with existing methods

DukeMTMC-reID dataset We perform comparison with the state-of-the-art methods in Table 1. The table has three parts based on the backbone network being used. First part compares models based on ResNet-50. The three comparative networks are KPM (Res-50) [17], PCB (Res-50)[20] and the proposed AANet-50. We outperform the best state-of-the-art method [20] in this category by 3.36% in mAP and 3.12% in Rank-1 accuracy.

The second comparison is based on networks using larger backbone models, which include both ResNet-101 and ResNet-152. The three comparative networks are GP-reID (Res-101) [1], SPReID (Res-152) [11] and the proposed AANet-152. Here, we again outperformed the state-of-the-art method [11] by 0.95% in mAP and 1.70% in Rank-1 accuracy.

The third comparison is based on networks from the second comparison, but this time with re-ranking [27]. We outperformed the state-of-the-art method [1] by 1.27% in mAP and 0.96% in Rank-1 accuracy.

Market1501 dataset We perform similar comparisons as in previous section in Table 2 using the Market1501 dataset. In the first comparison, which uses ResNet-50, the networks selected are KPM (Res-50) [17], PCB (Res-50)[20] and the proposed AANet-50. We outperformed the best state-of-the-art method [20] in this category by 0.85% in mAP and 0.09% in Rank-1 accuracy.

Methods	Market1501 mAP	Rank1	Rank10
PDC [18]	63.4	84.4	94.9
PL-Net [24]	69.3	88.2	-
DPFL [3]	73.1	88.9	-
GLAD [22]	73.9	89.9	-
KPM (Res-50)[17]	75.3	90.1	97.9
PCB (Res-50)[20]	81.6	93.8	98.5
Proposed AANet-50	**82.45**	**93.89**	**98.56**
GP-reID (Res-101) [1]	81.20	92.20	-
SPReID (Res-152) [11]	83.36	93.68	98.40
Proposed AANet-152	**83.41**	**93.93**	**98.53**
GP-reID (Res-101)[1]+RR	90.00	93.00	-
SPReID (Res-152)[11]+RR	90.96	94.63	97.65
Proposed AANet-152+RR	**92.38**	**95.10**	**97.94**

Table 2. Performance comparison with other state-of-the-art methods using Market1501 dataset. AANet-50 denotes AANet trained using ResNet-50. AANet-152 denotes AANet trained using ResNet-152. RR denotes Re-Ranking[27].

The second comparison is made using GP-reID (Res-101) [1], SPReID (Res-152) [11] and the proposed AANet, with either ResNet-101 or ResNet152. Here, we again outperformed the state-of-the-art method [11] by 0.05% in mAP and 0.25% in Rank-1 accuracy.

The third comparison is made on networks from the previous section but with re-ranking [27]. We outperform the state-of-the-art method [11] by 1.42% in mAP and 0.47% in Rank-1 accuracy. We believe that the attribute information is a key contributor in AANet's person re-ID performance.

4.2. Network Analysis

In this section, we study the effect of task loss weights and the size of the backbone network on the re-ID performance. We also review various training parameters.

Ablation Study In Table 3, we show the impact of the task loss weights on AANet accuracy performance using DukeMTMC-reID dataset. The global ID task, the part task, the attribute classification task and the attribute attention map task are denoted as $\mathcal{L}_g$, $\mathcal{L}_p$, $\mathcal{L}_a$ and $\mathcal{L}_{aa}$ respectively. As we add each of these relevant tasks to the network, the accuracy improves, which justifies the contribution of each task on the overall performance. When we use homoscedastic uncertainty learning to obtain the task loss weights $\mathcal{L}_g$, $\mathcal{L}_p$ and $\mathcal{L}_a$, the performance improves to 70.47% mAP and 85.44% Rank-1 accuracy. This result alone is enough to outperform the best state-of-the-art method using ResNet-50. With the integration of AAM, which provides more discriminatory features for learning, we improve the accuracy results to 72.56% mAP and 86.42% Rank-1 accuracy.

AANet-50 Task Loss	Task Weights				mAP %	Rank 1 %
$\mathcal{L}_g$	1	0	0	0	62.92	80.18
$\mathcal{L}_g + \mathcal{L}_p$	1	1	0	0	66.35	82.93
$\mathcal{L}_g + \mathcal{L}_p + \mathcal{L}_a$	1	1	1	0	67.28	83.29
$\mathcal{L}_g + \mathcal{L}_p \ + \mathcal{L}_a$	Uncertainty Learning				70.47	85.44
$\mathcal{L}_g + \mathcal{L}_p + \mathcal{L}_a + \mathcal{L}_{aa}$					**72.56**	**86.42**

Table 3. Performance comparisons of different combination of task losses using DukeMTMC-reID dataset, with and without uncertainty learning. The top three rows are AANet accuracy with equal weights to the tasks. Bottom two rows show the results with loss weights obtained from uncertainty learning.

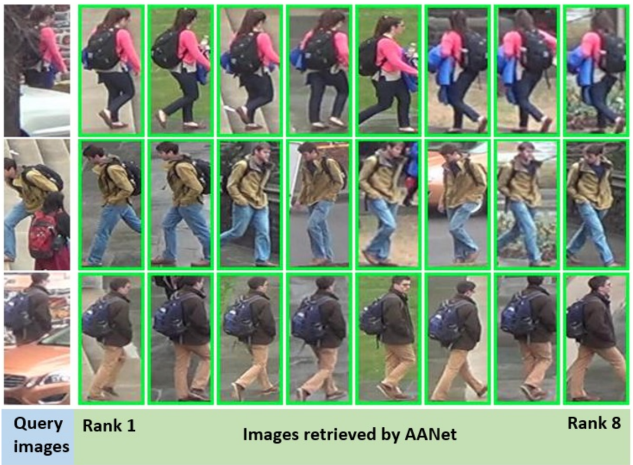

Figure 7. Three queries from the DukeMTMC-reID with eight retrieved images for each query.

Effect of Backbone Network The depth of the backbone network affects the accuracy performance of person re-ID. Deeper networks yield better result, and this is clearly shown in both Tables 1 and 2. Table 1 also shows that the our proposed smaller AANet-50 outperformed deeper SPReID (Res-152) [11] in Rank-1 accuracy by 0.47%, and GP-reID (Res-101) [1] by 1.22%. We achieved similar Rank-1 results on Market1501 dataset, with our AANet-50 outperforming those using deeper backbone networks.

Effect of Training Parameters Many tricks have been used in the literature to enhance accuracy [11] and [1]. In [11], the authors aggregate a total of 10 different datasets to generate $\sim$ 111k images and $\sim$ 17k identities for training and testing. In addition, multiple image sizes are used to train the network in different phases. In [1], authors use techniques such as pre-training before regression learning, large image size, hard triplet mining and deeper backbone network for good person re-ID. These are good practices. However, the proposed AANet uses smaller image size, simpler training process, and a shallower ResNet-50 architecture to outperform existing state-of-the-art.

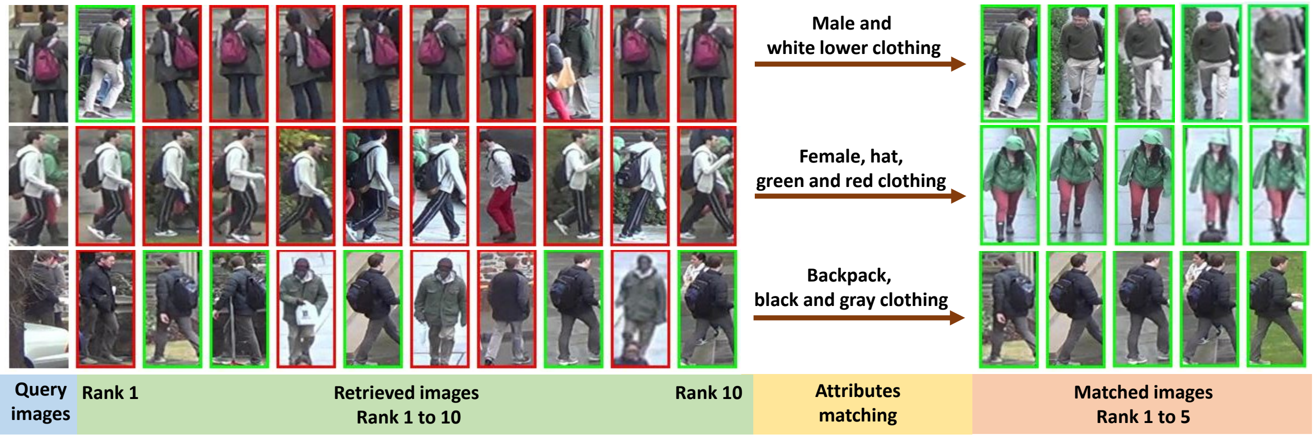

Figure 8. Three examples of how person attributes help in improving the image retrieval accuracy. These are challenging image queries that return many falsely accepted images. Since AANet returns each query and gallery images with predicted attributes, it provides an option for the user to use attribute matching to filter away the unwanted retrieved images. The useful attributes include gender. clothing colors, backpack, etc. Green box denotes same ID as query image. Red box denotes different ID from query image. Best viewed in color

Methods	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	mean
APR [15]	86.45	87.08	83.65	93.66	93.32	91.46	82.79	88.98	75.07	97.13	73.40	69.91	85.33
AANet-152	**92.31**	**88.21**	**86.58**	**94.45**	**94.24**	**94.83**	**87.77**	**89.61**	**79.72**	**98.01**	**77.08**	**70.81**	**87.80**

Table 4. Performance comparisons of attribute accuracy on Market1501 dataset.

5. Experimental Results Using Attribute

In this section, we illustrate how person attributes help in refining the retrieved images for person re-ID.

5.1. Retrieval results

We show three retrieval examples using AANet in Figure 7. Though there are some occlusions on the query subjects, for examples, cars and unwanted pedestrian, AANet has no problem retrieving correct images from the gallery set.

In Figure 8, we show some examples of challenging queries where the subjects are heavily occluded. This resulted in poor retrieval accuracy. The figure demonstrates how AANet provides an option for the user to filter away the incorrect retrievals by using predicted attributes from query and gallery images. Three examples are given, each with their own retrieval difficulties. First example is given in row one. More than half of the query subject is occluded by another pedestrian. Most computer vision methods will pick the unwanted pedestrian as subject of interest, and return wrong images. In this example, 9 out of 10 images are wrongly retrieved, which results in poor mAP performance. Through AANet's attribute matching, those wrong images can be filtered out easily without laborious manual filtering. The ranking of theses attribute matched images were 1, 19, 38, 78, 172 during first retrieval, indicating how difference they are to query image. Same challenging queries are given in row two and three. As in first example, attribute filtering are performed to return correct images up to rank 5.

5.2. Attribute Classification Performance

The accuracy of attribute classification of the proposed AANet is compared with APR [15] in Table 4. APR [15] is provided by Lin *et al*., the author who annotated the DukeMTMC-reID and Market1501 datasets with person attributes. Since AANet employs localized attribute features to enhance network learning, we obtained better representations and outperforms APR in every attribute prediction.

6. Conclusions

In this paper we propose a novel architecture to incorporate attributes based on physical appearance such as clothing color, hair, backpack etc. into a classication based person re-ID framework. The proposed Attribute Attenion Network (AANet) employs joint end-to-end learning and homoscedastic uncertainty learning for multitask loss fusion. The resulting network outperforms existing state-of-the-art re-ID methods on multiple benchmark datasets.

Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore.

References

[1] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person reidentification. *arXiv preprint arXiv:1801.05339*, 2018. 3, 6, 7

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 1

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person reidentification by deep learning multi-scale representations. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 2590–2600, 2017. 6, 7

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 6

[5] Guodong Ding, Salman Hameed Khan, Zhenmin Tang, and Fatih Porikli. Let features decide for themselves: Feature mask network for person re-identification. *CoRR*, abs/1711.07155, 2017. 6

[6] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multiscale aggregation for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[7] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *CoRR*, abs/1506.02142, 2015. 5

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 1, 2

[11] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018. 1, 3, 6, 7

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 5

[13] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115, 2017. 5

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. 2

[15] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *CoRR*, abs/1703.07220, 2017. 6, 8

[16] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 6

[17] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6886–6895, 2018. 6, 7

[18] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3980–3989, 2017. 2, 3, 7

[19] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3820–3828, 2017. 6

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 3, 6, 7

[21] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and crossimage representations for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1288–1296, 2016. 3

[22] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. GLAD: global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 420–428, 2017. 2, 3, 7

[23] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1249–1258, 2016. 3

[24] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Deep representation learning with part loss for person re-identification. *CoRR*, abs/1707.00798, 2017. 2, 3, 7

[25] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person reidentification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3239–3248, 2017. 3

[26] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1116–1124, 2015. 6

[27] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3652–3661, 2017. 6, 7

[28] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 4