# On the Importance of Appearance and Interaction Feature Representations for Person Re-identification

Richard Blythman*,[1], Andrea Zunino*,[1], Christopher Murray[1] and Vittorio Murino[1,2]
[1]Ireland Research Center, Huawei Technologies Co. Ltd., Dublin, Ireland
[2]Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy
{richard.blythman, christopher.murray}@huawei.com, {andrea.zunino, vittorio.murino}@iit.it

## Abstract

*In recent person re-identification (Re-ID) approaches, combining global and local appearance-based features has been shown to increase performance effectively. These types of models are often characterized by multiple branches that act as experts for specific local regions or global high-level semantic features. We argue that attention mechanisms can be useful for multi-branch Re-ID models by creating more robust representations based on the interaction of informative image features. In this paper, we investigate this idea and propose a novel multi-branch architecture with experts that learn distinct representations based on (i) the global image appearance and (ii) the interaction between features. Unlike former methods with local experts acting on partitions that are fixed a-priori, our feature interaction expert uses a novel attention-based pooling to automatically extract semantically-rich and discriminative features from different regions of a person image. Compared with existing attention-based algorithms, our method maintains the feature interaction information separately in order to discriminate between identities. Our approach achieves state-of-the-art performance across three popular benchmarks - CUHK03, Market1501 and MSMT17. Furthermore, saliency visualizations show that appearance and interaction experts learn complementary representations that attend to multiple discriminant regions, leading to improved classification ability.*

## 1. Introduction

Person re-identification (Re-ID) is a crucial component of smart cities, surveillance and ambient intelligence scenarios. Given a query image of a person of interest and a gallery of images corresponding to different identities (captured by cameras in different places at various times), the goal is to determine a match between the images based only on their whole appearance, and typically without using biometric cues.

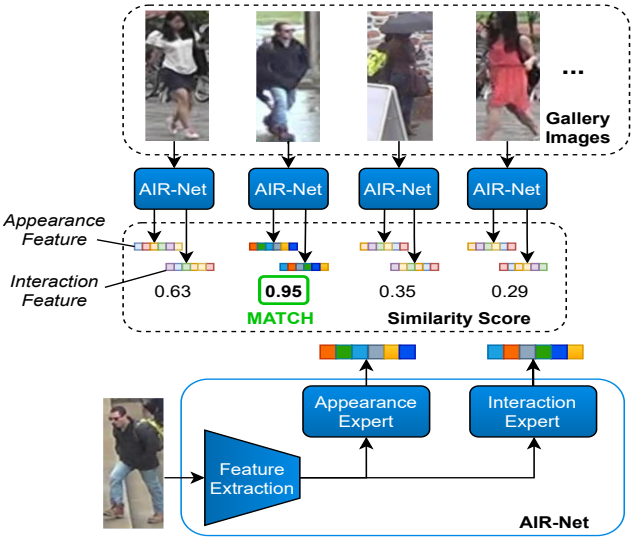


Figure 1. We propose a novel architecture with parallel branches that specialize independently as feature appearance and feature interaction experts. While the feature appearance branch builds a representation where the channels correspond to high-level semantic features, the feature interaction branch uses a novel attention-based mechanism to project the features associated with a person image to a semantic space based on the relationships between the features. Subsequently, images are matched considering extracted appearance and interaction features separately.

Traditional research focused on hand-crafted features based on color and texture [10, 1]. Recently, deep learning techniques - mainly based on convolutional neural networks (CNNs) - have been used to automatically learn features from large datasets. Global representation learning was the primary choice in Re-ID, where (i) the feature maps are extracted from query and gallery images, (ii) feature pooling is used to attain a global feature vector for each image, and (iii) matching is performed between the features of query and gallery images [30].

*Equal contribution.

Models that also extract local appearance representations that decompose the body into parts have performed consistently well in Re-ID [15, 32]. The aim of these approaches is to create more robust *local* representations that are more discriminative for overall person matching. For example, simple methods partition the feature maps in the spatial and channel dimension to create local experts specialized for specific regions or subsets of channels (loosely corresponding to semantic features) [4]. However, fixed spatial partitions may not align well with body parts due to variations in human poses and bounding box fit. Some methods have addressed this misalignment problem by using a pre-trained semantic segmentation model [15] or unsupervised clustering techniques [38] to parse body parts, but they result in more computational expense.

Another line of works has investigated how to strengthen feature representations by incorporating learning mechanisms known as attention into Re-ID models [19, 3, 5]. Recent studies have found that modeling the interactions between channels can help to selectively emphasize the most discriminative parts of the person, leading to improved performance on a number of tasks [13, 2].

By deeply investigating these methodologies, we realize that not only the strong features themselves are useful for discrimination, but that the *interaction* between the strong features also assume a vital role. We assert that this interaction information is complementary to the appearance-based features, in the same way that local appearance information supports global appearance information in part-based models. We argue that each person has a unique signature of interacting features (*e.g.*, red bag together with blue jeans), which is useful for distinguishing between identities. However, this interaction information is typically used for re-scaling [13] or mixing [2] with the original appearance features, which hides and likely limits its discriminative ability.

To investigate the capabilities of interaction features for re-identification, we propose the AIR-Net architecture. The novelty of our approach lies in the use of parallel branches that specialize independently as feature appearance and feature interaction experts, as shown in Figure 1. The first branch uses global max pooling to build a traditional appearance representation. For the second branch, we realize that the structure of existing attention modules limits their ability to match based on interaction features, and thus we introduce a novel attention-based mechanism that differs from a standard spatial attention block. While existing modules are always used to augment the original feature map by combining the appearance and interaction features [13, 2, 7], our attention module instead maintains the information related to discriminant high-level feature interactions separately. Since an expert branch devoted to analyzing the signatures of feature interaction must also build a representation of the features themselves, we also include a parallel global max pooling operation in the interaction branch. The various representations output by the branches are then used separately for the subsequent matching. In fact, we argue that there are additional benefits to matching person images directly in the semantic interaction space in addition to the traditional appearance space. We show the capability of our approach by an extensive ablation study employing separate sets of appearance and interaction features. This is also empirically proved by visualizing the saliency of features extracted by the proposed attention mechanism, which shows that the interaction representations focus on multiple areas of the person image that are more discriminant and characteristic of an identity.

To the best of our knowledge, the newly-proposed multi-branch model is the first to fully exploit the potential of attention-based interaction features that are learned automatically, allowing for more fine-grained discrimination of people in the Re-ID scenario. Our solution addresses the misalignment problem of part-based models by extracting semantically-significant regions of person images, while simultaneously reducing computational cost.

We evaluate the model on three popular benchmarks where our model achieves state-of-the-art results for Re-ID in the closed-set scenario. In summary, we have made three major contributions:

- We put forward a novel multi-branch Re-ID model that learns distinct and discriminative representations based on both semantically-significant appearance features as well as interaction of features, using parallel branches acting as specialized experts.
- We devise a new learnable attention-based pooling module as the primary component of the feature interaction branch. Unlike popular part-based models where the partitions are fixed a priori, our approach extracts semantically-discriminant regions of person images automatically. Compared with existing attention modules, ours maintains the information of strong high-level feature interactions separately, rather than fusing it with the original features. This mechanism allows the model to focus on discerning the signatures of interactions between informative parts of an image, proving to be highly effective for re-identification.
- We show that learning both appearance and interaction representations brings state-of-the-art results across three mainstream benchmarks - CUHK03, Market1501 and MSMT17.

## 2. Related Work

### 2.1. Person Re-identification

Person re-identification in the closed-set scenario addresses the problem of identifying individuals across a set

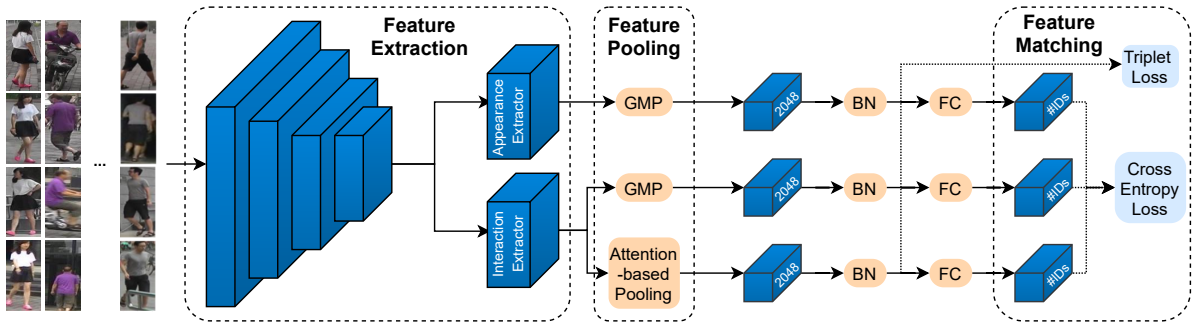


Figure 2. Our proposed Re-ID system composed of (i) feature extraction (ii) feature pooling, and (iii) feature matching. Feature extraction uses a multi-branch architecture consisting of an (a) appearance expert and (b) interaction expert. Feature pooling uses a combination of global max pooling (GMP) and our attention-based pooling. Feature matching is performed on each individual representation.

of non-overlapping cameras.

Recently, with the advent of deep learning and publicly-available Re-ID datasets of large capacity, efforts in this domain have focused on end-to-end deep learning solutions. One of the first approaches from Yi *et al*. [31] was to train a CNN with two sub-networks with a terminal cosine distance layer for determining the similarity between inputs. More recent state-of-the-art efforts however, have opted for using a backbone network pre-trained on ImageNet [8] as initial feature extraction module. An architectural approach improving the network for Re-ID is that proposed by Zhou *et al*. [37], who created a new multi-resolution bottleneck. Chen *et al*. [4] investigated the use of rough spatial and channel feature partitioning in higher layers to retain important secondary features. Martinel *et al*. [20] proposed a pyramidal network with multiple resolution learning objectives and spatial partitions to improve the network's discriminative ability.

Another Re-ID strategy using deep learning models is to incorporate semantic information of human parts using an auxiliary network. Safraz *et al*. [22] created an embedding using both fine and coarse pose information. Building upon DensePose [11], Zhang *et al*. [32] extracted dense pose information from 24 regions of a person. Zhu *et al*. [38] proposed a self pseudo-label generation method for human-part classification in order to learn more discriminative feature representations associated to human parts. While these approaches are effective, there is always a significant computational cost associated with incorporating extra semantic information, whether it be from clustering, a separate network for pose estimation or part segmentation.

We are inspired by effective part-based approaches for Re-ID that use local appearance-based representations (typically consisting of spatial and channel partitions fixed a priori) to complement global appearance information. While our method retains a global appearance expert, the local expert branches (consisting of spatial and channel partitions that are fixed a priori) are replaced with a feature interaction expert, which uses a learnable channel attention mechanism to automatically learn a complementary feature representation for discriminating between person identities. Compared with techniques that use clustering algorithms or separate pre-trained segmentation and human pose models, ours is computationally more flexible and efficient in addressing the part misalignment problem.

## 2.2. Attention-based Models

Attention-based mechanisms have become widespread and are used to tackle a variety of downstream tasks. In general, the primary objective of traditional attention modules is to rescale the original features to focus resources towards the most informative parts of images. Wang *et al*. [25] proposed Residual Attention Network that uses an encoder-decoder to predict the attention maps. Squeeze-and-Excitation network (SENet) [13] introduced a self-attention function on channels to perform feature recalibration, where global information can be used to learn to strengthen useful features. The Convolutional Block Attention Module (CBAM) used channel attention [28] that included max-pooled features in addition to average-pooled features.

Later works exploit attention to aggregate global context using long-range dependencies. The non-local neural network [26] set out to learn an attention map based on the affinity of query and key positions of an image, that was subsequently used to aggregate the features of all positions. Cao *et al*. [2] observed that the global context of non-local neural networks was actually independent of query position, and proposed Global Context Network (GCNet) that learned query-independent attention maps. Furthermore, this work generalized existing attention modules by abstracting the functional steps: (i) a context modeling stage, where the features of all positions are aggregated to form a global descriptor, (ii) a feature transform module to model the relationship between channels, and (iii) a fusion module that merges the global context feature into features of all positions. The Fully Attentional Block (FAB) [24] discarded the first step of the above general framework to maintain

some spatial information for mining useful features.

Another group of works exploit a similar mechanism to project the feature maps to a space where the interaction between the semantic features may be modeled and exploited. The global reasoning (GloRe) unit [7] was used to project the features to nodes in an interaction space where the relations between a number of disjoint regions were reasoned on using graph convolutions. The visual transformer [29] used a similar operation to extract visual "tokens" that represent semantic concepts in the image, and modeled relationships between them using a transformer.

Attention has also been successfully applied in Re-ID with the motivation that it can selectively emphasize the discriminative appearance-based image features. Harmonious Attention CNN (HA-CNN) [19] used hard region-level attention and soft pixel-level attention to enhance the convolutional response maps. Wang *et al*. [24] proposed a multi-task attention network that maintains spatial structure in addition to channel attention. Chen *et al*. [3] proposed an attention module to model complex high-order statistical information in images. Chen *et al*. [5] proposed ABD-Net, which introduced novel channel and spatial attention modules, as well as a soft orthogonalization constraint for de-correlating feature embeddings and hidden layer weights to promote diversity in feature representations. Chen *et al*. [6] introduced an architecture with layered attention mechanisms, as well as a novel module for aggregating low-level and high-level features. Zhang *et al*. [33] presented another attention module which uses learned pairwise affinity metrics between feature nodes. Fang *et al*. [9] devised a nested attention unit to be used in a bilinear attention network [16].

Different to previous works, we put forward that it is possible to create more robust representations of person images by maintaining the feature interaction information learned by an attention mechanism separate from standard global appearance-based features. To this end, we devise a new learnable attention-based pooling module that facilitates the matching directly in a semantic space that represents the relationships (*i.e*., interactions) between the features, and show that this new method improves performance compared to existing attention-based modules [2].

# 3. Method

In this section, we describe our proposed network for learning appearance and interaction representations (AIR-Net) for Re-ID. We decompose the system into a number of stages that are typical of Re-ID systems, namely (i) feature extraction, (ii) feature pooling, and (iii) feature matching between the query and gallery images, where AIR-Net operates mainly at the first two stages. In contrast with existing attention-based architectures [2, 33], we use a pair of parallel feature extractors that share early layers while maintaining separate representations for optimizing the network, as shown in Figure 2. The first branch focuses on extracting a global feature representation based on the overall person *appearance*. The second branch aims at encoding a representation based on the *interaction* between the features described by the feature map channels. This branch uses both global max pooling and attention-based pooling, where the first facilitates learning of the features themselves so that the second can analyze the interaction of these features. Finally, the images are matched separately based on the unique signatures of feature appearance and interaction.

## 3.1. Feature Extraction

Part-based approaches [4] have consistently shown strong performance for Re-ID. Separately, attention mechanisms have recently shown much promise [6, 33]. We are inspired by these methods to propose a novel multi-branch architecture where the feature extraction sub-networks aim to specialize on (i) the features representing the whole person appearance, and (ii) the interaction between strong high-level semantic features associated with the channels of the feature map. The overall architecture is shown in Figure 2. We use a ResNet-style backbone, which is typically broken down into five stages based on the sequential downsampling operations. Following recent works in Re-ID, we do not downsample the feature map at the end of the 4th stage of the network $X_0$. Rather, we duplicate the weights of the 5th stage to form two branches, which we refer to as the appearance feature extractor $F_a$ and the interaction feature extractor $F_i$. The resulting features are then given by:

$$X_a = F_a(X_0), \tag{1a}$$

$$X_i = F_i(X_0), \tag{1b}$$

where $X_a, X_i \in \mathbb{R}^{C\times H\times W}$, and $C$, $H$ and $W$ are the channel, height and width dimensions of the feature map.

## 3.2. Feature Pooling

The purpose of the feature pooling step is to capture global context of the image within the feature vector representation. Most studies use first-order pooling, such as average and max pooling, which compute the average and the maximum over individual spatial dimensions of the encoded features, respectively.

We suggest that attention-based pooling is complementary to standard global pooling methods. Recent studies on attention use global information based on the interactions between channels to selectively emphasize certain informative features [13, 2]. However, existing modules either aggregate or weight the original features with this interaction information. We suggest that the interaction information is complementary to the original features and that the individual characteristics of the alternative representations

are weakened when the features are amalgamated. Thus, we propose that distinct feature pooling operations be performed on each feature extractor branch to account for this (see Figure 2). For the feature maps $X_a$, $X_i$ from the feature extraction stage, Global Max Pooling (GMP) is performed on the feature map of the first branch (see Figure 3(a)):

$$A_1 = GMP(X_a), \tag{2}$$

where $GMP$ is global max pooling that computes the maximum over individual spatial dimensions of the coded features, and $A_1 \in \mathbb{R}^{C\times 1}$ is the appearance feature vector. GMP encourages subsequent feature matching to focus on the most discriminative part of the feature map [28]. For the second branch, the spatial dimensions of the feature $X_i$ are combined to give $\bar{X}_i \in \mathbb{R}^{C\times HW}$ before being passed to the attention-based pooling module, as depicted in Branch 2B of Figure 3(b), defined as:

$$I = (\bar{X}_i W_v)^T (\bar{X}_i W_u), \tag{3}$$

where learnable weights $W_u \in \mathbb{R}^{C\times 1}$ are used to collapse the channel dimension to a single strong semantic feature, and weights $W_v \in \mathbb{R}^{C\times C}$ model the channel interdependencies. The resulting interaction feature vector $I$ is a linear combination of the original features. Existing attention modules [13, 2] would mix this interaction information with the appearance-based features (see Figure 3(c)):

$$M = f(X_i, I), \tag{4}$$

where $M$ is the mixed feature vector, and $f$ is a fusion function such as aggregation or weighting.

However, we suggest that it is beneficial to maintain the attention-based global representations separately for subsequent matching. The original features $X_i$ are also pooled using another GMP operation, to give a second appearance feature vector $A_2 = GMP(X_i)$, that may differ from $A_1$ as a result of the weights in the second branch that are not shared (see Branch 2A of Figure 3(b)). This is consistent with existing part-based approaches where global pooling is included in every local expert branch [4]. This results in multiple complementary global feature vectors, that combine the unique advantages of the conventional global max pooling and attention-based pooling mechanisms. The ability to attend to different locations of the input enables the second branch to localize discriminative features automatically compared with hand-crafted partitions. Compared with a single fused feature representation, the individual differences between the alternative representations are more easily discerned during matching, improving the network's discriminative ability.

### 3.3. Feature Matching

We claim that there are benefits to matching person images directly in the interaction space, in addition to the traditional appearance space. Hence, the individual contributions of the feature representations - computed from the previous feature extraction and pooling steps - are considered during subsequent identification and metric learning, as shown in Figure 2.

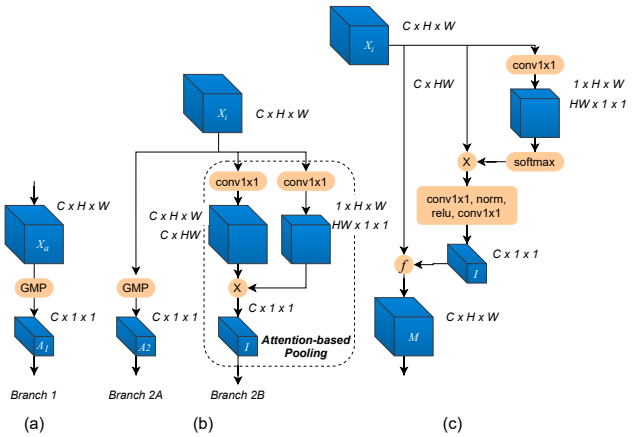


Figure 3. Comparison of (a) feature pooling of appearance feature extractor (Branch 1), (b) feature pooling of interaction feature extractor (Branch 2A & 2B), and (c) GC block [2] that fuses the interaction feature descriptor with the original features.

Given an input image $x_j$ with label $y_j$, the predicted probability of $x_j$ being recognized as class $y_j$ is encoded with a softmax function, represented by $p(y_j|x_j)$. The identity loss is then computed by the cross-entropy loss:

$$L_{ce} = -\sum_{j=1}^{N} y_j log(p(y_j|x_j)), \tag{5}$$

where $N$ is the number of training samples in each batch.

For the triplet loss, we consider the positive and negative batch-hard choice [12]. We generate batches by randomly sampling $P$ classes of human identities, and $K$ images of each class. The triplet loss is given by:

$$L_{tri} = \sum_{k=1}^{P}\sum_{j=1}^{K} max(0, D_{max}(F(x_j^k), F(x_p^k)) + m + \\ -D_{min}^{k\neq w}(F(x_j^k), F(x_n^w))), \tag{6}$$

where $x_j^k$ represents a data sample corresponding to the $j$-th image of the $k$-th person in the batch, $x_p^k$ and $x_n^w$ indicate the positive and negative samples in each batch (with $k \neq w$), $m$ denotes the triplet distance margin, and $D_{max}$ and $D_{min}$ are defined to consider the hard pairs as the maximum and minimum cosine distances between the representations $F$ (*i.e.* $A_1$, $A_2$ or $I$) of $x_j^k$ and the positive and negative samples in the batch, respectively. The full loss is given by:

$$L_{tot} = \frac{1}{N_{ce}}\sum_{j=1}^{N_{ce}} L_{ce} + \frac{1}{N_{tri}}\sum_{j=1}^{N_{tri}} L_{tri}, \tag{7}$$

| Method | CUHK03 | | | | Market1501 | | MSMT17 | |
|---|---|---|---|---|---|---|---|---|
| | Labeled | | Detected | | | | | |
| | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP |
| ABD-Net (ResNet-50) [5] | - | - | - | - | 95.6 | 88.3 | 82.3 | 60.8 |
| PyrNet (DenseNet) [20] | 71.6 | 68.3 | 68.0 | 63.0 | 93.6 | 81.7 | - | - |
| BAT-net (GoogLeNet) [9] | 78.6 | 76.1 | 76.2 | 73.2 | 95.1 | 87.4 | 79.5 | 56.8 |
| SCR (ResNet-50) [4] | 84.8 | 81.4 | 82.2 | 77.6 | 95.7 | 89.0 | - | - |
| RGA-SC (ResNet-50) [33] | 81.1 | 77.4 | 79.6 | 74.5 | **96.1** | 88.4 | 80.3 | 57.5 |
| SCSN (ResNet-50-CBAM) [6] | **86.8** | 84.0 | **84.7** | 81.0 | 95.7 | 88.5 | 83.8 | 58.5 |
| ISP (HRNet) [38] | 76.5 | 74.1 | 75.2 | 71.4 | 95.3 | 88.6 | - | - |
| AIR-Net (ResNet-50) | 85.8 $\pm$ 0.3 | 83.2 $\pm$ 0.1 | 83.0 $\pm$ 0.1 | 80.4 $\pm$ 0.1 | 95.2 $\pm$ 0.1 | **89.3** $\pm$ 0.2 | 84.2 $\pm$ 0.3 | 63.8 $\pm$ 0.1 |
| AIR-Net (IBN-ResNet-50) | **86.8** $\pm$ 0.6 | **84.8** $\pm$ 0.4 | 84.3 $\pm$ 0.8 | **82.0** $\pm$ 0.4 | 95.2 $\pm$ 0.1 | **89.3** $\pm$ 0.1 | **84.7** $\pm$ 0.1 | **64.5** $\pm$ 0.2 |

Table 1. Performance (%) comparisons between our approach and the state-of-the-art methods on CUHK03, Market1501 and MSMT17 datasets. For our AIR-Net strategy, we report the average and standard deviation results over three runs. The adopted backbone is reported in brackets for each approach. The **top** results are highlighted in bold for each metric.

where the number of summed feature representations for cross-entropy and triplet losses $N_{tri} = N_{ce} = 3$.

# 4. Experiments

In this section, we first introduce the Re-ID benchmark datasets that are used to evaluate our proposed approach. The architecture and setup considered in our experiments are then described. Finally, the results are discussed along with an ablation study showing each individual contribution to the overall performance by our proposed multi-branch architecture. Saliency map visualizations are also provided to better interpret the results.

## 4.1. Datasets and Protocols

We consider the most popular ReID bechmark datasets following recent works [9, 33, 6].

**CUHK03** [18]: Following the CUHK03-NP (New Protocol), this dataset consists of 14,097 images of 1,467 people from 10 different cameras divided into a training set of 767 individuals and a testing set of 700 individuals. The dataset provides two types of annotations The labeled version consists of 7,368 training images, 1,400 query images and 5,328 gallery images. The detection version includes 7,365 training images, 1,400 query images and 5,332 gallery images.

**Market-1501** [34]: This dataset contains 32,668 labeled images of 1,501 individuals in total acquired by 6 different cameras. The dataset is is divided into a training set of 12,936 images of 751 individuals, and a test set of 19,732 images of 750 people (with 3,368 query images and 16,364 gallery images).

**MSMT17** [27]: This is the most recent, challenging, and largest publicly-available person Re-ID dataset. It includes 126,441 images of 4,101 identities captured by 15 different cameras, considering both outdoor and indoor scenarios. It is divided into a training set of 32,621 images of 1,041 individuals, and a test set of 93,820 images of 3,060 people (with 11,659 query images and 82,161 gallery images).

Following common practice in the Re-ID problem, we use the mean average precision (mAP) and the cumulative matching characteristics (CMC) at Rank-1 to evaluate the performance of our proposed method.

## 4.2. Architecture and Setup

**Network**: We test our approach using both ResNet-50 and IBN-ResNet-50 [21] backbones pre-trained on ImageNet [8]. The latter is a simple variant of ResNet-50 with an instance normalization module. For ResNet-50 we add global context (GC) blocks [2] after the third and fourth stages of the backbone to facilitate a fairer comparison with other state-of-the-art models that use similar non-local attention units in the backbone [6, 33]. We change the stride of conv4_1 from 2 to 1 to exploit larger spatial information in the feature maps. The 5th stage of the backbone (the layers after conv4_1) is duplicated to give 2 independent branches that do not share weights. After performing a combination of global max pooling and attention-based pooling, we apply batch normalization [14] and a fully connected layer in each branch.

**Training**: We resize all images to $384 \times 192$ pixel resolution. Following recent methods, we deploy random horizontal flipping and random erasing [36] on the images for data augmentation. The batch size is set to 64 with a random selection of 16 identities and 4 images for each identity. We adopt the Adam optimizer [17] to train all models for 200 epochs with a weight decay of $5 \times 10^{-4}$. The initial learning rate is set to $3.5 \times 10^{-4}$, which is reduced to $3.5 \times 10^{-5}$ after 40 epochs and $3.5 \times 10^{-6}$ after 70 epochs. Each branch of our proposed model is optimized with cross-entropy and triplet losses. The margin of the triplet loss $m$ is set to 0.2. During evaluation, we concatenate all the feature vectors after the final batch normalization layers as the attention-based global representation for images in query and gallery sets. We run each experiment three times and report the average results with standard deviation.

| Model | # parameters |
|---|---|
| ABDNet (ResNet-50) [5] | $\sim$ 85M |
| SCR (ResNet-50) [4] | $\sim$ 53M |
| RGA-SC (ResNet-50) [33] | $\sim$ 42M |
| SCSN (ResNet-50-CBAM) [6] | $\sim$ 110M |
| ISP (HRNet) [38] | $\sim$ 44M |
| Baseline (IBN-ResNet-50) | $\sim$ 24M |
| **AIR-Net** (IBN-ResNet-50) | $\sim$ **42M** |

Table 2. Comparison of total number of parameters for AIR-Net and recently proposed models for Re-ID in closed-set scenario. The adopted backbones are reported in brackets for each approach.

### 4.3. Comparison with State-of-the-Art Methods

We compare the performance of our proposed AIR-Net with recent state-of-the-art methods across three popular datasets [1] in Table 1. We note that the compared works use a variety of backbones including ResNet-50, ResNet-50-CBAM, DenseNet, GoogLeNet and HRNet. We report results for both the simpler and the more powerful backbone - ResNet-50 and IBN-ResNet-50 models, respectively. These results demonstrate that our method outperforms the current state-of-the-art in mAP across all datasets. With respect to the CMC metric at Rank-1, AIR-Net achieves the highest result for CUHK03 Labeled and MSMT17, and just shy of the top performances for CUHK03 Detected and Market1501. We highlight that our approach tends to achieve the largest improvements on less-saturated metrics of the datasets, with an increase in mAP of 3.7% for the most recent, largest, and most challenging MSMT17 dataset. This suggests that AIR-Net may be even more beneficial for datasets with a larger amount of images in the training set. Our results with ResNet-50 backbone still achieves competitive results while outperforming the other methods which fairly utilize the same backbone.

While SCSN [6] achieves the second-top performance overall (with the top and joint-top Rank-1 results on two datasets), we note that this architecture uses a multi-faceted approach that includes spatial and channel attention within the backbone, non-local multi-stage fusion between the feature maps of different stages and a cascaded multi-stage suppression sub-network that is designed to iteratively capture less prominent features. In particular, we argue that the post-processing of the feature representations by the multi-stage suppression sub-network is uniquely suited to improving Rank-1 score. In addition, SCSN contains almost 3 times the number of parameters as shown in Table 2, which compares the size of our model with the most recent approaches proposed in the Re-ID closed-set scenario. Our multi-branch attention-based approach adds a relatively small overhead on top of the baseline model and contains fewer total parameters than other state-of-the-art models, while achieving the top results across all datasets. This further shows the validity of our approach that takes account of both feature appearance and interaction information.

[1]While we are aware that DukeMTMC-reID [35] is no longer publicly available due to improper collection procedures, we test our approach on it for completeness. Here also, AIR-Net outperforms the current state-of-the-art [4] in both metrics, obtaining 82.0% mAP and 91.2% in CMC at rank-1 as compared to 81.4% and 91.1%, respectively.

### 4.4. Ablation Study

In this section, we evaluate the effectiveness of using multiple expert branches and maintaining the appearance-based and interaction-based feature representations separately for subsequent matching. Furthermore, we compare the performance of our attention-based pooling approach with a popular existing attention module [2]. The ablation study is reported in Table 3. Starting with an IBN-ResNet-50 performance as a baseline (line 1), we replaced the global max pooling (GMP) with ABP to directly compare attention-based pooling versus standard global pooling. We find that the ABP block brings the highest performance increase when added after the 5th stage of the backbone. We note a consistent improvement in line 2, as the attention-based interaction features are learned rather than handcrafted. We then analyze the effects of incorporating a standard attention-based module - the GC block [2] - in a single branch architecture, followed by GMP (line 3). Since the GC block aggregates information based on the interactions between channels, we see a performance increase compared to the baseline and ABP block.

Secondly, we investigate the use of a multi-branch strategy [4], where the local part-based experts are replaced by an existing attention module [2], used to automatically aggregate information from discriminative regions of person images. A further improvement is observed compared to the single-branch variants, demonstrating the importance of multiple branches to create complementary features (line 4). However, the performance remains sub-optimal since the existing attention block fuses the learned feature interaction information with the original appearance features. We analyze the replacement of the existing attention block with our attention-based pooling module, which shows an overall improvement over all the datasets. This suggests that using feature interaction information independently in the second branch is beneficial (line 5).

Nonetheless, we argue that an expert branch devoted to analyzing the signatures of feature interaction must also build a representation of the features themselves. Hence, we add another GMP operation in the second branch (Branch 2A) to facilitate this learning. We complete the ablation with the last comparison in lines 6 and 7, which add the GC and ABP blocks in Branch 2B respectively. The complete AIR-Net shows a large improvement compared with the other variations, suggesting that discriminating between identities using a unique representation based on feature interaction brings a clear advantage in performance (line

| Variants | CUHK03 | | | | Market1501 | | MSMT17 | |
|---|---|---|---|---|---|---|---|---|
| | Labeled | | Detected | | | | | |
| | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP |
| 1 - Baseline | 74.0 ± 0.4 | 71.9 ± 0.2 | 70.3 ± 0.5 | 68.7 ± 0.5 | 94.7 ± 0.3 | 86.3 ± 0.1 | 73.4 ± 0.2 | 48.4 ± 0.3 |
| 2 - Baseline with ABP | 74.9 ± 0.4 | 72.0 ± 0.1 | 72.3 ± 0.8 | 69.7 ± 0.8 | 94.4 ± 0.2 | 86.0 ± 0.3 | 76.9 ± 0.2 | 52.4 ± 0.3 |
| 3 - Baseline with GC | 77.2 ± 0.2 | 74.6 ± 0.2 | 73.4 ± 0.2 | 71.4 ± 0.1 | 95.2 ± 0.3 | 87.3 ± 0.2 | 77.1 ± 0.2 | 53.0 ± 0.2 |
| 4 - // branch 1 GMP + GC | 83.8 ± 0.4 | 82.3 ± 0.4 | 81.1 ± 0.5 | 78.8 ± 0.5 | **95.3** ± 0.2 | 88.6 ± 0.1 | 79.9 ± 0.1 | 56.5 ± 0.1 |
| 5 - // branch 1 GMP + ABP | 84.8 ± 0.7 | 82.7 ± 0.6 | 80.9 ± 0.6 | 78.9 ± 0.6 | 95.1 ± 0.1 | 89.2 ± 0.1 | 83.0 ± 0.2 | 61.3 ± 0.2 |
| 6 - // branch 2 GMP + GC | 84.9 ± 0.4 | 82.8 ± 0.2 | 80.5 ± 0.7 | 78.6 ± 0.4 | 95.2 ± 0.1 | 88.6 ± 0.1 | 82.5 ± 0.1 | 60.9 ± 0.1 |
| 7 - AIR-Net | **86.8** ± 0.6 | **84.8** ± 0.4 | **84.3** ± 0.8 | **82.0** ± 0.4 | 95.2 ± 0.1 | **89.3** ± 0.1 | **84.7** ± 0.1 | **64.5** ± 0.2 |

Table 3. Ablation study for evaluating the effectiveness of using multiple expert branches and comparing existing attention modules - the global context (GC) block - with our attention-based pooling (ABP) module. All variants use the IBN-ResNet-50 backbone. The average and standard deviation results over three runs are reported.

7). We highlight the large improvement gained by AIR-Net over the standard baseline for all the considered datasets (*e.g.*, a 16.3% increase in mAP for MSMT17).

### 4.5. Visualizations of Saliency Maps

To provide a better interpretation of our proposed model, we use the popular Grad-CAM [23] saliency approach to visualize where our method is focusing on the images. Grad-CAM can highlight image cues that the network considers important for a specific prediction. We qualitatively compare the saliency visualizations of the baseline and our model in Figure 4. In particular, we select samples from the MSMT17 dataset and visualize each individual branch for our AIR-Net. In general, the saliency maps for feature of Branch 1 and Branch 2A are similar to the baseline, since the same GMP operation is used. This is expected since the primary purpose of Branch 2A is to facilitate learning of feature representations so that their interactions can be modeled in Branch 2B. Nonetheless, slightly larger regions of activation are noticeable for Branch 2A (*e.g.*, the area near the head of the person for the last three samples), owing to the different weights learnt in the parallel branch. Interestingly, the feature interaction branch Branch 2B appears to consistently focus on multiple discriminative small regions of the person (*e.g.*, head, shoes, hood, suitcase) at the same time. This reinforces the idea that this branch specializes in the interaction between strong features. The difference in saliency extracted from Branch 2B compared with those of Branch 1 and Branch 2A supports the fact that the parallel feature representations of AIR-Net extract complementary information from the images.

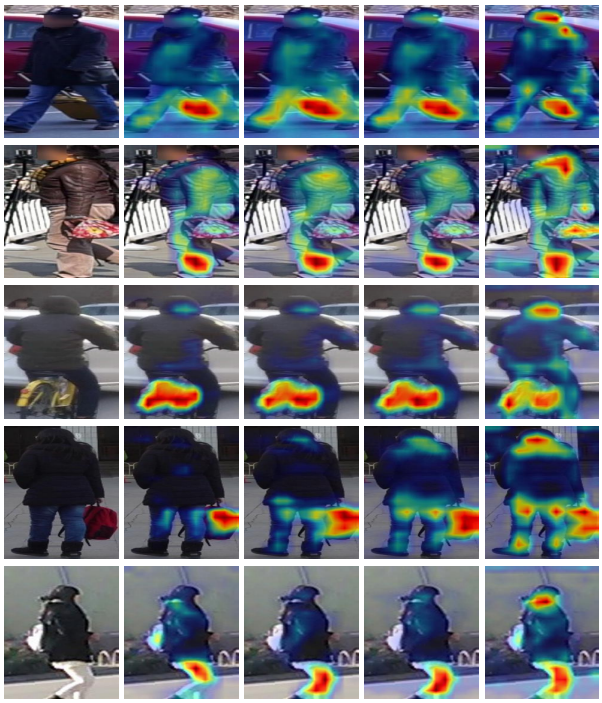


Figure 4. The first column shows the original images from the MSMT17 dataset and the remaining columns show the Grad-CAM saliency visualization for our baseline and proposed AIR-Net, fine-tuned on the MSMT17 dataset. AIR-Net-*n* is the saliency visualization of the *n*-th branch of our model. By individually optimizing the branches, AIR-Net is able to jointly consider appearance-based features of Branches 1 and 2A and interaction-based features of Branch 2B. Notably, the feature interaction Branch 2B appears to consistently focus on multiple discriminative regions of the person (*e.g.*, head, shoes, hood, suitcase).

## 5. Conclusions

We proposed a novel multi-branch architecture, composed of parallel appearance and interaction feature extractors, for person Re-ID. We argue that existing attention modules are sub-optimal since the interaction information is mixed with the original features, reducing the discriminative ability of the network. Hence, we designed an attention-based pooling module that maintains the global representation of feature interaction for subsequent matching between person images. We performed rigorous evaluation and ablation studies, finding that our method outperforms recent state-of-the-art models on the popular public datasets, while also reducing computational expense.

## References


[1] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino. Person re-identification by discriminatively selecting parts and features. In *Eur. Conf. Comput. Vis. Worksh.*, 2015.

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Int. Conf. Comput. Vis. Worksh.*, 2019.

[3] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Int. Conf. Comput. Vis.*, 2019.

[4] Hao Chen, Benoit Lagadec, and Francois Bremond. Learning discriminative and generalizable representations by spatial-channel partition for person re-identification. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020.

[5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. ABD-Net: Attentive but diverse person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[6] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[7] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

[9] P. Fang, J. Zhou, S. Roy, L. Petersson, and M. Harandi. Bilinear attention networks for person retrieval. In *Int. Conf. Comput. Vis.*, 2019.

[10] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Eur. Conf. Comput. Vis.*, 2008.

[11] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

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

[15] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[16] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Adv. Neural Inform. Process. Syst.*, 2018.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[20] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Aggregating deep pyramidal representations for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

[21] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Eur. Conf. Comput. Vis.*, 2018.

[22] M. Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[24] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Eur. Conf. Comput. Vis.*, pages 365–381, 2018.

[25] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[27] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Eur. Conf. Comput. Vis.*, 2018.

[29] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[30] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Int. Conf. Pattern Recog.*, 2014.

[32] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[33] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[34] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, 2015.

[35] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[36] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.

[37] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Int. Conf. Comput. Vis.*, 2019.

[38] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Eur. Conf. Comput. Vis.*, 2020.