Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12732-3

Consistent attentive dual branch network for person re-identification

Asad Munir[1] · **Niki Martinel**[1] · **Christian Micheloni**[1]

Received: 13 January 2021 / Revised: 17 May 2021 / Accepted: 21 February 2022
© The Author(s) 2022, corrected publication 2022

Abstract

Several recent person re-identification methods are focusing on learning discriminative representations by designing efficient metric learning loss functions. Other approaches design part based architectures to compute an informative descriptor based on local features from semantically coherent parts. Few efforts learn the relationship between distant similar regions and parts by adjusting them to their most feasible positions with the help of soft attention. However, they focus on calibrating distant similar parts features and ignore to learn the noise (blur) free and distinct feature representations as the person re-identification datasets contain degraded images. To tackle these issues, we propose a novel Consistent Attention Dual Branch Network (CadNet) that has ability to model long-range dependencies (correlations) between channels as well as feature maps. We adopt multiple classifiers trained to learn the most discriminative global features for a unique representation of a person. Correlation between channels are consistently computed by using channel attention mechanism to make the learned feature noise free and distict from noisy and blurry data. Feature correlations interpret the relationship between distant similarities in the images computed by the self attention mechanism. The proposed CadNet significantly enhances the performance with respect to the baseline on the person re-identification benchmarks.

Keywords Person Re-Identification · Attentive networks · Person search

1 Introduction

The Problem of Person re-identification (re-id) is, given a probe, to retrieve person's images from gallery sets acquired by the same or other cameras. The task of re-id is an essential

✉ Asad Munir
asad.munir@uniud.it

Niki Martinel
niki.martinel@uniud.it

Christian Micheloni
christian.micheloni@uniud.it

[1] Department of Computer Science, University of Udine, Udine, Italy

Published online: 19 March 2022

component of intelligent surveillance systems [4, 22] and its importance is progressively improving in research. The presence of highly variable factors like illumination, resolution, clothing, view angle, human pose, occlusions and background in the images make re-id a very challenging task.

Another complication with respect to classical classification tasks is that in re-id the identities set (classes) mismatch between the training and the testing stages. Precisely, image classification task has similar classes (i.e. person identities) in training and testing sets while re-id has different identities in both sets. Therefore, the task of re-id requires a strong and discriminative feature descriptor to distinguish unseen, in the testing set, similar images belonging to new identities.

With the development of neural networks and deep learning algorithms, the ConvNets [11, 16, 32], originally well designed for image classification tasks, perform impressively by providing a discriminative feature representations for person images. Such a representation capability outperforms the traditional handcrafted low-level features by a large margin. To exploit ConvNets in Re-Id solutions, a research trend aims to design better metric learning loss functions [6, 13, 21] such as triplet loss, triplet hard loss, quadruplet loss, etc. for a better description of the person's image. These loss functions enlarge and reduce the inter-class and intra-class variations respectively, thus improve the generalization capability of the model. The performance of such metric learning based loss functions is highly controlled by the sampling method and by hard sample mining techniques. On the other hand, many approaches [23, 42, 52] address the person re-identification task as a general image classification task. The basic idea of these studies is computing the cross entropy softmax classification loss for the person's images. While testing, these classification based approaches compute the distance matrix from the output features of the images to distinguish the person identities. Due to the mismatch between training targets and testing targets, the performance of metric learning loss functions becomes inferior in re-id task. To overcome this issue, we propose a multi classifiers training instead of a general single classifier to learn the most discriminative features from person images. The effectiveness of the proposed multi classifiers learning is presented in the ablation study section of the paper.

Recently, part based models [30, 34, 44, 46–48, 53] have represent the state of art performance in person re-id by learning part based local feature representations from the person's image. Some of these methods [34, 47] compute a strong discriminative representation of the image by splitting it into several body parts and then evolve the local features from all the parts into a single representation. Other approaches [44, 48] horizontally partition the deep neural networks feature maps into several parts to learn more informative and fine-grained salient features in individual local parts. They distinguish one identity from an other by using discriminative cues from these parts. To learn salient part features, such methods require well aligned body parts for the same person. This is one of the main drawbacks due to lack of part consistency.

Lately, many attention-based approaches [5, 18, 19, 25, 39, 40, 50, 53] have been proposed to overcome such partitioning and misalignment issues occurring in part based techniques. Attention is a powerful tool to perform spatial localization in the neural networks to interpret their decisions. AACN [40] tackle the misalignment and occlusions issues that occur in the re-identification tasks by masking out the undesirable background with pose guided attention mechanism. Others [19, 50] focus on better matching features and essential attention regions by learning superior attention maps. Self-attention helps to compute the features correlations [43] by providing more weight to similar parts in the image and modeling long-range dependencies in a statistically efficient way. However, the drawbacks of these approaches are the lack of learning the key-part features due to randomness part selection and in considering noise (e.g. blur) effect in the learned features since most of the re-identification images are blurry and noisy.

To overcome aforementioned issues, thus to enhance the final matching relevant region features should be computed as well as the feature correlation. For such a purposes we propose the introduction of a self attention module. In addition, to address the issue of noise, e.g. blur, thus to learn noise free features, e.g. features learned from sharp patches, we propose a consistent computation of channels correlation of multi scale features by exploiting the channel attention module. The proposed Consistent Attention Dual Branch Network (CadNet) is a modified version of the self and channel attention network (SCAN) [24] work such that noise free, salient and discriminative features are learned. The overall scheme of the proposed approach is shown in Fig. 1 and our contributions with respect to the SCAN [24] are:

- The channels correlation are used at the end of each stage instead for every residual connection, thus reducing their number. This is motivated by the hypothesis that the features computed at the end of each stage are a better representation of the image. Such an hypothesis is experimentally supported by the fact that computing the channels correlation at this stage enhances the performance and reduces the computational cost. In addition, these inter dependencies make the learned features robust to noise (e.g. blur) thus contributing to improve the matching score.
- A dual branch mechanism composed by a residual and an attentive branch is introduced. The former aims to provide *noise free* features while the latter provides similarities between patches at different location in the matching images (e.g. a backpack carried by hand in the probe and on shoulders on the gallery images). The final representation is the concatenation of both branches.

With the above mentioned changes, the performance of CadNet significantly improves on person re-id benchmarks as compared to SCAN [24] as well as to other state of the art methods.

2 Related work

2.1 Metric and classification losses

Deep neural networks, compared to hand-crafted features, perform better by learning the required features and metrics from the data using suitable loss functions. Ding et al [10] proposed a triplet loss to compute the relative distance between different images. Chen et al. [6] adopted quadruplet loss to enlarge inter-class variations and reduce intra-class ones. Yu et al. [41] proposed a soft hard sample mining technique which assigns weights to hard samples. Many research efforts address the person re-id problem as an image classification problem. Some [1, 17] compute the cross entropy loss for the images pairs by taking paired images as input. Others, like [34, 35], propose margin based losses or adopt simple classification losses for each part by splitting an image into multiple parts.

Unlike all the above methods, we compute multiple cross entropy losses from each added classifier allowing to achieve significantly higher performance than those of a standard single classifier. These multiple losses are then used along with triplet loss.

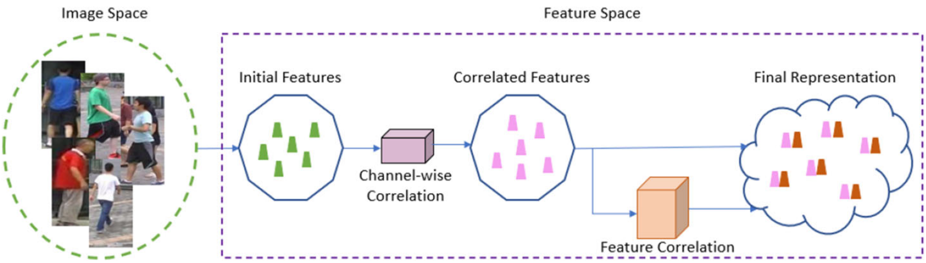

Fig. 1 The explanation of the mechanism of the proposed approach

2.2 Part based deep neural networks

Different works [30, 34, 44, 46–48] introduced local part based feature representations to enhance the re-id performance. Some methods perform inaccurate part localization by directly splitting images into local stripes. Other approaches deal with the alignment of these local parts by pose estimation and region proposal generation. Zhao et al. [48] proposed a part-aligned network for a better partitioning of the body parts. Sun et al. [34] introduced a uniform partition strategy to partition the person image into horizontal stripes. Zhang et al.[44] computed local and global losses by partitioning the body parts into horizontal stripes. Shu et al. [30] learned part based features by dividing images into parts and used a network to assign weights to each part.

Unlike these part based methods, attention based methods directly learn local features correlation from the image data. There is no need to split the image into stripes because the attention method computes the correlations between features patches and hence provides better relationship between different parts among the images.

2.3 Attention based models

Attention-based methods are used to manage localization and misalignment issues in the images. Liu et al. [19] proposed the HydraPlus network which provides better representation of the images by learning low-level attentive features. Li et al. [18] proposed a network for a multi-scale feature representation which simultaneously learns hard region-level and soft pixel-level attentive features. Other approaches [40, 50] focused on better matching the features and proposed attentive learning to learn appropriate attention maps. Xiang et.al [39] proposed a scheme to fuse feature from multiple regions and use soft attention to assign weights for each region. Zhong et al. [53] solved the alignment problem by using attention mechanism to mix global-local features for more stable pedestrian descriptors. Unlike these works, SCAN [24] introduced the self and channel attention mechanisms to weight more the some features and to make them noise free by constantly computing the channels correlation to avoid noise effects. These improve the features matching when the images have similarities at different locations. SCAN focuses on attentive representations and ignores the original features which may still have useful information.

In this work, we modified the SCAN [24] model by applying a dual branch mechanism which preserves the original features along with attentive representations. We also improve the channels correlation by computing them at key locations in the network.

3 Consistent attentive dual branch network

3.1 Problem definition and notations

Let a set of n training images $\{I_i\}_{i=1}^{n}$ with corresponding identities labels $\{y_i\}_{i=1}^{n}$ be acquired by a camera network. The task of person re-identification is to, given a probe, retrieve the person's images from the galleries of different cameras. The problem of person re-id is usually treated as an image classification task when using cross entropy loss. The difference between these two tasks is that training and testing classes (person identities) are identical in image classification while different in re-id. With the help of classifiers, the most discriminative features for each person are learned from the dataset. During testing, these features are used to compute the distance matrix between the probe and the persons identities to achieve the person re-identification.

3.2 Overview

To make a better matching in person re-id, strong and discriminative feature representations of person's images are required by the neural networks. To learn such representations, neural networks are trained in a supervised fashion by using data of persons with known identities. In the testing stage, the features of unseen persons are extracted to match with other unknown persons. The presence of unknown identities significantly reduces the matching performance. The training mechanism of neural networks plays an important role to learn from the person's image specific things (cloths, handbags, etc) that are important features to disambiguate between different people. We propose a training mechanism that exploits 4 classifiers. The predictions from all the classifiers are merged to make the final decision. We name it multi classifier (Multi-C) training. During training, several convolution operations takes place across multiple channels of the features produced by the network's layers. The final output of network's layers are the sum through all the channels. This induces channel dependencies in the learned features. Such channel dependencies cause to miss the tiny effective details in the output features especially in case of person re-id since the images are blurry and noisy. We compute channels correlation to enhance the convolution features at every stage of the network so that the network is able to increase its sensitivity to missing informative features due to degraded data. Another aspect is that the convolution operations have just local information, hence they miss the long range similarities present in the images. These long range similarities has an essential impact in re-id when matching images. We capture these similarities with features correlations which can be exploited by self attention mechanism. The details of computing channel and feature correlations are explained in the next sections.

3.3 Proposed network architecture

Recent research works have shown that Convolutional Neural Networks (CNNs) efficiently learn deeper and robust feature representations from images and are precise to train if they have shorter connections between layers. Leveraging on such outcomes, we adopt the ResNet-50 [11] as our backbone network with several additions and modifications. We adopt multi classifiers training with multiple fully connected layers which are shown as classifiers in Fig. 2 to predict the identity of the person in the input probe image. The gradients from all added classifiers are gathered at the previous 1×1 convolution layer and force that layer to learn the most discriminative global features. Such features are used to compute

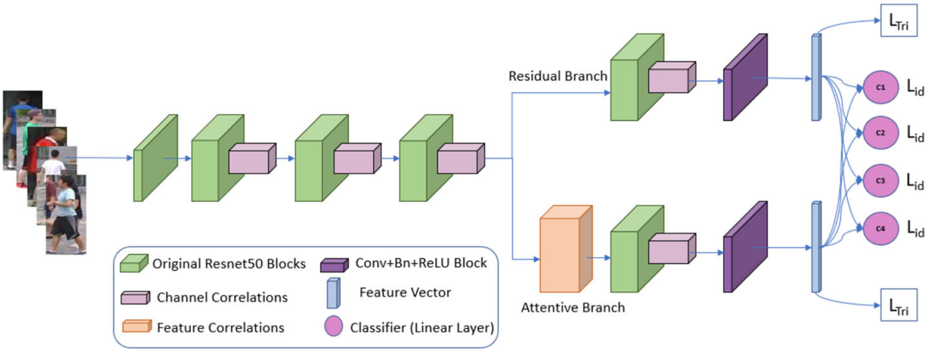

Fig. 2 Overview of the proposed Network. $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the identity loss. Features from both residual and attentive branches are fed to multiple classifiers having shared weights

the distance matrix to overcome the issue of identity mismatching in the testing stage. Since the convolution layers have a local receptive fields then the learned global features depend on the local neighbourhood similarities and ignore the long-range dependencies. To capture the similar parts at different regions in the image and to work with re-id degraded data, we compute feature and channels correlation with the help of self and channel attention mechanisms to learn noise free and salient features. These two modifications are expressed as channels correlation and feature correlations in Fig. 2. After the third stage of the backbone network [11] we designed two branches named residual and attentive branches. We added the self attention module at the start of the attentive branch because self attention produces better results when the spatial size of features is small [43]. Channels correlation are computed after every block of the ResNet-50. The details for computing channel and feature correlations are explained in the next sections. The resulting proposed network (CadNet) is shown in Fig. 2 and is trained with cross entropy losses (L_{id}) from all classifiers and the triplet loss.

3.4 Channels correlation

Since person re-identification is applied to surveillance cameras, commonly real scenarios and used datasets consist of blurry and noisy images. Most of the existing methods are unable to grasp deep salient features from them. To build a stronger descriptor against such a degradation, noise free and distinct feature learning is required. To fulfill this objective, we introduce several channel attention modules to compute channels correlation consistently during the feature learning process.

Let $K = [k_1, k_2, ..., k_C]$ be the learned set of filter kernels for C output channels with k_l being the parameters of the l^{th} filter in a general convolution operation. The output from this convolution operation can be written as $U = [u_1, u_2, ..., u_C]$, where

$$u_l = k_l * X = \sum_{n=1}^{C'} k_l^n * x^n \tag{1}$$

In the above equation, $k_l = [k_l^1, k_l^2, ..., k_l^{C'}]$, $X = [x^1, x^2, ..., x^{C'}]$ (X being the input feature maps and C' is the number of input channels). The convolution operation is denoted by $*$ and 2D spatial kernel k_c^n represents a single channel of k_l which interacts with the corresponding channel of X. The output of the convolutional layers is obtained through a channel-wise sum of the computed feature values. Therefore, the channel dependencies are introduced along with the spatial correlation captured by the convolutional filters in the learned weights. We follow the work in [15] for computing these channel dependencies (correlations) but apply them at compacted features (convolutional block) instead at residual connections (used in [15]).

Each unit of the output U is unable to exploit contextual information outside of its region because the convolution operation has a local receptive field. To resolve this issue, global spatial information is squeezed into a channel descriptor. This operation generates channel-wise statistics and is achieved by using global average pooling. A statistic $z \in \mathbb{R}^C$ can be generated by shrinking U through the spatial dimension $H \times W$. The l^{th} element of z, computed by global average pooling, can be written as:

$$z_l = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_l(i, j) \tag{2}$$

For better modeling channel-wise dependencies, the learned function must have the ability to capture the nonlinear interaction between channels and permit multiple channels to oppose one-hot activation. The sigmoid activation fulfills these requirements and can be written as:

$$n = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{3}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ are the parameters of the dimensionality reduction layer and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are the parameters of dimensionality-increasing layer while δ denotes the ReLU function and r is the reduction ratio [1]. Two 1 × 1 convolution layers implement W_1 and W_2 around the non-linearity. The final output of the channel attention is obtained by rescaling the output U by means of the activations:

$$\bar{x}_l = n_l \cdot u_l \tag{4}$$

where $\bar{X} = [\bar{x}_1, \bar{x}_2, ..., \bar{x}_C]$. The dot product refers to channel-wise multiplication of feature maps $u_l \in \mathbb{R}^{H \times W}$ and the scalar n_l. The overall operation of the channel attention for computing channels correlation is shown in Fig. 3 and it helps to boost feature discrimination.

3.5 Feature correlations

In the existing works, most of the designed models for person re-id consist of convolutional layers. These convolutional layers are computationally unable to model long-range dependencies and distant similarities in the images because the convolution operation computations are bound to local neighbourhoods. To efficiently model the relationships between widely separated spatial regions, we adapt a non local model [38] to compute features correlations with self-attention in a convolutional framework. The introduction of this module builds a strong descriptor along with the original feature and hence improves the matching by capturing similarities at different image locations. The image features from previous hidden layers $x \in \mathbb{R}^{C \times N}$ are first split into two feature spaces f, g with the help of 1×1 convolution layers. C is the number of channels and $N = H \times W$. The attention is computed such that $f(x) = W_f x$ and $g(x) = W_g x$ with learnable weight matrices

[1] Please refer to the experiments section for the evaluation of r

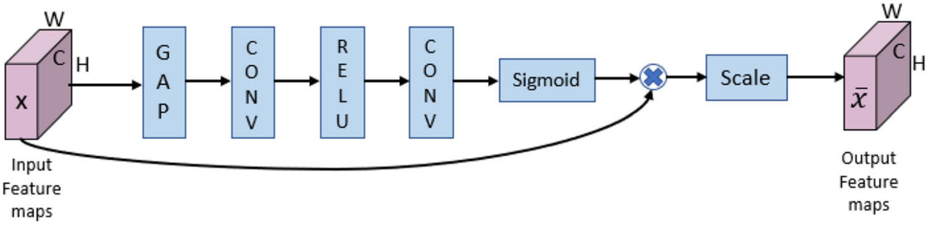

Fig. 3 Computation of channels correlation via channel attention module which consists of global average pooling, 1 × 1 convolution, ReLU and sigmoid layers

$W_f \in \mathbb{R}^{\bar{C}\times C}$ and $W_g \in \mathbb{R}^{\bar{C}\times C}$ ($\bar{C} = \frac{C}{r}$ and r is the reduction ratio). The attention map $\alpha_{j,i}$ for $s_{ij} = f(x_i)^T g(x_j)$ is computed as:

$$\alpha_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N} exp(s_{ij})} \tag{5}$$

where $\alpha_{j,i}$ denotes the extent to which the model contributes in the i^{th} location when synthesizing the j^{th} region. The attention layer's output is $o = (o_1, o_2, ..., o_j, ..., o_N) \in \mathbb{R}^{C\times N}$ and

$$o_j = v\left(\sum_{i=1}^{N} \alpha_{j,i} h(x_i)\right) \tag{6}$$

where $h(x_i) = W_h x_i$ and $v(x_i) = W_v x_i$.

$W_h \in \mathbb{R}^{\bar{C}\times C}$, and $W_v \in \mathbb{R}^{\bar{C}\times C}$ in (6) are implemented using 1 × 1 convolution layers and are the learned weight matrices. $\bar{C}$ is the number of channels after reduction C/r. To avoid memory usage, we set $r = 16 (i.e., \bar{C} = C/16)$ in our experiments. The output of the attention layer is added back to the input feature map. The final output is written as:

$$y_i = \gamma o_i + x_i \tag{7}$$

where γ is a learnable scalar parameter and is initialized as 0. In the start, γ encourages the network to rely on the cues in the local neighbourhood and then gradually learns to assign more weight to non-local evidence. We append the self attention module at the start of the attentive branch which is shown in Fig. 2. The operation of computing the features correlation through self attention is shown in Fig. 4.

4 Experiments

4.1 Datasets

We performed our experiments and evaluated the proposed network on two person re-id benchmark datasets, market-1501 [49] and DukeMTMC-reID [26]. We adopted rank-1 accuracy, rank-5 accuracy and mean average precision (mAP) as our evaluation metrics. We used the standards splits for training and testing identities. The details about the two datsets are:

Market-1501 dataset has 32668 images of 1501 person identities automatically detected from six disjoint cameras. The training set consists of 12936 images of 751 identities. The

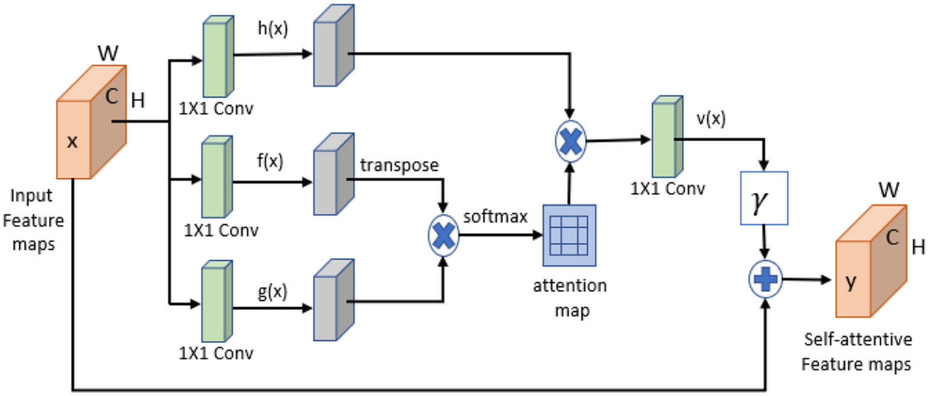

Fig. 4 Computation of feature correlations via self attention module. The dimension of the output (self-attention) features is the same of the input because they are the input to the next residual block of the ResNet-50 Network

query set has 3368 probe images of 750 identities and the gallery set has 19732 images with 750 identities.

DukeMTMC-reID dataset contains manually annotated boxes generated by eight cameras. It is composed by 36411 images of 1404 identities. There are 16522 images of 702 identities in the training set. The query and testing sets have 2228 and 17661 images of 702 identities, respectively.

4.2 Implementation details

The backbone of the proposed CadNet network consists of a ResNet-50 network and is implemented using Pytorch. We trained CadNet on a nvidia RTX2080Ti gpu. Following the work of R-FCN [8], we modified the stride (stride=1) of the last downsampling block before the global average pooling to make the spatial size of convolution features larger. We used global max pooling on these features instead of global average pooling. A 1×1 convolution layer followed by Batch normalization, Rectified Linear Unit (ReLU) and dropout layers are appended after the max pooling to reduce the size of features from 2048 to 1024. We add several modifications in the network as specified in Section 3.2. The channels correlation of features at each stage are computed by the channel attention modules embedded throughout the network. The two branches of the proposed CadNet provide two feature vectors of length 1024 and are trained separately (without concatenation) by using shared multiple classifiers. The concatenation of the features from two branches defines the final representation vector of length 2048 which is used for feature matching. We optimized the network by using Adam optimizer with momentum 0.9. The initial learning rate is set to $3e-4$ and is divided by 10 after 80 epochs. We trained our model for 140 epochs with a batch size of 64. Photometric distortions [14] and the AlexNet-style color augmentation [12] are applied to 256×128 sized images followed by random horizontal flipping and random erasing data augmentations. The dropout probability is set to 0.5 and the weight decay is $5e-4$. The ResNet-50 baseline training time, on the exploited testing configuration, is 2.5 hours for Market-1501 and 3 hours for DukeMTMC-reID. The proposed CadNet converges in 3.5

hours for DukeMTMC-reID and takes 3 hours for Market-1501 dataset to train. The training time for the CadNet is comparable with respect to the baseline while the performance is significantly higher than the baseline.The inference time is identical for both baseline and CadNet (0.175 sec per batch).

4.3 Comparison with state-of-art methods

The results of the proposed CadNet along with the comparison with other state of the art methods on market-1501 and DukeMTMC-reID datasets are presented in Table 1 and Table 2. Unlike the other state of the art methods, the proposed CadNet introduce multi classifiers training mechanism which enhance the performance. The gradients from the each added classifiers are gathered at previous convolution layer and make that layer to learn more and more refined and discriminative features with each addition. With 4 classifiers

Table 1 Comparisons of the proposed CadNet to the state-of-the-art re-id methods on Market-1501

Methods	Reference	Rank-1(%)	Rank-5(%)	mAP
SpindleNet [47]	CVPR17	76.9	91.5	-
Part-Aligned [48]	ICCV17	81.0	92.0	63.4
HydraPlus-Net [19]	ICCV17	76.9	91.3	-
LSRO [52]	ICCV17	84.0	-	66.1
SVDNet [33]	ICCV17	82.3	92.3	62.1
DPFL [7]	ICCV17	88.9	92.3	73.1
PSE [27]	CVPR18	87.7	94.5	69.0
HA-CNN [18]	CVPR18	91.2	-	75.5
AACN [40]	CVPR18	85.9	-	66.9
MLFN [2]	CVPR18	90.0	-	74.3
DuATM [31]	CVPR18	91.4	97.1	76.6
DKP [29]	CVPR18	90.1	96.7	75.3
GCSL [3]	CVPR18	93.5	-	81.6
PCB [34]	ECCV18	92.3	97.2	77.4
Part-aligned [48]	ECCV18	91.7	96.9	79.6
SGGNN [28]	ECCV18	92.3	96.1	82.8
Mancs [37]	ECCV18	93.1	-	82.3
IDCL [42]	CVPRW19	93.9	97.8	80.5
PyrNet [21]	CVPRW19	93.6	98.2	81.7
AWPCN [30]	MMTA20	94.0	-	82.1
MMHPN [46]	MMTA20	94.6	-	83.4
APA [53]	MMTA20	93.6	-	81.7
MSMP [36]	NC20	93.7	-	81.2
CASN(IDE) [50]	CVPR19	92.0	-	78.0
SFT [20]	ICCV19	93.4	97.4	82.7
Baseline(ResNet-50)	-	92.6	-	78.6
SCAN [24]	ICPR20	94.2	97.8	83.6
CadNet(Proposed)	-	94.6	98.0	85.2

The top 1 and 2 results are mentioned in red and blue

Table 2 Comparisons to the state-of-the-art re-id methods on DukeMTMC-reID dataset

Methods	Reference	Rank-1(%)	Rank-5(%)	mAP
Verif-Identif [51]	TOMM18	68.9	-	49.3
LSRO [52]	ICCV17	67.7	-	47.1
SVDNet [33]	ICCV17	76.7	86.4	56.8
DPFL [7]	ICCV17	73.2	-	60.6
PSE [27]	CVPR18	79.8	89.7	62.0
HA-CNN [18]	CVPR18	80.5	-	63.8
AACN [40]	CVPR18	76.8	-	59.2
MLFN [2]	CVPR18	81.0	-	62.8
DuATM [31]	CVPR18	81.8	90.2	68.6
DKP [29]	CVPR18	80.3	89.5	63.2
GCSL [3]	CVPR18	84.9	-	69.5
PCB+RPP [34]	ECCV18	83.3	-	69.2
Part-aligned [48]	ECCV18	84.4	92.2	69.3
SGGNN [28]	ECCV18	81.1	88.4	68.2
Mancs [37]	ECCV18	84.9	-	71.8
IDCL [42]	CVPRW19	84.7	-	69.4
CASN(IDE) [50]	CVPR19	84.5	-	67.0
AWPCN [30]	MMTA20	85.7	-	74.1
APA [53]	MMTA20	84.7	-	69.4
MSMP [36]	NC20	84.4	-	70.4
Baseline(ResNet-50)	-	82.8	-	65.2
SCAN [24]	ICPR20	85.3	92.7	71.0
CadNet(Proposed)	-	86.3	92.8	72.7

The highest and second highest results are shown in red and blue

we got the highest performance and further addition of classifiers starts reducing the scores because of the classifiers errors which are also getting added for each classifier. To handle the blurriness and noise in the data, the proposed network computes channels correlation continuously at various spatial sizes. These correlations produce the noise free feature maps from degraded data [9, 45] and proceed them towards the classifiers. The sharpness in features make them easy to distinguish from each other and hence improve the matching scores. For further refinement of features, the proposed network adopts a dual branch mechanism. The contribution of the residual branch in the performance of the CadNet is to provide the noise free and discrimative features of the image which enlarge the difference between two different identities. The attentive branch merge the information from the distant image location which has higher contributions in the prediction of person's identity. The concatenation of the attentive features with residual features amplify the information in the learned feature vectors. The final representation from the proposed method produce better matching between to person and accelerate the performance. The effects of each of the components on the results of the proposed method are explained in Section 4.5. In all experiments, we reported the results of the other methods directly from their papers. The results in Table 1 and Table 2 show the superior performance of the proposed method as compared to other state of the art methods. Unlike other methods, the learned features with the proposed CadNet consist of distant similarities and hence provide better and unique representation of the person. Therefore, the performance of the proposed method is higher than the others.

4.4 Visual results

The visual/qualitative results from the proposed network are illustrated in Fig. 5. We used the trained CadNet model to obtain the feature representations of all the images and then followed [54] to compute the visual results. We computed the class activation maps [54] for both the datasets and present them visually. The proposed network is unable to compute the class activation maps at the last convolution block because we reduce the size of the features to 1024. The last block returns 2048 feature maps and the input to the classifiers is 1024. Since they have different sizes thus we calculated class activation maps at the third convolution block where the network is split into two branches. In Fig. 6 first row shows the original images from Market1501 dataset and second row represents the corresponding class activation maps. Similarly, third and fourth rows demonstrate the class activation maps for DukeMTMC-reID dataset. Class activation maps represents how much each image region contributes in the prediction of classes (person identities) probabilities. The highest contribution in the predictions is carried out by red regions while the blue regions represents the lowest contribution (or no contribution). The images clearly show that the proposed solution takes into higher consideration regions belonging to persons while is discarding the background. This behaviour contributes to improve the quantitative performance.

4.5 Ablation study

4.5.1 Effect of classifiers

To evaluate the contribution of single classifier versus multiple classifiers, we modified the ResNet-50 baseline to get unbiased results. In this view, we trained the ResNet-50 [11] baseline network with different number of classifiers and reported the rank1 accuracy and mAP for both the datasets. In particular, we used a ResNet-50 pretrained on ImageNet dataset and removed its last linear layers. New linear layers according to the number of

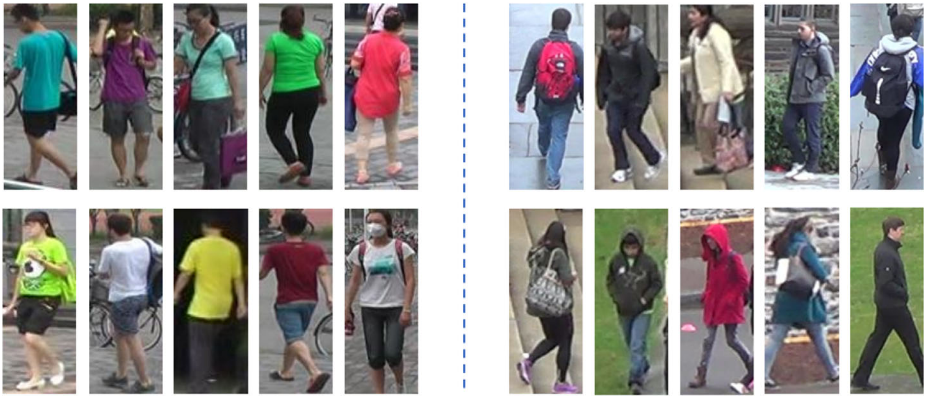

Fig. 5 Images from the two datasets are shown in this figure. The images on the right and left side of the dashed line are taken from Market1501 and DukeMTMC-reID datsets respectively

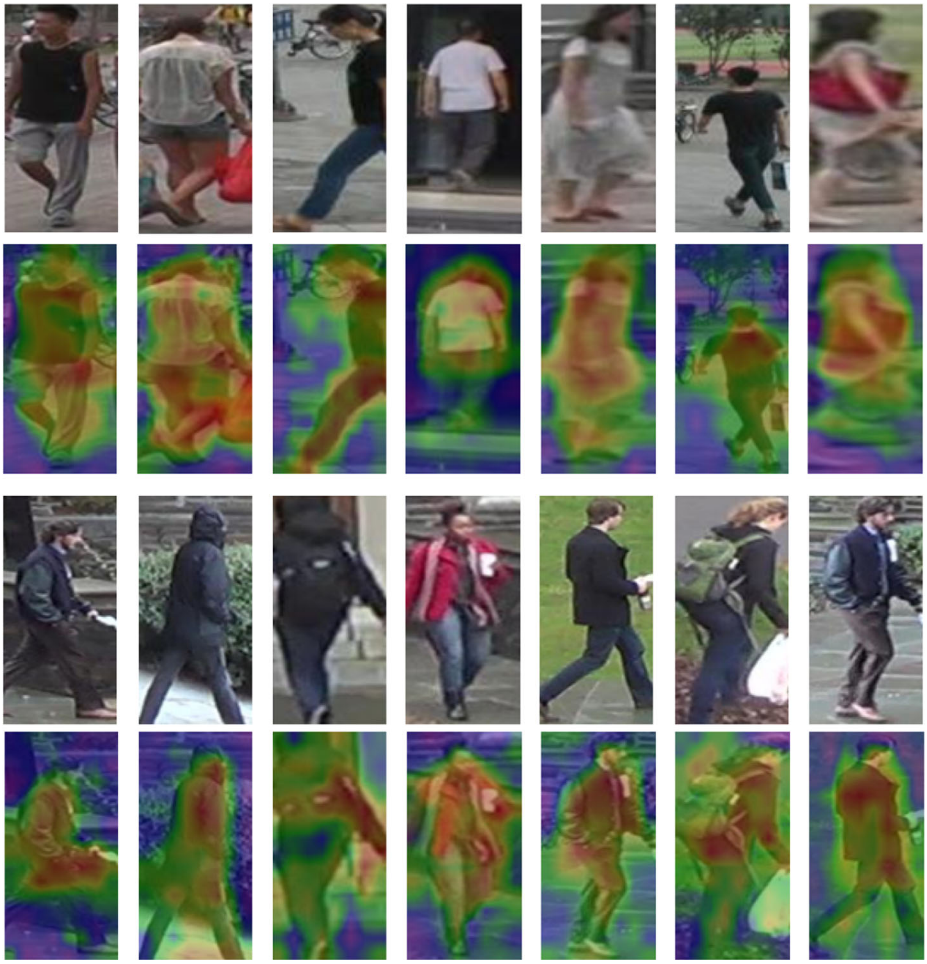

Fig. 6 Class activation maps obtained with the proposed method. First and third rows are the original images and second and fourth rows consists of corresponding class activation maps for Market1501 and DukeMTMC-reID datasets respectively

classes present in the datasets have been appended before training on the re-id datasets. Figure 7 shows the contribution of different classifier layer on the ResNet-50 baseline. Both the measurements rank1 accuracy and mAP linearly increase until 4 classifiers. Then the slopes reaches a plateau or decrease gently. Since the highest performance has been reached with 4 classifiers we exploited such a number of layers in the CadNet solution and in the aforementioned/comparison results.

4.5.2 Parameters selection

Most of the parameters and specifications expressed in Section 3.2 are used throughout our experiments and are gathered from the previous standard person re-identification techniques. Instead, for the parameter *r* used in the self attention module can be set to 2,4,8,16.

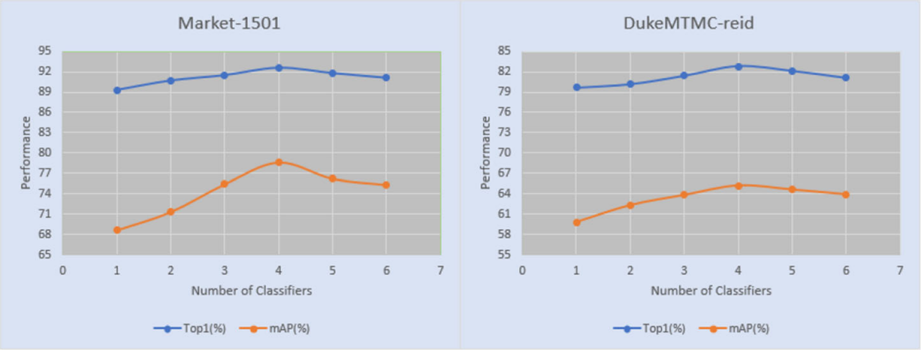

Fig. 7 Effect of number of classifiers on the performance of the network on two benchmarks. Both the measurements represent peak values when 4 classifiers are selected for each datasets

The parameter r is the division factor to generate patches from the input features. We reported the results by selecting multiple sizes of the generated patches in self attention module. The performance is slightly effected with different values of r because the number of channels $\bar{C}$ are reduced to C/r. The impact of different values for r is shown in Fig. 8. Evaluating such information, we chose $r = 16$ (i.e., $\bar{C}$ = C/16) in all our experiments. Such a selection not only improves qualitative performance but also reduces the computational costs and improves memory efficiency.

4.5.3 Component analysis

To illustrate the effectiveness of the proposed contributions, we provide a component analysis for the proposed network. First, we performed the separation of the three components (Channels Correlation CC, Feature Correlation FC and multi classifiers Multi-C) according to SCAN [24] and reported the results in Table 3. The ResNet-50 baseline proposed in Section 4.5.1 (one classifier) has been exploited as performance reference (the first row). Second row shows the impact of the exploitation of four classification layers. Such first two rows show the numerical values of points 1 and 4 of Fig.7.

Third and forth rows in Table 3 demonstrate the contributions of the channel attention (CA) and self attention (SA) modules. SCAN [24] is the ResNet-50 4-C with both modules. Third and forth row of Table 3 show performance of SCAN without SA and CA respectively. The performance improvement of the new exploitation of the channel attention with respect

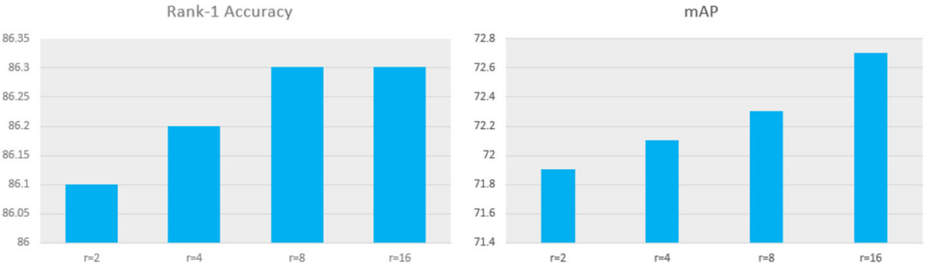

Fig. 8 The impact of choosing different values for the reduction ratio r on DukeMTMC-reID dataset. Left side represents the rank-1 accuracy scores and right side shows the mean average precision (mAP)

Table 3 Component Analysis of the proposed CadNet on Market-1501 and DukeMTMC-reID datasets in terms of mAP(%) and top-1 accuracy(%)

Networks	Components			Market1501		DukeMTMC-reID	
	CC	FC	Multi-C	mAP	R1	mAP	R1
ResNet-50 1-C baseline	×	×	×	68.6	89.3	59.8	79.7
ResNet-50 4-C	×	×	✓	78.6	92.6	65.2	82.8
SCAN [24] - SA	✓	×	✓	81.6	93.5	68.9	83.9
SCAN [24] - CA	×	✓	✓	80.8	93.8	68.2	84.5
SCAN [24]	✓	✓	✓	83.6	94.2	71.0	85.3
CadNet(residual branch)	✓	×	✓	84.1	94.4	71.7	85.5
CadNet (attentive branch)	✓	✓	✓	84.4	94.4	71.9	85.7
CadNet (proposed)	✓	✓	✓	**85.2**	**94.6**	**72.7**	**86.3**

CC and FC represents the channel and feature correlations respectively

The bold entries show the final results with all the proposed components insertion to form CadNet

to SCAN [24] show the superiority of CadNet(residual branch) in row 6 with respect to SCAN row 5. To evaluate the CadNet(attentive branch), we trained the proposed network shown in Fig. 2 with only such a branch (e.g. residual branch has been removed). The results are in row 7 of Table 3. Finally, both branches have been used (row 8) to show the results of the complete proposed solution. Each of the residual and attentive branches improve the performance over SCAN [24] separately but the combination of both branches have a greater effect on the performance. This implies that the mixture of all the proposed contributions together in the form of CadNet provides a stronger descriptor.

With the proposed placement of channels correlation (CC), $R1$ and mAP are increased by 0.9% and 2.5% for Market1501 and 1.6% and 2.8% for DukeMTMC-reID as compared to SCAN. Similarly, the dual branch design with the proposed CC and feature correlation (FC) improves $R1$ and mAP by 0.4% and 1.6% for Market1501 and 1.0% and 1.3% for DukeMTMC-reID.

5 Conclusions

We proposed a novel Consistent Attentive Dual Branch Network with multiple classifiers for Person Re-Identification (CadNet). We exploited a multi classifiers training strategy in which each classifier contributes in distinguishing between identities and helps the model to learn the most discriminative and unique features for each person. Due to blurry and noisy person re-id data, general re-id models misses small and tiny details. The introduction of channels correlation makes the learned features noise free and highlights these small details to build a stronger descriptor. Channels correlation are computed through channel attention module consistently at multiple positions in the network to flow the tiny information towards the final representations. Local and non-local similarities in the person images are computed by the two branches, respectively residual and attentive ones, and merged to create a strong, unique and discriminative feature representation of each person. This has been shown to improve the matching score between two persons. Spectral normalization is applied while computing channel and self correlations to stabilize the training dynamics for the convergence of the model. The visual results show the participation of each tiny component of person in predicting the identity. The proposed CadNet learns small details that help to significantly enhance the person re-id performance with respect to other state of the art methods as shown on two widely adopted benchmarks datasets. The proposed network only focused on learning similarities/dissimilarities present for a single person. Cross correlations can be introduced in the future work to learn distinguishing features between two persons.

Acknowledgements This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866).

Funding Open access funding provided by Università degli Studi di Udine within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: CVPR
2. Chang X, Hospedales TM, Xiang T (2018) Multi-level factorisation net for person re-identification. In: CVPR
3. Chen D, Xu D, Li H, Sebe N, Wang X (2018) Group consistent similarity learning via deep crf for person re-identification. In: CVPR
4. Chen J, Li K, Deng Q, Li K, Philip SY (2019) Distributed deep learning model for intelligent video surveillance systems with edge computing. IEEE Transactions on Industrial Informatics
5. Chen L, Yang H, Xu Q, Gao Z (2020) Harmonious attention network for person re-identification via complementarity between groups and individuals. Neurocomputing
6. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR
7. Chen Y, Zhu X, Gong S (2017) Person re-identification by deep learning multi-scale representations. In: ICCVW
8. Dai J, Li Y, He K, Sun J (2016) R-fcn: Object detection via region-based fully convolutional networks. In: NIPS
9. Dai T, Cai J, Zhang Y, Xia ST, Zhang L (2019) Second-order attention network for single image super-resolution. In: CVPR
10. Ding S, Lin L, Wang G, Chao H (2015) Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn 48(10):2993–3003
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
13. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:1703.07737
14. Howard AG (2013) Some improvements on deep convolutional neural network based image classification. arXiv:1312.5402
15. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: CVPR
16. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS
17. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR

18. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: CVPR
19. Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017) Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV
20. Luo C, Chen Y, Wang N, Zhang Z (2019) Spectral feature transformation for person re-identification. In: ICCV
21. Martinel N, Luca Foresti G, Micheloni C (2019) Aggregating deep pyramidal representations for person re-identification. In: CVPRW
22. Micheloni C, Remagnino P, Eng HL, Geng J (2010) Intelligent monitoring of complex environments. IEEE Intelligent Systems
23. Munir A, Martinel N, Micheloni C (2020) Multi branch siamese network for person re-identification. In: ICIP
24. Munir A, Martinel N, Micheloni C (2020) Self and channel attention network for person re-identification. In: ICPR
25. Munir A, Micheloni C (2020) Self attention based multi branch network for person re-identification. In: Splitech
26. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV
27. Saquib Sarfraz M, Schumann A, Eberle A, Stiefelhagen R (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: CVPR
28. Shen Y, Li H, Yi S, Chen D, Wang X (2018) Person re-identification with deep similarity-guided graph neural network. In: ECCV
29. Shen Y, Xiao T, Li H, Yi S, Wang X (2018) End-to-end deep kronecker-product matching for person re-identification. In: CVPR
30. Shu X, Yuan D, Liu Q, Liu J (2020) Adaptive weight part-based convolutional network for person re-identification. Multimed Tools Appl 79(31):23617–23632
31. Si J, Zhang H, Li CG, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR
32. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
33. Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: ICCV
34. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV
35. Varior RR, Shuai B, Lu J, Xu D, Wang G (2016) A siamese long short-term memory architecture for human re-identification. In: ECCV
36. Wang C, Song L, Wang G, Zhang Q, Wang X (2020) Multi-scale multi-patch person re-identification with exclusivity regularized softmax. Neurocomputing 382:64–70
37. Wang C, Zhang Q, Huang C, Liu W, Wang X (2018) Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In: ECCV
38. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: CVPR
39. Xiang S, Fu Y, Chen H, Ran W, Liu T (2020) Multi-level feature learning with attention for person re-identification. Multimed Tools Appl 79(43):32079–32093
40. Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. In: CVPR
41. Yu R, Dou Z, Bai S, Zhang Z, Xu Y, Bai X (2018) Hard-aware point-to-set deep metric for person re-identification. In: ECCV
42. Zhai Y, Guo X, Lu Y, Li H (2019) In defense of the classification loss for person re-identification. In: CVPRW
43. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: ICML
44. Zhang X, Luo H, Fan X, Xiang W, Sun Y, Xiao Q, Jiang W, Zhang C, Sun J (2017) Alignedreid: Surpassing human-level performance in person re-identification. arXiv:1711.08184
45. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks
46. Zhang Y, Liu S, Qi L, Coleman S, Kerr D, Shi W (2020) Multi-level and multi-scale horizontal pooling network for person re-identification. Multimed Tools Appl 79(39):28603–28619
47. Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X (2017) Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR
48. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: ICCV
49. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: ICCV

50. Zheng M, Karanam S, Wu Z, Radke RJ (2019) Re-identification with consistent attentive siamese networks. In: CVPR
51. Zheng Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person reidentification. TOMM 14(1):1–20
52. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV
53. Zhong W, Jiang L, Zhang T, Ji J, Xiong H (2020) A part-based attention network for person re-identification. Multimed Tools Appl 79:22525–22549
54. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: CVPR

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.