Relation-Aware Global Attention for Person Re-identification

Zhizheng Zhang[1*] Cuiling Lan[2†] Wenjun Zeng[2] Xin Jin[1] Zhibo Chen[1†]
[1]University of Science and Technology of China [2]Microsoft Research Asia
{zhizheng,jinxustc}@mail.ustc.edu.cn {culan,wezeng}@microsoft.com chenzhibo@ustc.edu.cn

Abstract

For person re-identification (re-id), attention mechanisms have become attractive as they aim at strengthening discriminative features and suppressing irrelevant ones, which matches well the key of re-id, i.e., discriminative feature learning. Previous approaches typically learn attention using local convolutions, ignoring the mining of knowledge from global structure patterns. Intuitively, the affinities among spatial positions/nodes in the feature map provide clustering-like information and are helpful for inferring semantics and thus attention, especially for person images where the feasible human poses are constrained. In this work, we propose an effective Relation-Aware Global Attention (RGA) module which captures the global structural information for better attention learning. Specifically, for each feature position, in order to compactly grasp the structural information of global scope and local appearance information, we propose to stack the relations, i.e., its pairwise correlations/affinities with all the feature positions (e.g., in raster scan order), and the feature itself together to learn the attention with a shallow convolutional model. Extensive ablation studies demonstrate that our RGA can significantly enhance the feature representation power and help achieve the state-of-the-art performance on several popular benchmarks. The source code is available at https://github.com/microsoft/Relation-Aware-Global-Attention-Networks.

1. Introduction

Person re-identification (re-id) aims to match a specific person across different times, places, or cameras, which has drawn a surge of interests from both industry and academia. The challenge lies in how to extract discriminative features (for identifying the same person and distinguishing different persons) from person images where there are background clutter, diversity of poses, occlusion, *etc*.

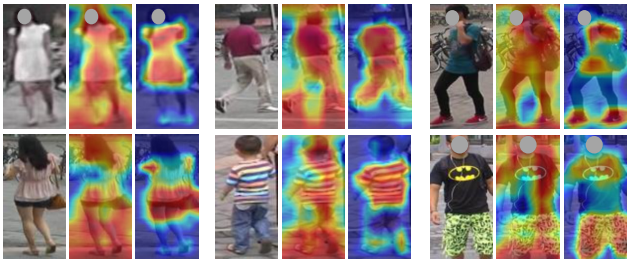


Figure 1. Comparison of the learned attention between (b) spatial attention of CBAM [38] without exploring relations, and (c) our proposed attention which captures global scope relations and mines from such structural information. (a) The original image [1].

Recently, many studies resort to attention design to address the aforementioned challenges in person re-id by strengthening the discriminative features and suppressing interference [22, 39, 20, 12, 11, 6]. Most of the attentions are learned by convolutions with limited receptive fields, which makes it hard to exploit the rich structural patterns in a global scope. One solution is to use large size filters in the convolution layer [38]. The other solution is to stack deep layers [33] which increases the network size greatly. Besides, the studies in [24] show that the effective receptive field of CNN only takes up a fraction of the full theoretical receptive field. These solutions cannot ensure the effective exploration of global scope information (*e.g.*, global scope contents and corresponding positional geometry) for effective person re-id.

Moreover, the non-local neural network is proposed in [35] to allow the collection of global information by weighted summation of the features from all positions to the target position, where the connecting weight is calculated by the pairwise relation/affinity. Actually, for a target feature position, its pairwise relations with all the feature nodes/positions could contain valuable structural information of a global scope, *e.g.*, clustering-like pattern (through *pairwise affinities* and *position information*). However, the non-local network overlooks the exploration

*This work was done when Zhizheng Zhang was an intern at MSRA.
†Corresponding author.

[1]All faces in the images are masked for anonymization.

of such rich global information. It only simply uses the learned relations/affinities as the weights to aggregate the features. Such a deterministic manner of using relations (*i.e.*, weighted sum) has weak mining capability and lacks sufficient adaptability. Cao *et al*. observe that the learned connecting weights of non-local block are target position invariant [5], which is not as adaptive as expected. We believe it is important to mine knowledge from the relations through a modeling function and leverage such valuable global scope structural information to infer attention.

In this paper, we propose an effective Relation-Aware Global Attention (RGA) module to efficiently learn discriminative features for person re-id. RGA explicitly explores the global scope relations for mining the structural information (clustering-like information). This is helpful for implicitly inferring semantics and thus attention. Fig. 1 shows our learned attention on the person re-id images. Thanks to the introduction and mining of global scope relations, our attention can focus on the discriminative human body regions. As illustrated in Fig. 2 (c), for each feature node, *e.g.*, a feature vector of a spatial position on a feature map, we model the pairwise relations of this node with respect to all the nodes and compactly stack the relations as a vector (which represents the global structural information) together with the feature of the node itself to infer the attention intensity via a small model. In this way, we take into account both the appearance feature and its global scope relations, to determine the feature importance from a global view. This mechanism is also consistent with the perception of human in finding discriminative features: making a global scope comparison to determine the importance.

In summary, we have made two major contributions:

- We propose to globally learn the attention for each feature node by taking a global view of the relations among the features. With the global scope relations having valuable structural (clustering-like) information, we propose to mine semantics from relations for deriving attention through a learned function. Specifically, for a feature node, we build a compact representation by stacking its pairwise relations with respect to all feature nodes as a vector and mine patterns from it for attention learning.
- We design a relation-aware global attention (RGA) module which compactly represents the global scope relations and derives the attention based on them via two convolutional layers. We apply such design to spatial (RGA-S) and channel dimensions (RGA-C) and demonstrate its effectiveness for person re-id.

We conduct extensive ablation studies to demonstrate the effectiveness of the proposed RGA in finding discriminative features and suppressing irrelevant ones for person re-id. Our scheme empowered by RGA modules achieves the state-of-the art performance on the benchmark datasets CUHK03 [21], Market1501 [46], and MSMT17[37].

2. Related Work

2.1. Attention and Person Re-id

Attention aims to focus on important features and suppress irrelevant features. This well matches the goal of handling aforementioned challenges in person re-id and is thus attractive. Many works learn the attention using convolutional operations with small receptive fields on feature maps [32, 45, 22, 6]. However, intuitively, to have a good sense of whether a feature node is important or not, one should know the features of global scope which facilitates the comparisons needed for decision.

In order to introduce more contextual information, Wang *et al*. and Yang *et al*. stack many convolutional layers in their encoder-decoder style attention module to have larger receptive fields [33, 40]. Woo *et al*. use a large filter size of 7×7 over the spatial features in their Convolutional Block Attention Module (CBAM) to produce a spatial attention map [38]. In [42], a non-local block [35] is inserted before the encoder-decoder style attention module to enable attention learning based on globally refined features. Limited by the practical receptive fields, all these approaches are not efficient in capturing the large scope information to globally determine the spatial attention.

Some works explore the external clues of human semantics (pose or mask) as attention or to use them to guide the learning of attention [39, 28, 29, 44]. The explicit semantics which represent human structures is helpful for determining the attention. However, the external annotation or additional model for pose/mask estimation is usually required.

In this paper, we intend to explore the respective global scope relations for each feature node to learn attention. The structural information in the relation representation which includes both *affinity* and *location information* is helpful for learning semantics and infer attention.

2.2. Non-local/Global Information Exploration

Exploration of non-local/global information has been demonstrated to be very useful for image denoising [3, 8, 4], texture synthesis [10], super-resolution [14], inpainting [2], and even high level tasks such as image recognition, object segmentation [35] and action localization [7]. Non-local block in [35] aims at strengthening the features of the target position via aggregating information from all positions. For each target position/node, to obtain an aggregated feature, they compute a weighted summation of features of all positions (sources), with each weight obtained by computing the pairwise relation/affinity between the source feature node and target feature node. Then, the aggregated feature is added to the feature of the target position to form the output. Cao *et al*. visualized the target position specific connecting weights of source positions and surprisingly observed that the connecting weights are *not* specific to the

target positions [5], *i.e.*, the vector of connecting weights is invariant to the target positions where a connecting weight from a source position is actually only related to the feature of this source position. Simplified non-local block [5] exploits such target position invariant characteristic and determines each connecting weight by the source feature node only, which achieves very close performance as the original non-local. Note that the aggregated feature vector which is added to each target position is thus the same for different target positions and there is a lack of target position specific adaptation.

Even though non-local block also learns pairwise relations (connecting weights), the global scope structural information is not well exploited. They just use them as weights to aggregate the features in a deterministic manner and have not mined the valuable information from the relations. Different from non-local block, we aim to dig more useful information from a stacked relation representation and derive attention from it through a learned model. Our work is an exploration on how to make better use of relations and we hope it will inspire more works from the research community.

3. Relation-Aware Global Attention

For discriminative feature extraction in person re-id, we propose a Relation-aware Global Attention (RGA) module which makes use of the compact global scope structural relation information to infer the attention. In this section, we first give the problem formulation and introduce our main idea in Subsec. 3.1. For CNN, we elaborate on the designed spatial relation-aware global attention (RGA-S) in Subsec. 3.2 and channel relation-aware global attention (RGA-C) in Subsec. 3.3, respectively. We analyze and discuss the differences between our attention and some related approaches in Subsec. 3.4.

3.1. Formulation and Main Idea

Generally, for a feature set $\mathcal{V} = \{\mathbf{x}_i \in \mathbb{R}^d, i = 1, \cdots, N\}$ of N correlated features with each of d dimensions, the goal of attention is to learn a mask denoted by $\mathbf{a} = (a_1, \cdots, a_N) \in \mathbb{R}^N$ for the N features to weight/mask them according to their relative importance. Note that we also refer to a feature vector as feature node or feature.

Two common strategies are used to learn the attention value a_i of the i^{th} feature vector, as illustrated in Fig. 2 (a) and (b). **(a) Local attention**: the attention for a feature node is determined locally, *e.g.*, applying a shared transformation function $\mathcal{F}$ on itself, *i.e.*, $a_i = \mathcal{F}(\mathbf{x}_i)$ [32]. However, such local strategies do not fully exploit the correlations from a global view and ignore the global scope structural information. For vision tasks, deep layers [33] or large-size kernels [38] are used to remedy this problem. **(b) Global attention**: one solution is to use all the feature nodes (*e.g.* by concatenation) together to jointly learn attention, *e.g.*, using fully connected operations. However, this is usually computationally inefficient and difficult to optimize, as it requires a large number of parameters especially when the number of features N is large [23].

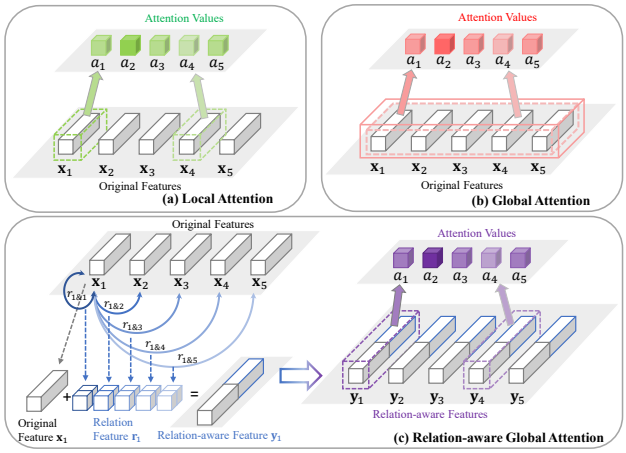
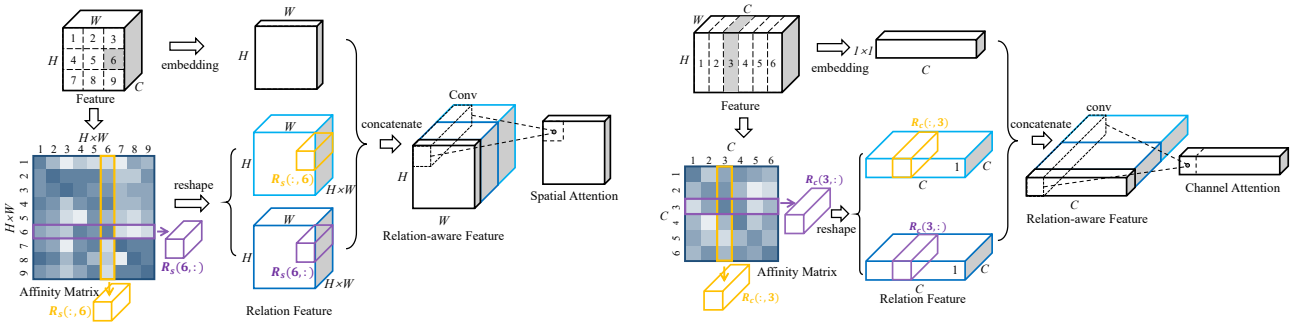


Figure 2. Illustration of learning attention values $a_1, \cdots, a_5$ for five feature vectors/nodes $\mathbf{x}_1, \cdots, \mathbf{x}_5$. (a) Local attention: learn attention locally (*e.g.*, based on individual feature as shown). (b) Global attention: learn attention jointly from all the 5 feature vectors (*e.g.*, by concatenating them together). (c) Proposed relation-aware global attention: learn attention by taking into account the global relation information. For the i^{th} (here $i = 1$) feature vector, the global scope relation information is represented by stacking the pairwise relations $\mathbf{r}_i = [r_{i,1}, \cdots, r_{i,5}, r_{1,i}, \cdots, r_{5,i}]$. Note that $r_{i\&j} = [r_{i,j}, r_{j,i}]$. Unlike (a) that lacks global awareness and (b) that lacks explicit relation exploration, our proposed attention is determined through a learned function with the global scope relations which contain structural information as input.

In contrast to these strategies, we propose a relation-aware global attention that enables i) the exploitation of global structural information and knowledge mining, and ii) the use of shared transformation function for different individual feature positions to derive the attention. For re-id, the latter makes it possible to globally compute the attention by using local convolutional operations. Fig. 2 (c) illustrates our basic idea for the **proposed relation-aware global attention**. The main idea is to exploit the pairwise relation (*e.g.* affinity/similarity) of the current (i^{th}) feature node with all the feature nodes, respectively, and stack them (with some fixed order) to compactly represent the global structural information for the current feature node. Specifically, we use $r_{i,j}$ to represent the affinity between the i^{th} feature and the j^{th} feature. For the feature node $\mathbf{x}_i$, its affinity vector is $\mathbf{r}_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,N}, r_{1,i}, r_{2,i}, \cdots, r_{N,i}]$. Then, we use the feature itself and the pairwise relations, *i.e.*, $\mathbf{y}_i = [\mathbf{x}_i, \mathbf{r}_i]$, as the feature used to infer its attention through a *learned* transformation function. Note that $\mathbf{y}_i$ contains global information.



(a) Spatial Relation-Aware Global Attention

(b) Channel Relation-Aware Global Attention

Figure 3. Diagram of our proposed Spatial Relation-aware Global Attention (RGA-S) and Channel Relation-aware Global Attention (RGA-C). When computing the attention at a feature position, in order to grasp information of global scope, we stack the pairwise relation items, *i.e.*, its correlations/affinities with all the feature positions, and the unary item, *i.e.*, the feature of this position, for learning the attention with convolutional operations.

Mathematically, we denote the set of features and their relations by a graph $G = (\mathcal{V}, \mathcal{E})$, which comprises the node set $\mathcal{V}$ of N features, together with an edge set $\mathcal{E} = \{r_{i,j} \in \mathbb{R}, i = 1, \cdots, N \text{ and } j = 1, \cdots, N\}$. The edge $r_{i,j}$ represents the relation between the i^{th} node and the j^{th} node. The pairwise relations for all the nodes can be represented by an affinity matrix $R \in \mathbb{R}^{N \times N}$, where the relation between node i and j is $r_{i,j} = R(i, j)$. $\mathbf{r}_i = [R(i, :), R(:, i)]$, where $R(i, :)$ denotes the i^{th} row of R and $R(:, i)$ denotes the i^{th} column of R.

Discussion: For the i^{th} feature node $\mathbf{x}_i$, its corresponding relation vector $\mathbf{r}_i$ provides a compact representation to capture the global structural information, *i.e.*, both the position information and pairwise affinities with respect to all feature nodes. With the pairwise relation values denoting the similarity/affinity between every feature node and the current feature node while their locations in the relation vector denoting the positions (indexes) of the feature nodes, the relation vector reflects *the clustering states and patterns* of all the nodes with respect to the current node, which benefits the global determination of the relative importance (attention) of $\mathbf{x}_i$. With such affluent structural information/patterns contained, we propose to mine from the relations for effectively learning attention through a modeling function. The structural patterns of person re-id images span in a learnable space considering the feasible poses are constrained by the human physical structure.

3.2. Spatial Relation-Aware Global Attention

Given an intermediate feature tensor $X \in \mathbb{R}^{C \times H \times W}$ of width W, height H, and C channels from a CNN layer, we design a spatial relation-aware attention block, namely RGA-S, for learning a spatial attention map of size $H \times W$. We take the C-dimensional feature vector at each spatial position as a feature node. All the spatial positions form a graph G_s of $N = W \times H$ nodes. As illustrated in Fig. 3 (a), we raster scan the spatial positions and assign their identification number as $1, \cdots, N$. We represent the N feature nodes as $\mathbf{x}_i \in \mathbb{R}^C$, where $i = 1, \cdots, N$.

The pairwise relation (*i.e.* affinity) $r_{i,j}$ from node i to node j can be defined as a dot-product affinity in the embedding spaces as:

$$r_{i,j} = f_s(\mathbf{x}_i, \mathbf{x}_j) = \theta_s(\mathbf{x}_i)^T \phi_s(\mathbf{x}_j), \tag{1}$$

where θ_s and ϕ_s are two embedding functions implemented by a 1×1 spatial convolutional layer followed by batch normalization (BN) and ReLU activation, *i.e.* $\theta_s(\mathbf{x}_i) = ReLU(W_\theta \mathbf{x}_i)$, $\phi_s(\mathbf{x}_i) = ReLU(W_\phi \mathbf{x}_i)$, where $W_\theta \in \mathbb{R}^{\frac{C}{s_1} \times C}$ and $W_\phi \in \mathbb{R}^{\frac{C}{s_1} \times C}$. s_1 is a pre-defined positive integer which controls the dimension reduction ratio. Note that BN operations are all omitted to simplify the notation. Similarly, we can get the affinity from node j to node i as $r_{j,i} = f_s(\mathbf{x}_j, \mathbf{x}_i)$. We use the pair $(r_{i,j}, r_{j,i})$ to describe the bi-directional relations between $\mathbf{x}_i$ and $\mathbf{x}_j$. Then, we represent the pairwise relations among all the nodes by an affinity matrix $R_s \in \mathbb{R}^{N \times N}$.

For the i^{th} feature node, we stack its pairwise relations with all the nodes in a certain fixed order (*e.g.*, raster scan order), *i.e.*, node identities as $j = 1, 2, \cdots, N$, to obtain a relation vector $\mathbf{r}_i = [R_s(i, :), R_s(:, i)] \in \mathbb{R}^{2N}$. For example, as in Fig. 3 (a), the sixth row and the sixth column of the affinity matrix R_s, *i.e.* $\mathbf{r}_6 = [R_s(6, :), R_s(:, 6)]$, is taken as the relation features for deriving the attention of the sixth spatial position.

To learn the attention of the i^{th} feature node, besides the pairwise relation items $\mathbf{r}_i$, we also include the feature itself $\mathbf{x}_i$ to exploit both the global scope structural information relative to this feature and the local original information. Considering these two kinds of information are not in the same feature domain, we embed them respectively and con-

catenate them to get the spatial relation-aware feature $\widetilde{\mathbf{y}}_i$:

$$\widetilde{\mathbf{y}}_i = [pool_c(\psi_s(\mathbf{x}_i)), \varphi_s(\mathbf{r}_i)], \tag{2}$$

where ψ_s and φ_s denote the embedding functions for the feature itself and the global relations, respectively. They are both implemented by a spatial 1×1 convolutional layer followed by BN and ReLU activation, *i.e.*, $\psi_s(\mathbf{x}_i) = ReLU(W_\psi \mathbf{x}_i)$, $\varphi_s(\mathbf{r}_i) = ReLU(W_\varphi \mathbf{r}_i)$, where $W_\psi \in \mathbb{R}^{\frac{C}{s_1} \times C}$, $W_\varphi \in \mathbb{R}^{\frac{2N}{2s_1} \times 2N}$. $pool_c(\cdot)$ denotes global average pooling operation along the channel dimension to further reduce the dimension to 1. Then $\widetilde{\mathbf{y}}_i \in \mathbb{R}^{1+N/s_1}$. Note that other convolution kernel size (*e.g.* 3×3) can also be used. We found they achieve very similar performance so that we use 1×1 convolutional layer for lower complexity.

The global scope relations contain affluent structural information (*e.g.*, clustering-like state in feature space with semantics), we propose to mine valueable knowledge from them for inferring attention through a learnable model. We obtain the spatial attention value a_i for the i^{th} feature/node through a modeling function as:

$$a_i = Sigmoid(W_2 ReLU(W_1 \widetilde{\mathbf{y}}_i)), \tag{3}$$

where W_1 and W_2 are implemented by 1×1 convolution followed by BN. W_1 shrinks the channel dimension with a ratio of s_2 and W_2 transforms the channel dimension to 1.

3.3. Channel Relation-Aware Global Attention

Given an intermediate feature tensor $X \in \mathbb{R}^{C \times H \times W}$, we design a relation-aware channel attention block, namely RGA-C, for learning a channel attention vector of C dimensions. We take the $d = H \times W$-dimensional feature map at each channel as a feature node. All the channels form a graph G_c of C nodes. We represent the C feature node as $\mathbf{x}_i \in \mathbb{R}^d$, where $i = 1, \cdots, C$.

Similar to spatial relation, the pairwise relation $r_{i,j}$ from node i to node j can be defined as a dot-product affinity in the embedding spaces as:

$$r_{i,j} = f_c(\mathbf{x}_i, \mathbf{x}_j) = \theta_c(\mathbf{x}_i)^T \phi_c(\mathbf{x}_j), \tag{4}$$

where θ_c and ϕ_c are two embedding functions that are shared among feature nodes. We achieve the embedding by first spatially flattening the input tensor X into $X' \in \mathbb{R}^{(HW) \times C \times 1}$ and then using a 1×1 convolution layer with BN followed by ReLU activation to perform a transformation on X'. As illustrated in Fig. 3 (b), we obtain and then represent the pairwise relations for all the nodes by an affinity matrix $R_c \in \mathbb{R}^{C \times C}$.

For the i^{th} feature node, we stack its corresponding pairwise relations with all the nodes to have a relation vector $\mathbf{r}_i = [R_c(i,:), R_c(:,i)] \in \mathbb{R}^{2C}$, to represent the global structural information.

To infer the attention of the i^{th} feature node, similar to the derivation of the spatial attention, besides the pairwise relation items $\mathbf{r}_i$, we also include the feature itself $\mathbf{x}_i$. Similar to Eq. (2) and (3), we obtain the channel relation-aware feature $\mathbf{y}_i$ and then the channel attention value a_i for the i^{th} channel. Note that all the transformation functions are shared by nodes/channels. There is no fully connected operation across channels.

3.4. Analysis and Discussion

We analyze and discuss the differences from other related approaches. Moreover, we discuss the joint use of the spatial and channel RGA and their integration strategies.

RGA vs. CBAM[38]. Most of the attention mechanisms in CNN are actually local attention, which determines the attention of a feature position using local context [38, 33, 32, 22]. Taking the representative attention module CBAM [38] as an example, it uses a convolution operation of filter size 7×7 followed by sigmoid activation function to determine the attention of a spatial feature position. Therefore, only $7 \times 7 = 49$ neighboring feature nodes are exploited to determine the attention of the center position. In contrast, for our spatial RGA (RGA-S), for a spatial feature position, we jointly exploit the feature nodes at all spatial positions to globally determine the attention. We achieve this through simple 1×1 convolutional operations on the vector of stacked relations.

RGA vs. Non-local (NL) [35] and Simplified NL [5]. Non-local block [35] exploits the global context to refine the feature at each spatial position. For a target feature position, to obtain an aggregated feature which is then added to the original feature for refinement, they compute a weighted summation of features of source positions. Even though there is structural information from the pairwise relations, non-local overlooks the exploration of such valuable information and only uses the relations as weights for feature aggregation through such a deterministic manner. As observed and analyzed by Cao *et al*. [5], the connecting weights in non-local block are *invariant to the target positions*, with each connecting weight *locally* determined by the source feature node itself. Therefore, the vector of the connecting weights is the same for different target positions, so is the corresponding aggregated feature vector. This results in a lack of target position specific adaptation. In contrast, in our RGA, even though we similarly make use of the pairwise relations, our intention is rather different which is to *mine* knowledge from the global scope structural information of the relations through a learned modeling function.

Usage of RGA-S and RGA-C. RGA-S and RGA-C can be plugged into any CNN network in a plug-and-play fashion. We can use RGA-S or RGA-C alone, or jointly use them in sequence (*e.g.*, apply RGA-C following RGA-S which is denoted as RGA-SC) or in parallel (RGA-S//C).

4. Experiments

4.1. Implementation Details and Datasets

Network Settings. Following the common practices in re-id [41, 1, 43], we take ResNet-50 [15] to build our baseline network and integrate our RGA modules into the RetNet-50 backbone for effectiveness validation. Similar to [30, 43], the last spatial down-sampling operation in the conv5_x block is removed. In our experiments, we add the proposed RGA modules after all of the four residual blocks (including conv2_x, conv3_x, conv4_x and conv5_x). For brevity, we also refer to the scheme as *RGA*. Within RGA modules, we set the ratio parameters s_1 and s_2 to be 8. We use both identification (classification) loss with label smoothing [31] and triplet loss with hard mining [16] as supervision. Note that we do not implement re-ranking [49].

Training. We use the commonly used data augmentation strategies of random cropping [36], horizontal flipping, and random erasing [50, 36, 32]. The input image size is 256×128 for all the datasets. The backbone network is pre-trained on ImageNet [9]. We adopt the Adam optimizer to train all models for 600 epochs with the learning rate of 8×10^{-4} and the weight decay of 5×10^{-4}.

Datasets and Evaluation Metrics. We conduct experiments on three public person re-id datasets, *i.e.*, CUHK03 [21], Market1501 [46], and the large-scale MSMT17 [37]. We follow the common practices and use the cumulative matching characteristics (CMC) at Rank-1 (R1) and mean average precision (mAP) to evaluate the performance.

4.2. Ablation Study

Following the common practice, we perform the ablation studies on two representative datasets CUHK03 (with the Labeled bounding box setting) and Market1501.

RGA related Models vs. Baseline. Table 1 shows the comparisons of our spatial RGA (*RGA-S*), channel RGA (*RGA-C*), their combinations, and the baseline. We observe that:

1) Either *RGA-S* or *RGA-C* significantly improves the performance over *Baseline*. On CUHK03, *RGA-S*, *RGA-C*, and the sequentially combined version *RGA-SC* significantly outperform *Baseline* by **5.7%**, **6.6%**, and **8.4%** respectively on mAP, and **5.5%**, **5.5%**, and **7.3%** respectively on Rank-1 accuracy. On Market1501, even though the performance of *Baseline* is already very high, *RGA-S* and *RGA-C* improve the mAP by 3.8% and 4.2%, respectively.

2) For learning attention, even without taking the visual features (Ori.), *i.e.*, feature itself, as part of the input, using the proposed global relation representation itself (*RGA-S w/o Ori.* or *RGA-C w/o Ori.*) significantly outperforms *Baseline*, by *e.g.* 5.0% or 5.9% in mAP accuracy on CUHK03.

3) For learning attention, without taking the proposed global relation (Rel.) as part of the input, the scheme *RGA-S w/o Rel.* or *RGA-C w/o Rel.* are inferior to our scheme *RGA-S* or *RGA-C* by 2.4% or 1.9% in mAP accuracy on CUKH03. Both 2) and 3) demonstrate that global scope relation representation is very powerful for learning attention.

Table 1. Performance (%) comparisons of our models with the baseline, and the effectiveness of the global relation representation (Rel.) and the feature itself (Ori.). w/o: without.

Model		CUHK03(L)		Market1501	
		R1	mAP	R1	mAP
Baseline	ResNet-50	73.8	69.0	94.2	83.7
Spatial	RGA-S w/o Rel.	76.8	72.3	94.3	83.8
	RGA-S w/o Ori.	78.2	74.0	95.4	86.7
	RGA-S	**79.3**	**74.7**	**96.0**	**87.5**
Channel	RGA-C w/o Rel.	77.8	73.7	94.7	84.8
	RGA-C w/o Ori.	78.1	74.9	95.4	87.1
	RGA-C	**79.3**	**75.6**	**95.9**	**87.9**
Both	RGA-S//C	77.3	73.4	95.3	86.6
	RGA-CS	78.6	75.5	95.3	87.8
	RGA-SC	**81.1**	**77.4**	**96.1**	**88.4**

4) The combination of the spatial RGA and channel RGA achieves the best performance. We study three ways of combination: parallel with a fusion (*RGA-S//C*), sequential spatial-channel (*RGA-SC*), sequential channel-spatial (*RGA-CS*). *RGA-SC* achieves the best performance, 2.7% and 1.8% higher than *RGA-S* and *RGA-C*, respectively, in mAP accuracy on CUHK03. Sequential architecture allows the later module to learn attention based on modulated features resulting from its preceding attention module, which makes the optimization easier.

RGA vs. Other Approaches. For fairness of comparison, we re-implement their designs on top of our baseline and show the results in Table 2.

1) Spatial attention. *CBAM-S* [38] uses a large filter size of 7×7 to learn attention while *FC-C* [23] uses fully connection over the (channel-pooled) spatial feature maps. Non-local(*NL*) [35] takes pairwise relations/affinities as the weights to obtain an aggregated feature for refinement. *SNL* is a simplified scheme of non-local [5], which determines the weight for aggregation using only the source feature itself. NL ignores the mining of the global scope structural information from the relations and only uses them for weighted sum. In contrast, our RGA aims to mine from the relations. It is observed that the weights for aggregation in schemes *NL* and *SNL* are invariant to the target positions [5]. Thanks to the exploration of global structural information and its mining through learnable modeling function, our *RGA-S* achieves the best performance, which is about 2% better than the others in mAP accuracy on CUHK03(L).

To better understand the difference between the non-local *NL* [35] and our *RGA-S*, we visualize their learned pairwise relation/affinity values with respect to three randomly selected target positions in Fig. 4. We find that the relation values are target position invariant for the non-local

Table 2. Performance (%) comparisons of our attention and other approaches, applied on top of our baseline.

Methods		CUHK03 (L)		Market1501	
		R1	mAP	R1	mAP
Baseline	ResNet-50	73.8	69.0	94.2	83.7
Spatial	CBAM-S [38]	77.3	72.8	94.8	85.6
	FC-S [23]	77.0	73.0	95.2	86.2
	NL [35]	76.6	72.6	95.6	87.4
	SNL [5]	77.4	72.4	95.7	87.3
	RGA-S (Ours)	**79.3**	**74.7**	**96.0**	**87.5**
Channel	SE [18]	76.3	71.9	95.2	86.0
	CBAM-C [38]	76.9	72.7	95.3	86.3
	FC-C [23]	77.4	72.9	95.3	86.7
	RGA-C (Ours)	**79.3**	**75.6**	**95.9**	**87.9**
Both	CBAM-CS[38]	78.0	73.0	95.0	85.6
	FC-S//C [23]	78.4	73.2	94.8	85.0
	RGA-SC (Ours)	**81.1**	**77.4**	**96.1**	**88.4**

scheme (top row), which is similar to the observation made by Cao *et al*. [5]. In contrast, thanks to the learned modeling function applied on the relation vector and appearance feature, it drives the pairwise relation function (see Eq. (1)) to better model the relations and makes the learned relations target position adaptive in our scheme (bottom row). For a target position, we observe the feature positions which have similar semantics are likely to have large relation/affinity values. This indicates that our attention model has mined helpful knowledge, *e.g*., clustering-like patterns in semantic space, from the relations for inferring attention.

2) Channel attention. In Squeeze-and-Excitation module (*SE* [18]), they use spatially global average-pooled features to compute channel-wise attention, by using two fully connected (FC) layers with the non-linearity. In comparison with *SE*, our *RGA-C* achieves 3.0% and 3.7% gain in Rank-1 and mAP accuracy. *CBAM-C* [38] is similar to (*SE*) [18] but it additionally uses global max-pooled features. Similarly, *FC-C* [23] uses a FC layer over spatially average pooled features. Before their pooling, the features are further embedded through 1×1 convolutions. Thanks to the exploration of pairwise relations, our scheme *RGA-C* outperforms *FC-C* [23] and *SE* [18], which also use global information, by 1.9% and 3.0% in Rank-1 accuracy on CUHK03. On Market1501, even though the accuracy is already very high, our scheme still outperforms others.

3) Spatial and channel attention. When both spatial and channel attentions are utilized, our models consistently outperform using the channel attention alone or using the spatial attention alone.

Parameters. As shown in Table 3, the number of parameters of the scheme *RGA-S* is less than the *NL* scheme, while the number of parameters of the scheme *RGA-C* is about 2% to 6% larger than other schemes.

Influence of Embedding Functions. We use asymmetric embedding functions (see Eq. (1)) to model the directional relations ($r_{i,j}, r_{j,i}$) between node i and node j. We compare it with symmetric embedding and no embedding in Table 4. We observe that using the feature directly (w/o Embedding) or using symmetric embedding function also significantly outperforms *Baseline* but is clearly inferior to using asymmetric embedding. It indicates that the main improvements come from our new design of relation-based attention learning, in which better relation modeling will deliver better performance. Using asymmetric embedding functions leaves more optimization space.

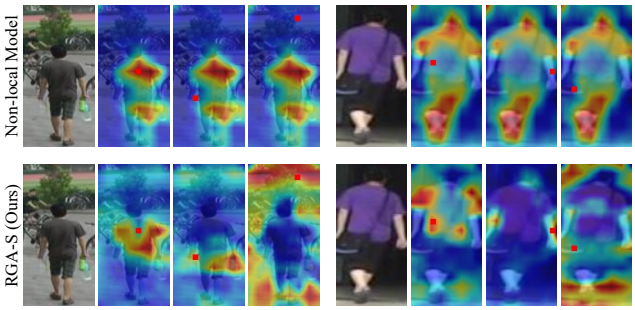


Figure 4. Each three subimages visualize the connecting weights (relation values) from all positions w.r.t three target positions (marked by red squares), for non-local scheme (top row) and our RGA-S scheme (bottom row). For the color intensity, red indicates a large value while blue indicates a small one. We observe that the weights are invariant to the target positions for non-local model but adaptive in our RGA-S. For a target position, the positions with similar semantics usually have large relation values in our RGA-S, which reflects clustering-like patterns.

Table 3. Number of parameters for different schemes (Million).

Baseline	Spatial				Channel			
	CBAM-S	FC-S	NL	RGA-S	CBAM-C	FC-C	SE	RGA-C
25.1	26.1	26.9	30.6	28.3	26.4	26.4	27.6	28.1

Table 4. Influence of embedding functions on performance (%).

Model	CUHK03 (L)		Market1501	
	R1	mAP	R1	mAP
Baseline	73.8	69.0	94.2	83.7
w/o Embedding	78.6	75.2	95.2	87.3
Symmetric	79.4	75.2	95.6	87.4
Asymmetric (Ours)	81.1	77.4	96.1	88.4

Which ConvBlock to Add RGA-SC? We compare the cases of adding the RGA-SC module to different residual blocks. The RGA-SC brings gain on each residual blocks and adding it to all blocks performs the best. Please refer to the supplementary for more details.

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

Table 5 shows the performance comparisons of our relation-aware global attention models (RGA-SC) with the state-of-the-art methods on three datasets. In comparison with the attention based approaches [28, 25, 39, 19] which leverage human semantics (*e.g*. foreground/background,

Table 5. Performance (%) comparisons with the state-of-the-arts on CUHK03, Market1501 and MSMT17.[2]

	Method	CUHK03				Market1501		MSMT17	
		Labeled		Detected					
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Attention -based	MGCAM (CVPR18) [28]	50.1	50.2	46.7	46.9	83.8	74.3	-	-
	AACN (CVPR18) [39]	-	-	-	-	85.9	66.9	-	-
	SPReID (CVPR18) [19]	-	-	-	-	92.5	81.3	-	-
	HA-CNN (CVPR18) [22]	44.4	41.0	41.7	38.6	91.2	75.7	-	-
	DuATM (CVPR18) [27]	-	-	-	-	91.4	76.6	-	-
	Mancs (ECCV18) [32]	69.0	63.9	65.5	60.5	93.1	82.3	-	-
	MHN-6(PCB) (ICCV19) [6]	77.2	72.4	71.7	65.4	95.1	85.0	-	-
	BAT-net (ICCV19) [11]	78.6	76.1	76.2	73.2	95.1	84.7	79.5	56.8
Others	PCB+RPP (ECCV18) [30]	63.7	57.5	-	-	93.8	81.6	68.2	40.4
	HPM (AAAI19) [13]	63.9	57.5	-	-	94.2	82.7	-	-
	MGN(w flip) (MM19) [34]	68.0	67.4	66.8	66.0	95.7	86.9	-	-
	IANet (CVPR19) [17]	-	-	-	-	94.4	83.1	75.5	46.8
	JDGL (CVPR19) [47]	-	-	-	-	94.8	86.0	77.2	52.3
	DSA-reID (CVPR19) [43]	78.9	75.2	78.2	73.1	95.7	87.6	-	-
	OSNet (ICCV19) [51]	-	-	72.3	67.8	94.8	84.9	78.7	52.9
Ours	Baseline	73.8	69.0	70.5	65.5	94.2	83.7	75.7	51.5
	RGA-SC	**81.1**	**77.4**	**79.6**	**74.5**	**96.1**	**88.4**	**80.3**	**57.5**

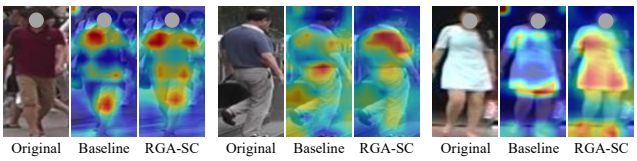


Figure 5. Grad-CAM visualization according to gradient responses: *Baseline* vs. *RGA-SC*.

human part segmentation) and those [22, 27, 32] which learn attention from input images themselves, our *RGA-SC* significantly outperforms them. On the three datasets CUHK03(L)/CUHK03(D), Market1501, and the large-scale MSMT17, in comparison with all other approaches, our scheme *RGA-SC* achieves the best performance which outperforms the second best approaches by 1.3%/1.3%, 0.8%, and 0.7% in mAP accuracy, respectively. The introduction of our RGA-SC modules consistently brings significant gain over our *Baseline*, *i.e.*, **8.4**%/**9.0**%, **4.7**%, and **6.0**% in mAP accuracy, respectively.

4.4. Visualization of Attention

Similar to [38], we apply the Grad-CAM [26] tool to the baseline model and our model for the qualitative analysis. Grad-CAM tool can identify the regions that the network considers important. Fig. 5 shows the comparisons. We can clearly see that the Grad-CAM masks of our RGA model cover the person regions better than the baseline model. The modulation function of our attention leads the network to focus on discriminative body parts.

We visualize the learned spatial attention mask in Fig. 1. The attention focuses on the person and ignores the background. In comparison with the attention approach of CBAM [38] which does not exploit relations, our attention more clearly focuses on the body regions with discriminative information, which benefits from our mining of knowledge from the global scope structural information (where they present clustering-like patterns in semantic space (see the bottom row in Fig. 4). Note that we observe that the head is usually ignored. That is because the face usually has low resolution and is not reliable for differentiating different persons. More visualization results including those on different layers can be found in the supplementary.

5. Conclusion

For person re-id, in order to learn more discriminative features, we propose a simple yet effective Relation-Aware Global Attention module which models the global scope structural information and based on this to infer attention through a learned model. The structural patterns provide some kind of global scope semantics which is helpful for inferring attention. Particularly, for each feature position, we stack the pairwise relations between this feature and all features together with the feature itself to infer the current position's attention. Such feature representation facilitates the use of shallow convolutional layers (*i.e.* shared kernels on different positions) to globally infer the attention. We apply this module to the spatial and channel dimensions of CNN features and demonstrate its effectiveness in both cases. Extensive ablation studies validate the high efficiency of our designs and state-of-the-art performance is achieved.

Acknowledgements

This work was supported in part by NSFC under Grant U1908209, 61632001 and the National Key Research and Development Program of China 2018AAA0101400.

[2]We do not include results on DukeMTMC-reID [48] since this dataset is not publicly released anymore.

References

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

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *TOG*, volume 28, page 24. ACM, 2009. 2

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, volume 2, pages 60–65, 2005. 2

[4] Antoni Buades, Bartomeu Coll, and J-M Morel. Non-local color image denoising with convolutional neural networks. In *CVPR*, 2017. 2

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019. 2, 3, 5, 6, 7

[6] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019. 1, 2, 8

[7] Peihao Chen, Chuang Gan, Guangyao Shen, Wenbing Huang, Runhao Zeng, and Mingkui Tan. Relation attention for temporal action localization. *TMM*, 2019. 2

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *TIP*, 16(8):2080–2095, 2007. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[10] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *ICCV*, volume 2, pages 1033–1038. IEEE, 1999. 2

[11] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *ICCV*, pages 8030–8039, 2019. 1, 8

[12] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. AAAI, 2019. 1

[13] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. *AAAI*, 2019. 8

[14] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *ICCV*, pages 349–356. IEEE, 2009. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 6

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, pages 9317–9326, 2019. 8

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 7

[19] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 7, 8

[20] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 1

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2, 6

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 1, 2, 5, 8

[23] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 3, 6, 7

[24] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, pages 4898–4906, 2016. 1

[25] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *ICME*, 2019. 7

[26] 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 *ICCV*, pages 618–626, 2017. 8

[27] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 8

[28] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2, 7, 8

[29] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 2

[30] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. 2018. 6, 8

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[32] 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 *ECCV*, 2018. 2, 3, 5, 6, 8

[33] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017. 1, 2, 3, 5

[34] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granu-

larities for person re-identification. *ACM Multimedia*, 2018. 8

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1, 2, 5, 6, 7

[36] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018. 6

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018. 2, 6

[38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 1, 2, 3, 5, 6, 7, 8

[39] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018. 1, 2, 7, 8

[40] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognition*, 86:143–155, 2019. 2

[41] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 6

[42] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 2

[43] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 6, 8

[44] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 2

[45] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017. 2

[46] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 6

[47] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 8

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017. 8

[49] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 6

[50] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6

[51] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. *ICCV*, 2019. 8