# Relation-Aware Global Attention

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@mail.ustc.edu.cn {culan, wezeng}@microsoft.com
jinxustc@mail.ustc.edu.cn chenzhibo@ustc.edu.cn

## Abstract

*Attention mechanism aims to increase the representation power by focusing on important features and suppressing unnecessary ones. For convolutional neural networks (CNNs), attention is typically learned with local convolutions, which ignores the global information and the hidden relation. How to efficiently exploit the long-range context to globally learn attention is underexplored. In this paper, we propose an effective Relation-Aware Global Attention (RGA) module for CNNs to fully exploit the global correlations to infer the attention. Specifically, when computing the attention at a feature position, in order to grasp information of global scope, we propose to stack the relations, i.e., its pairwise correlations/affinities with all the feature positions, and the feature itself together for learning the attention with convolutional operations. Given an intermediate feature map, we have validated the effectiveness of this design across both the spatial and channel dimensions. When applied to the task of person re-identification, our model achieves the state-of-the-art performance. Extensive ablation studies demonstrate that our RGA can significantly enhance the feature representation power. We further demonstrate the general applicability of RGA to vision tasks by applying it to the scene segmentation and image classification tasks resulting in consistent performance improvement.*

## 1. Introduction

Besides the fields of machine translation [3], speech recognition [10], attention mechanism has been demonstrated effective in many vision fields such as image captioning [49, 8], classification [42, 47], object detection [2], person re-identification [56, 32, 26, 41, 30, 33]. Attention tells "where" (*i.e.*, attentive spatial location) and "what" (*i.e.*, attentive channels) to focus to enhance the task-orientated representation power of features.

When facing a visual scene, human can efficiently pay attention to the interested parts [11]. Attention [42, 47, 8, 41] intends to simulate the human visual system to pay different attention to different features. It acts as a modulator that suppresses the unnecessary features and reinforces the important ones, by multiplying a mask. It balances the relative importance among different components. For CNNs, convolutional operations with limited receptive fields are usually used to compute the attention [8, 47, 42, 41]. However, locally learned attention is hard to reach optimal representativeness. One solution is to use large filter size in the convolution layer [47]. The other solution is to stack deep layers [42] which increases the network size greatly. Besides, the studies in [34] show that the effective receptive field of CNN only takes up a fraction of the full theoretical

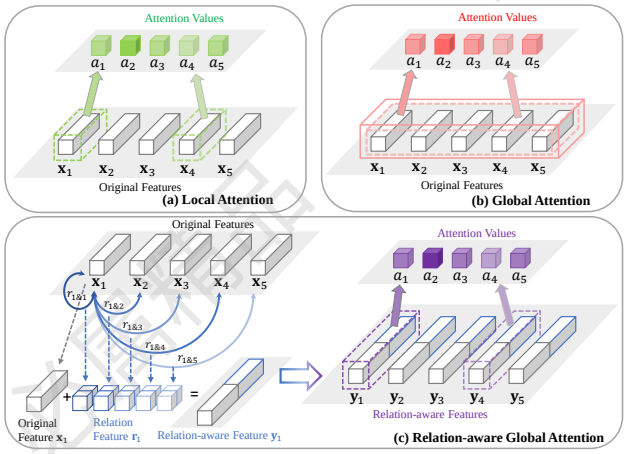



Figure 1. Illustrations of learning attention values $a_1, \cdots, a_5$ for five feature vectors $\mathbf{x}_1, \cdots, \mathbf{x}_5$. (a) Local attention: learn attention locally (*e.g.*, based on individual feature as shown). (b) Global attention: learn attention jointly from all the 5 feature vectors. (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 relation information is represented by 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 of global awareness and (b) that may require a large number of parameters, our proposed attention exploits global information and facilitates the use of shared function for deriving the attention individually.

*This work is done when Zhizheng Zhang is an intern at MSRA.

receptive field. These solutions cannot ensure the effective capture of global information, thus limiting the exploration of correlations among features for attention learning.

Inspired by the common sense that people often know the importance of something through comparison, in this paper, we propose a simple yet effective Relation-Aware Global Attention (RGA) module to globally learn the attention. For each node (*e.g.*, a feature vector of a spatial position), we take the pairwise relations of all the nodes with respect to the current node to represent the global structure information to learn attention. As illustrated in Fig. 1 (c), for each node (feature vector), we grasp global information from all nodes by calculating their pairwise relations *i.e.*, correlation/similarities, and then stack these pairwise relation values together with the original feature to be the feature of this node for deriving its attention. With global structure information embedded in each local position, convolution operations are used to infer the attention intensity.

We showcase the effectiveness of the RGA in the task of person re-identification, where abundant previous works strive to address the challenges by leveraging attention designs [56, 32, 26, 41, 30, 33]. Extensive ablation studies demonstrate the effectiveness of our design in comparison with other attention mechanisms. The models with our proposed attention appended achieve the state-of-the art performance on the benchmark datasets CUHK03 [29], Market1501 [57], and MSMT17[46]. Moreover, when applied on top of the strong non-local neural networks, our RGAM can further increase the accuracy. This also indicates our relation-aware global attention is complementary to the non-local filter idea [44].

To demonstrate the generality of the proposed relation-aware global attention modules, we further conduct experiments for the task of scene segmentation on the popular dataset Cityscapes [12], and image classification on the CIFAR dataset [27]. Experimental results show that our global attention models consistently improve the performance over the baselines.

In summary, we have made three major contributions:

- We propose to globally determine the attention by taking a global view of the mutual relations among the features. Specifically, for a feature node, we propose a compact representation by stacking the pairwise relations with respect to this node and the feature itself.
- We design a relation-aware global attention (RGA) module based on the relation-based compact representation and shallow convolutional layers. We apply such design to spatial and channel dimensions respectively, and demonstrate the effectiveness of this global attention module in both cases and the their combination case.
- Extensive ablation studies on person re-identification demonstrate the effectiveness of our proposed RGA, which provides the state-of-the-art performance. We also demonstrate the general applicability of RGA to other vision tasks such as scene segmentation and image classification.

## 2. Related Work

**Attention.** Attention aims to focus on important features and suppress irrelevant features. Intuitively, to have a good sense of whether a local region is important or not, one should know the global scene. For CNNs, most of the current selective attention modules learn attention from limited local contexts. Wang *et al.* [42] propose an encoder-decoder style attention module by stacking many convolutional layers. In [53], a non-local block [44] is inserted before the encoder-decoder style attention module to enable attention learning based on globally refined features. In the Convolutional Block Attention Module [47], a convolution layer with a large filter size of $7\times7$ is applied over the cross-channel pooled spatial features to produce a spatial attention map. Limited by the practical receptive fields, all these approaches are not efficient in capturing the large scope information to globally determine the spatial attention. In [33], for spatial attention model, two FC layers are applied to the cross-channel average pooled feature map to generate a spatial attention map. Each spatial position corresponds to a FC node with unshared parameters. This may require a large number of parameters and make the training difficult. Hu *et al*. [24] use two FC layers over spatially average-pooled features to compute channel-wise attention. All these methods do not explicitly exploit the inter-channel (or spatial) relations to enable the assessment of relative importance.

We address these problems by proposing a Relation-Aware Global Attention module. For each feature position, we embed the pairwise relation between this position and all the positions to capture the global information. Then, two convolutional layers with translation invariant property are used to efficiently exploit the global structural information.

**Non-local/Global Information Exploration.** Exploration of non-local/global information has been demonstrated to be very useful for image denoising [6, 13, 7], texture synthesis [15], super-resolution [18], inpainting [5], and even high level tasks such as image recognition and object segmentation [44]. In [6, 44], they adopt the non-local mean idea which computes a weighted summation of the non-local and local pixels/features as the refined representation of the target pixels/features. The weight value connecting every two positions represents their relationship and is calculated from the similarity/correlation of the pair. All the positions (nodes) with the mutually connected edges construct a global graph. In [19], the graph structure (adjacency) is used as part of the features for graph matching for fewshot 3D action recognition.

Note for one node, all its edges and the corresponding lo-

cation information of the connected nodes deliver the global relationships with all the nodes. This information can tell the clustering states and spatial clustering patterns. In this work, we leverage this global structural information associated with a node to globally learn the attention (that will be used to module the feature of this node) using convolutional layers within general CNN frameworks. This is in contrast to the weighted summarization operation in non-local mean. We will experimentally demonstrate their complementary roles in Section 4.

## 3. Relation-Aware Global Attention

For attention, to have a good sense of the importance of a feature, one should know the others for an objective assessment. Thus, the global information is essential. We thus propose a relation-aware global attention module which makes use of the structural relation information. In the following, we first give the problem formulation and present our main idea in Subsection 3.1. For CNNs, we decouple the attention into a spatial relation-aware global attention and a channel relation-aware global attention and elaborate them in Subsection 3.2 and 3.3, respectively. Finally, we simply introduce the joint using of them in Subsection 3.4.

### 3.1. Problem 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, our goal is to learn a mask denoted by $\mathbf{a} = (a_1, \cdots, a_N) \in \mathbb{R}^N$ for the $N$ features to weight them according to their relative importance. The updated $i^{th}$ feature through attention is $\mathbf{z}_i = a_i \mathbf{x}_i$.

To learn the attention value $a_i$ of the $i^{th}$ feature vector, there are two common strategies as illustrated in Fig. 1 (a) and (b). **(a) Local attention**: each feature determines its attention locally, *e.g.*, using a shared transformation function $\mathcal{F}$ on itself, *i.e.*, $a_i = \mathcal{F}(\mathbf{x}_i)$. This local strategy does not fully exploit the correlations with other features [8]. For vision tasks, deep layers [42] or large-sized kernels [47] are used to make the attention learning more global. **(b) Global attention**: one straightforward solution is to use all the features together to jointly learn attention, *e.g.*, using fully connected operations. However, this is usually computationally expensive and requires a large number of parameters especially when the number $N$ of features is large [33].

In contrast to these strategies, we propose a relation-aware global attention that enables i) the exploitation of global structural information, and ii) the use of shared transformation function for different individual features to derive the attention. For visual tasks, the later makes it possible to globally compute the attention by using local convolutional operations. We illustrate our basic idea in Fig. 1 **(c) Proposed relation-aware global attention**.

The main idea is to exploit the pairwise relation (*e.g.* affinity/similarity) related to the $i^{th}$ feature to represent this feature node's global structural information. Specifically, we use $r_{i,j}$ to represent the affinity between the $i^{th}$ feature and the $j^{th}$ feature. For the $i^{th}$ feature $\mathbf{x}_i$, 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 using a shared transformation function. Note that $\mathbf{y}_i$ contains global information.

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 $\mathcal{E} = \{r_{i,j} \in \mathbb{R}^1, i = 1, \cdots, N \text{ and } j = 1, \cdots, N\}$. 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 $r_{i,j} = R(i,j)$. $\mathbf{r}_i = [R(i,:), R(:,i)]$.

### 3.2. Spatial Relation-Aware Attention

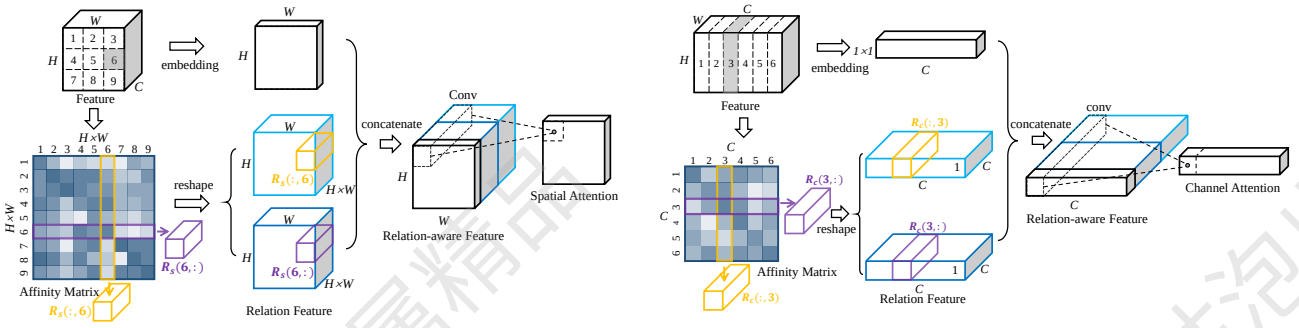
Given an intermediate feature map (tensor) $X \in \mathbb{R}^{C \times H \times W}$ of width $W$, height $H$, and $C$ channels from a CNN layer, we design a spatial relation-aware attention block, namely RGA-S, for learning a spatial attention map of size $H \times W$. We take the $C$-dimensional feature 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. 2 (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 $\theta_s$ and $\phi_s$ are two embedding functions implemented by a $1 \times 1$ 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 a 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 for all the nodes by an affinity matrix $R_s \in \mathbb{R}^{N \times N}$.

For the $i^{th}$ feature/node, we collect its pairwise relations with all the nodes, *i.e.*, a relation vector $\mathbf{r}_i = [R_s(i,:), R_s(:,i)] \in \mathbb{R}^{2N}$, to represent the global structural information. As illustrated in Fig. 2 (a), the sixth row and the sixth column of the affinity matrix $R_s$, *i.e.* $\mathbf{r}_6 = [R_s(6,:), R_s(:,6)]$,




(a) Spatial Relation-Aware Global Attention

(b) Channel Relation-Aware Global Attention

Figure 2. 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.

is taken as the relation features for deriving the attention of the sixth spatial position.

To infer 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 make use of both the global mutual relations and the local original information. Considering these two kinds of information are not in the same feature domain, we embed them respectively and concatenate them to have 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 $\psi_s$ and $\varphi_s$ denote the embedding functions for the feature itself and the global relations, respectively. They are both implemented by a spatial $1 \times 1$ convolutional layer followed by Batch Normalization and ReLU activation, *i.e.*, $\psi_s(\mathbf{x}_i) = ReLU(W_\psi \mathbf{x}_i)$, $\varphi_s(\mathbf{r}_i) = ReLU(W_\varphi \mathbf{x}_i)$, 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 be 1. Then $\widetilde{\mathbf{y}}_i \in \mathbb{R}^{N/s_1+1}$. Note that other convolution kernel sizs *e.g.* $3\times3$ can also be used. We found they achieve very similar performance and we will use $1 \times 1$ convolutional layer for lower complexity.

The spatial attention value $a_i$ for the $i^{th}$ feature/node is then obtained by:

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

where $W_1$ and $W_2$ are implemented by $1 \times 1$ convolution followed by batch normalization. $W_1$ shrinks the channel dimension with a ratio of $s_2$ and $W_2$ transforms the channel dimension to 1.

Note that all these operations are achieved by convolution operations and the global relations are also exploited.

### 3.3. Channel Relation-Aware Attention

Given an intermediate feature map (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 vector 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 (*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:

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

where $\theta_c$ and $\phi_c$ are two embedding functions that are shared among feature nodes. We achieve the embedding by first spatially flattening input tensor $X$ into $X' \in \mathbb{R}^{(HW) \times C \times 1}$ and then using a $1 \times 1$ convolution layer with batch normalization followed by ReLU activation to perform a transformation on $X'$. As illustrated in Fig. 2 (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 collect its pairwise relations with all the nodes, *i.e.*, a relation vector $\mathbf{r}_i = [R_c(i,:), R_c(:,i)] \in \mathbb{R}^{2C}$, to represent the global structural information. As illustrated in Fig. 2 (b), the third row and the third column of the affinity matrix $R_c$, *i.e.* $\mathbf{r}_3 = [R_c(3,:), R_c(:,3)]$, is taken as the relation features for deriving the attention of the third channel node.

To infer the attention of the $i^{th}$ feature/node, similar to the derviation of 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 connection operations across channels.

### 3.4. Joint Relation-Aware Global Attention

The spatial attention RGA-S and channel attention RGA-C play complementary roles. Both can be applied in any stage of convolution networks and trained in an end-to-end manner without any additional auxiliary supervision. They can be used alone or combined. We suggest to jointly use them in a sequential manner because of its lower training difficulty relative to the parallel manner. Take the sequential spatial-channel combination as example. Given an intermediate feature map (tensor) $X$, after the modulation by the spatial RGA, we get the feature map $\hat{X}$. Afterwards, channel RGA is derived from $\hat{X}$ and applied on $\hat{X}$. We will discuss the results of using each alone versus in combination, and sequential aggregation vs. parallel aggregation in the next section.

## 4. Experiments

We showcase the effectiveness of our Relation-Aware Global Attention on the person re-identification task in Subsection 4.1. Extensive ablation studies demonstrate the effectiveness of our designs. Our models achieve the state-of-the-art performance on the CUHK03 [29], Market1501 [57] and MSMT17[46] datasets. As extension experiments, we also investigate our models on scene segmentation in Subsection 4.2 and image classification in Subsection 4.3.

### 4.1. Experiments on Person Re-identification

Person re-identification (re-ID) aims to match a specific person in different occasions from the same camera or across multiple cameras. It has become increasingly popular in both research and industry community due to its application and research significance [58, 28]. Generally, given an input image, we employ a convolutional neural network to obtain a feature vector. Re-identification is to find the images with the same identity by matching the feature vectors of images (based on feature distance).

#### 4.1.1 Implementation Details and Datasets

**Network Settings**. Following the common practices in re-ID systems [4, 39, 52, 1, 54], we take ResNet-50 [21] to build our baseline network and apply our RGA modules to the RetNet-50 backbone for effectiveness validation. Similar to [39, 54], the last spatial down-sampling operation in the conv5_x block of ResNet-50 is removed. Except for special instructions, in our re-ID experiments, we adopt the proposed RGA modules after all of the four residual blocks (including conv2_x, conv3_x, conv4_x and conv5_x). Within RGA modules, we set the ratio parameters $s_1$ and $s_2$ to be 8. In all our re-ID experiments, we use both identification (classification) loss with label smoothing [40] and triplet loss with hard mining [23] as supervision signals. Note that we do not implement re-ranking [61] for clear comparisons in all our experiments. More details please see our supplementary.

**Training**. We use the commonly used data augmentation strategies of random cropping [45], horizontal flipping, and random erasing [62, 45, 41]. The input image size is set as $256 \times 128$ for all the datasets. The backbone network is pre-trained on ImageNet [14]. We adopt Adam optimizer with the learning rate of $8 \times 10^{-4}$ and the weight decay of $5 \times 10^{-4}$. Please refer to the supplementary for more details.

**Datasets and Evaluation Metrics**. We conduct experiments on four public person re-ID datasets, *i.e.*, CUHK03 [29], Market1501 [57], DukeMTMC-reID [59] and the large-scale MSMT17 [46]. To save space, the introduction of the datasets are placed in our supplementary. For CUHK03, we use the new protocol as used by [61, 60, 22]. 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.

#### 4.1.2 Ablation Study

We perform the ablation studies on the CUHK03 (with the Labeled bounding box setting) and Market1501 datasets.

**RGA related Models versus Baseline.** Table 1 shows the comparisons of our spatial RGA (*RGA-S*), channel RGA (*RGA-C*), mixed RGA with both channel and spatial RGA, and the baseline. We have the following observations.

**1)** Either our spatial RGA model (*RGA-S*) or channel RGA model (*RGA-C*) *significantly* improves the performance over our powerful baseline. On CUHK03, *RGA-S*, *RGA-C*, and the combined spatial and channel RGA model *RGA-SC* significantly outperform the baseline by **5.7%**, **5.9%**, and **7.5%** respectively on mAP accuracy, and **5.5%**, **5.2%**, and **6.6%** respectively on Rank-1 accuracy. On Market1501, even though the performance on baseline is already very high (*i.e.* 83.7% on mAP), our *RGA-S* and *RGA-C* improve the mAP accuracy by 3.1% and 3.5% respectively.

**2)** For learning attention, even without taking the visual features (Vis.), *i.e.*, feature itself, as part of the input, the proposed global relation representation itself (*RGA-S w/o Vis.* or *RGA-C w/o Vis.*) can significantly outperform the baseline's performance, *e.g.* 4.4% or 4.3% gain over baseline in Rank-1 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.* outperforms the baseline, but is inferior to our scheme *RGA-S* or *RGA-C* by 2.5% or 1.2% in Rank-1 accuracy on CUKH03.

**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*). Sequential spatial-channel *RGA-SC* achieves the

best performance, 1.8% and 1.6% higher than *RGA-C* and *RGA-S*, respectively, in mAP accuracy on CUHK03. Parallel optimization is more difficult than the sequential one.

Table 1. Performance (%) comparisons of our models with the baseline and the effectiveness of the global relation representation.

| 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 Vis. | 78.2 | 74.0 | **95.4** | 86.7 |
| | RGA-S | **79.3** | **74.7** | **95.4** | **86.8** |
| Channel | RGA-C w/o Rel. | 77.8 | 73.7 | 94.7 | 84.8 |
| | RGA-C w/o Vis. | 78.1 | **74.9** | **95.4** | 87.1 |
| | RGA-C | **79.0** | **74.9** | **95.4** | **87.2** |
| Both | RGA-S//C | 77.3 | 73.4 | 95.3 | 86.6 |
| | RGA-CS | 78.6 | 75.5 | 95.3 | 87.8 |
| | RGA-SC | **80.4** | **76.5** | **95.8** | **88.1** |

**RGA versus Other Attention Mechanisms.** Table 2 compares the performance of our attention modules with the other state-of-the-art attention designs. For fairness of comparison, we re-implement their attention designs on top of our baseline. **1) Channel attention.** There are several channel attention designs. In Squeeze-and-Excitation module (*SE*) [24], 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 2.7% and 3.0% gain in Rank-1 and mAP accuracy. *CBAM-C* [47] is similar to (*SE*) [24] but it additionally uses global max-pooled features. Similarly, *FC-C* [33] uses a FC layer over spatially average pooled features. Before their pooling, the features are further embedded through $1 \times 1$ convolutions. Thanks to the exploration of pairwise relations, our scheme *RGA-C(Ours)* outperforms *FC-C*[33] and *SE* [24] which also use global information by 1.6% and 2.7% in Rank-1 accuracy on CUHK03. On Market1501, even the accuracy is already very high, our scheme still outperforms others. **2) Spatial attention.** For spatial attention designs, *CBAM-S* [47] uses large filter size of $7\times7$ to learn attention while *FC-C*[33] uses fully connection over the spatial feature maps. Our scheme achieves the best performance, which is 2% better than the others in Rank-1 accuracy on CUHK03. **3) Spatial and channel attention.** When both spatial and channel attentions are utilized, our models consistently outperform both the channel attention and spatial attention.

**RGA versus Non-local Blocks.** Both non-local neural network [44] and our RGA utilize the local and non-local pairwise relations but with rather different purposes. The non-local blocks function in a way similar to the non-local means [44], which computes the response at a position as a weighted sum of the features at all positions. The pairwise relation, in terms of similarity/affinity, is used as the weight. In contrast, we leverage the collection of pairwise relations to represent the overall relation between the current position's feature and all the positions' features for inferring the attention. The two modules are complementary in concepts and functions. We experimentally demonstrate this and show results in Table 3. *Non-local* denotes the scheme after integrating non-local blocks [44] to our baseline and *Non-local + RGA-S* denotes that our spatial RGA modules are also integrated after the non-local blocks. On top of the non-local networks, the introduction of spatial-wise relation-aware global attention significantly further improves the performance, *i.e.*, by **4.0%** and **4.2%** in Rank-1 and mAP accuracy, respectively on CUHK03.

Table 2. Performance (%) comparisons of our attention and other attention designs, 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 [47] | 77.3 | 72.8 | 94.8 | 85.6 |
| | FC-S [33] | 77.0 | 73.0 | 95.2 | 86.2 |
| | RGA-S (Ours) | **79.3** | **74.7** | **95.4** | **86.8** |
| Channel | SE [24] | 76.3 | 71.9 | 95.2 | 86.0 |
| | CBAM-C [47] | 76.9 | 72.7 | 95.3 | 86.3 |
| | FC-C [33] | 77.4 | 72.9 | 95.3 | 86.7 |
| | RGA-C (Ours) | **79.0** | **74.9** | **95.4** | **87.2** |
| Both | CBAM-CS[47] | 78.0 | 73.0 | 95.0 | 85.6 |
| | FC-S//C [33] | 78.4 | 73.2 | 94.8 | 85.0 |
| | RGA-SC (Ours) | **80.4** | **76.5** | **95.8** | **88.1** |

Table 3. Performance(%) comparisons of non-local models and our spatial RGA to validate their complementary property.

| Model | CUHK03(L) | | Market1501 | |
|---|---|---|---|---|
| | R1 | mAP | R1 | mAP |
| Baseline | 73.8 | 69.0 | 94.2 | 83.7 |
| Non-local | 76.6 | 72.6 | **95.6** | 87.4 |
| RGA-S | 79.3 | 74.7 | 95.4 | 86.8 |
| Non-local + RGA-S | **80.6** | **76.8** | **95.6** | **88.0** |

**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 best. Please refer to the supplementary for more details.

#### 4.1.3 Comparison with the State-of-the-Art

Table 4 shows the performance comparisons of our relation-aware global attention models (RGA-SC) with the state-of-the-art methods on the commonly used three datasets. In comparison with the attention based approaches [38, 35, 48, 26] which leverage human semantics (*e.g.* foreground/background, human part segmentation) and those [30, 37, 41] which learn attention from input images themselves, our *RGA-CS* significantly outperforms. Even with-

Table 4. Performance (%) comparisons with the state-of-the-arts on CUHK03, Market1501 and DukeMTMC-reID.

| Method | | CUHK03 Labeled | | CUHK03 Detected | | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP |
| Basic-CNN (ResNet-50) | IDE(ECCV18) [39] | 43.8 | 38.9 | - | - | 85.3 | 68.5 | 73.2 | 52.8 |
| | Gp-reid(ArXiv18) [1] | - | - | - | - | 92.2 | 81.2 | 85.2 | 72.8 |
| Attention-based | MGCAM(CVPR18) [38] | 50.1 | 50.2 | 46.7 | 46.9 | 83.8 | 74.3 | - | - |
| | MaskReID(ArXiv18) [35] | - | - | - | - | 90.0 | 70.3 | 78.9 | 61.9 |
| | AACN(CVPR18) [48] | - | - | - | - | 85.9 | 66.9 | 76.8 | 59.3 |
| | SPReID(CVPR18) [26] | - | - | - | - | 92.5 | 81.3 | 84.4 | 71.0 |
| | HA-CNN(CVPR18) [30] | 44.4 | 41.0 | 41.7 | 38.6 | 91.2 | 75.7 | 80.5 | 63.8 |
| | DuATM(CVPR18) [37] | - | - | - | - | 91.4 | 76.6 | 81.8 | 64.6 |
| | Mancs(ECCV18) [41] | 69.0 | 63.9 | 65.5 | 60.5 | 93.1 | 82.3 | 84.9 | 71.8 |
| Others | PCB+RPP(ECCV18) [39] | 63.7 | 57.5 | - | - | 93.8 | 81.6 | 83.3 | 69.2 |
| | HPM(AAAI19) [17] | 63.9 | 57.5 | - | - | 94.2 | 82.7 | 86.6 | 74.3 |
| | MGN (MM18) [43] | 68.0 | 67.4 | 66.8 | 66.0 | 95.7 | 86.9 | **88.7** | **78.4** |
| | DSA-reID(CVPR19) [54] | 78.9 | 75.2 | **78.2** | 73.1 | 95.7 | 87.6 | 86.2 | 74.3 |
| Ours | Baseline | 73.8 | 69.0 | 70.5 | 65.5 | 94.2 | 83.7 | 83.4 | 71.2 |
| | RGA-SC | **80.4** | **76.5** | 77.4 | **73.3** | **95.8** | **88.1** | 86.1 | 74.9 |

Table 5. Performance (%) comparisons with the state-of-the-arts on MSMT17. Note that since MSMT17 is a newly released large-scale dataset, there are only a few works available for comparisons.

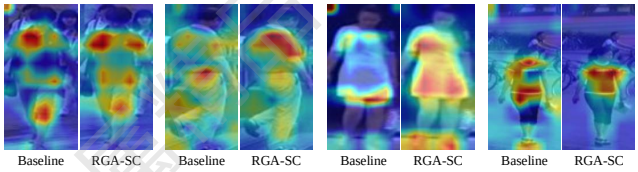
| Method | R1 | mAP |
|---|---|---|
| GoogLeNet(CVPR18) [46] | 47.6 | 23.0 |
| PDC(CVPR18) [46] | 58.0 | 29.7 |
| GLAD(CVPR18) [46] | 61.4 | 34.0 |
| Baseline | 75.7 | 51.5 |
| RGA-SC (Ours) | **81.3** | **56.3** |

out leveraging the human dense semantics as in [54], our schemes achieve superior performance on CUHK03 (with labeled bounding boxes) (1.3% gain in mAP) and Market1501, and competitive performance on CUHK03 (with detected bounding boxes) and DukeMTMC-reID. Both *MGN* [43] and *HPM* [17] ensemble local features at multiple granularities and are better than our scheme on DukeMTMC-reID. We believe we can achieve a better performance when we replace their backbone ResNet-50 by our RGA-SC network.

We also evaluate our model on the latest large-scale dataset MSMT17[46] which consists of 126,441 bounding boxes of 4,101 identities and show the results in Table 5. On this challenging dataset, our RGA model outperforms the baseline by 5.6% and 4.8% on Rank-1 and mAP accuracy respectively. We also achieve the best performance in comparision with the state-of-the-art approaches.

#### 4.1.4 Visualization of Attention

Similar to [47], we apply the Grad-CAM[36] 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.3 (a) 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.




(a) Grad-CAM[36] visualization according to gradient responses.


(b) Visualization of Spatial Attention Mask.

Figure 3. Visualization: (a) Grad-CAM[36] visualization: ResNet-50 baseline vs. ResNet-50 + RGA-SC; (b) The spatial attention mask produced by our proposed RGA-SC after conv5_x block.

We visualize the learned spatial attention mask in Fig.3 (b). The attention focuses on the person and ignores the background. We observe the head is usually ignored. That is because the face is not frontal or has low resolution and is not reliable for differentiating different persons.

### 4.2. Experiments on Scene Segmentation

The goal of scene segmentation is to predict the semantic categories including stuffs (*e.g.* road, sky) and objects (*e.g.* person, car) at pixel level. It favors many practical applications, such as automatic driving, robot sensing and image editing [16] and it also attracts a lot of interests in research committee. We evaluate our proposed RGA modules on scene segmentation task and show our analysis on the popular Cityscapes [12] dataset which consists of 5,000

images with 2048×1024 resolution.

**Implementation Details**. We use the ResNet-101 pre-trained on ImageNet with make some modifications (by replacing the convolutions within the last two blocks by dilated convolutions with dilation rates being 2 and 4) following PSPNet [55] as the backbone. We use the similar training setting as in [50].

**Experiments and Results**. PSPNet [55] uses pyramid pooling module (PPM) while DeepLabv3 [9] adopts atrous spatial pyramid pooling (ASPP) respectively to capture information of multiple scales to tackle multi-scale objects. By adding their modules to the ResNet-101 backbone, we have two stronger baseline schemes *ResNet-101 + PPM* [55] and *ResNet-101 + ASPP* [9]. On top of the three baselines, we add our proposed RGA-SC module to demonstrate its effectiveness, respectively. We apply our RGA-SC module on the feature map of the last convolutional layer and concatenate it with the original features of the baseline network. Table 6 shows the comparison results. We can see on the validation set, the introduce of our RGA-SC module brings an improvement of 3.7%, 1.9%, and 1.1%, respectively in comparison with their respective baselines.

**Discussion.** In our framework, for the tasks with the same sized images in a dataset, we do not need any special design. For tasks with variable image sizes in a dataset, *e.g*. COCO object detection dataset [31], we need to adopt RoIAlign [20] to pool the arbitrary sized spatial features to a fixed sized feature map to assure the length of features, where 1×1 convolution is performed on, is equal at dataset level. For the scene segmentation task, follow the common practice [55], the original images of size 1024×2048 are cropped of size as 769×769 for training while the original image size 1024×2048 is used for testing. We adopt RoIAlign [20] to pool the arbitrary sized spatial features to a fixed sized feature map to solve the varied sizes problem. More study on other tasks with various input sizes will be left as future work.

Table 6. Performance (mean IoU %) comparisons on the validation set of Cityscapes. We evaluate the effectiveness of our proposed RGA module on three baselines.

| Method | Mean IoU (Train.) | Mean IoU (Val.) |
|---|---|---|
| ResNet-101 | 83.8 | 74.8 |
| ResNet-101 + PPM [55] | 85.0 | 77.0 |
| ResNet-101 + ASPP [9] | 85.9 | 78.5 |
| ResNet-101 + RGA-SC | 86.1 | 78.5 |
| ResNet-101 + PPM + RGA-SC | 85.2 | 78.9 |
| ResNet-101 + ASPP + RGA-SC | 86.8 | 79.6 |

### 4.3. Experiments on Image Classification

We further investigate the effectiveness of the RGA on image classification task on CIFAR-10 and CIFAR-100 dataset [27]. Each image has 32×32 pixels.

**Implementation Details**. We conduct experiments based on two representative network architecture ResNet [21] and DenseNet [25]. For the ResNet, we apply a RGA-SC module after the first residual block which outputs a feature maps of size $32 \times 32 \times 16$, where 16 is the number of channels. We set the dimension-reduction ratio parameters $s_1$ and $s_2$ in our RGA-SC module to be 4. For the DenseNet, we apply a RGA-SC module with $s_1$ and $s_2$ set to 8 after the first dense block which outputs a feature maps of size $16 \times 16 \times 108$. During training, we adopt a standard data augmentation of cropping and mirroring as in [21, 25, 51].

**Experiment Results**. Tabel 7 shows the comparisons with the baselines. To demonstrate the reduction of classification error is due to the effectiveness of our proposed attention mechanism instead of the increase of parameters, we further compare them with the deeper baseline models with the same network structures but more parameters. We can see our design can achieve stable improvement in comparison to the baselines and the deeper models.

Table 7. Performance comparisons in terms of classification error (%) on CIFAR-10 (C-10) and CIFAR-100 (C-100).

| Model | Depth | Params | C-10 | C-100 |
|---|---|---|---|---|
| ResNet (Baseline) | 110 | 1.7M | 6.40 | 27.62 |
| ResNet | 218 | 3.5M | 6.26 | 25.71 |
| ResNet + RGA-SC | 110+4 | 3.0M | **5.91** | **25.36** |
| DenseNet (k=12, Baseline) | 100 | 7.0M | 4.08 | 20.97 |
| DenseNet (k=12) | 115 | 9.42M | **3.93** | 20.5 |
| DenseNet (k=12) + RGA-SC | 100+4 | 7.3M | 4.06 | **20.15** |
| DenseNet-BC (k=12, Baseline) | 100 | 0.8M | 4.47 | 22.78 |
| DenseNet-BC (k=12) | 115 | 0.97M | 4.41 | 22.13 |
| DenseNet-BC (k-12) + RGA-SC | 100+4 | 0.85M | **4.32** | **21.18** |

## 5. Conclusion and Future Work

In this paper, we propose a simple yet effective Relation-Aware Global Attention module for CNNs to exploit the global structural information, by leveraging the mutual relations among features. 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 convlutional 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 are achieved for the person re-identification task. For other vision tasks such as scene segmentation and image classification, our RGA module also shows its superiority over the baselines.

The future works include the applications of RGA to other vision tasks, *e.g*., object detection, pose estimation, as well as large scale image classification (*e.g*. on ImageNet).

## References


[1] J. Almazan, B. Gajic, N. Murray, and D. Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339*, 2018. 5, 7

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2014. 1

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[4] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu. Deepperson: Learning discriminative deep features for person re-identification. *arXiv preprint arXiv:1711.10658*, 2017. 5

[5] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009. 2

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

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

[8] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, pages 5659–5667, 2017. 1, 3

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 8

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *NeurIPS*, pages 577–585, 2015. 1

[11] M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201, 2002. 1

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 2, 8

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

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

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

[16] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 7

[17] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang. Horizontal pyramid matching for person re-identification. *arXiv preprint arXiv:1804.05275*, 2018. 7

[18] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, pages 349–356. IEEE, 2009. 2

[19] M. Guo, E. Chou, D.-A. Huang, S. Song, S. Yeung, and L. Fei-Fei. Neural graph matching networks for fewshot 3d action recognition. In *ECCV*, pages 653–669, 2018. 2

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 8

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 8

[22] L. He, Z. Sun, Y. Zhu, and Y. Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018. 5

[23] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[24] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2, 6

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 8

[26] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 1, 2, 6, 7

[27] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2, 8

[28] B. Lavi, M. F. Serj, and I. Ullah. Survey on deep learning techniques for person re-identification task. *arXiv preprint arXiv:1807.05284*, 2018. 5

[29] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2, 5

[30] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 1, 2, 6, 7

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 8

[32] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *TIP*, pages 3492–3506. 1, 2

[33] Y. Liu, Z. Yuan, W. Zhou, and H. Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 1, 2, 3, 6

[34] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, pages 4898–4906, 2016. 1

[35] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*, 2018. 6, 7

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 7

[37] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 6, 7

[38] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 6, 7

[39] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. 2018. 5, 7

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[41] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018. 1, 2, 5, 6, 7

[42] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017. 1, 2, 3

[43] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. *ACM Multimedia*, 2018. 7

[44] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2, 6

[45] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018. 5

[46] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 2, 5, 7

[47] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 1, 2, 3, 6, 7

[48] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018. 6, 7

[49] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 1

[50] Y. Yuan and J. Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 8

[51] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang. Interleaved group convolutions. In *ICCV*, pages 4373–4382, 2017. 8

[52] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 5

[53] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 2

[54] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 5, 7

[55] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 8

[56] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017. 1, 2

[57] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 5

[58] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 5

[59] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017. 5

[60] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *TCSVT*, 2018. 5

[61] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 5

[62] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 5