Relation Network for Person Re-identification

Hyunjong Park, Bumsub Ham[*]
School of Electrical and Electronic Engineering, Yonsei University
hyunpark@yonsei.ac.kr, bumsub.ham@yonsei.ac.kr

Abstract

Person re-identification (reID) aims at retrieving an image of the person of interest from a set of images typically captured by multiple cameras. Recent reID methods have shown that exploiting local features describing body parts, together with a global feature of a person image itself, gives robust feature representations, even in the case of missing body parts. However, using the individual part-level features directly, without considering relations between body parts, confuses differentiating identities of different persons having similar attributes in corresponding parts. To address this issue, we propose a new relation network for person reID that considers relations between individual body parts and the rest of them. Our model makes a single part-level feature incorporate partial information of other body parts as well, supporting it to be more discriminative. We also introduce a global contrastive pooling (GCP) method to obtain a global feature of a person image. We propose to use contrastive features for GCP to complement conventional max and averaging pooling techniques. We show that our model outperforms the state of the art on the Market1501, DukeMTMC-reID and CUHK03 datasets, demonstrating the effectiveness of our approach on discriminative person representations.

Introduction

Person re-identification (reID) is one of fundamental tasks in computer vision, with the purpose of retrieving a particular person from a set of pedestrian images, captured by multiple cameras. It has been getting a lot of attention in recent years, due to the wide range of applications including pedestrian detection (Zheng et al. 2017b) and multi-person tracking (Tang et al. 2017). This problem is very challenging, since pedestrians have different attributes (*e.g.*, clothing, gender, hair), and the pictures of them are taken under different conditions, such as illumination, occlusion, background clutter, and camera types. Remarkable advances in convolutional neural networks (CNNs) over the last decade allow to obtain person representations (Lin et al. 2017a; 2017b; Ge et al. 2018) robust to these factors of variations, especially for human pose, and they also enables learning met-

rics (Zhang, Xiang, and Gong 2016; Chen et al. 2017) for computing the similarities of person features.

Person reID methods using CNNs typically focus on extracting a global feature of a person image (Lin et al. 2017a; 2017b; Ge et al. 2018; Zhang, Xiang, and Gong 2016; Chen et al. 2017) to obtain a compact descriptor for an efficient retrieval. This, however, gives a limited representation, as the global feature may not account for intra-class variations (e.g., human pose, occlusion, background clutter). To address this problem, part-based methods (Zhao et al. 2017a; Su et al. 2017; Zheng et al. 2017a; Li, Zhu, and Gong 2018; Liu et al. 2017; Zhao et al. 2017b; Sun et al. 2018b; Fu et al. 2019) have been proposed. They extract local features from body parts (*e.g.*, arms, legs, torso), often together with the global feature of a person image itself, and aggregate them for an effective person reID. To leverage body parts, these approaches extract pose maps from off-the-shelf pose estimators (Zhao et al. 2017a; Su et al. 2017; Zheng et al. 2017a), compute attention maps to consider discriminative regions of interest (Li, Zhu, and Gong 2018; Liu et al. 2017; Zhao et al. 2017b), or slice person images into horizontal grids (Sun et al. 2018b; Fu et al. 2019). Part-level features provide better person representations than a global one, but aggregating the individual local features *e.g.*, by concatenating them without considering relations between body parts, is limited to represent an identity of a person discriminatively. In particular, this does not differentiate the identities of different persons that have similar attributes in corresponding parts between images, since part-based methods compute the similarity of corresponding part-level features independently.

In this paper, we propose to make each part-level feature incorporate information of other body parts to obtain discriminative person representations for an effective person reID. To this end, we introduce a new relation module exploiting one-vs.-rest relations of body parts. It accounts for the relations between individual body parts and the rest of them, so that each part-level feature contains information of the corresponding part itself and other body parts, supporting it to be more discriminative. As will be seen in our experiments, considering the relation between body parts provides better part-level features with a clear ad-

---

vantage over current part-based methods. We have observed that 1) directly using both global average and max pooling techniques (GAP and GMP) to obtain a global feature of a person image does not provide a performance gain, and 2) GMP gives better results than GAP. Based on this, we also present a global contrastive pooling (GCP) method to obtain better feature representations based on GMP, which adaptively aggregates GAP and GMP results of the entire part-level features. Specifically, it uses the discrepancy between the pooling results, and distill the complementary information to max pooled features in a residual manner. Experimental results on standard benchmarks, including the Market1501 (Zheng et al. 2015), DukeMTMC-reID (Ristani et al. 2016), and CUHK03 (Li et al. 2014), demonstrate the advantage of our approach for person reID. To encourage comparison and future work, our code and models are available online: https://cvlab-yonsei.github.io/projects/RRID/.

The main contributions of this paper can be summarized as follows: 1) We introduce a relation network for part-based person reID to obtain discriminative local features. 2) We propose a new pooling method exploiting contrastive features, GCP, to extract a global feature of a person image. 3) We achieve a new state of the art, outperforming other part-based reID methods by a large margin.

Related Work

Person reID. Several person reID methods based on CNNs have recently been proposed. They typically formulate the reID task as a multi-class classification problem (Zheng, Zheng, and Yang 2018), where person images of the same identity belong to the same category. A classification loss encourages the images of the same identity to be embedded nearby in feature space. Other reID methods additionally use person images of different identities for training, and enforces the feature distance of person images with the same identity to be smaller than that with different identities by a ranking loss. There are many attempts to obtain discriminative feature representations, *e.g.*, leveraging generative adversarial networks (GANs) to distill identity-related features (Ge et al. 2018), using attributes to offer complementary information (Lin et al. 2017b), or exploiting body parts to extract diverse person features (Zhao et al. 2017a; Zheng et al. 2017a; Su et al. 2017; Liu et al. 2017; Li, Zhu, and Gong 2018; Yao et al. 2019; Zhao et al. 2017b; Sun et al. 2018b; Fu et al. 2019; Wang et al. 2018).

Part-based methods enhance the discriminative capabilities of various body parts. We classify them into three categories: The first approach uses a pose estimator (or a landmark detector) to extract a pose map (Zhao et al. 2017a; Zheng et al. 2017a; Su et al. 2017). This requires an additional data with landmark annotations to train a pose estimator, and the retrieval accuracy of reID largely depends on the performance of the estimator. The second approach leverages body parts implicitly using an attention map (Li, Zhu, and Gong 2018; Liu et al. 2017; Zhao et al. 2017b; Yao et al. 2019), which can be achieved without auxiliary supervisory signals (*i.e.*, pose annotations). It provides a feature representation robust to background clutter, focusing on the regions of interest, but the attended regions may

not contain discriminative body parts. The third approach also exploits body parts implicitly dividing person images into horizontal grids of multiple scales (Sun et al. 2018b; Fu et al. 2019; Wang et al. 2018). It assumes that person pictures, localized by off-the-shelf object detectors (Felzenszwalb et al. 2008), generally have the same body parts for particular grids (*e.g.*, legs on the lower parts of person images). This is, however, problematic when the detectors do not localize the persons tightly. Our method belongs to the third category. In contrast to other methods, we aggregate local features while considering relations between body parts, rather than exploiting them directly. Furthermore, we introduce a GCP method to obtain a global feature of a person image, providing discriminative person representations.

Relation network. Exploiting relational reasoning (Santoro et al. 2017; Baradel et al. 2018; Sun et al. 2018a; Sung et al. 2018) is important for many tasks requiring the capacity to reason about dependencies between different entities (*e.g.*, objects, actors, scene elements). Many works have been proposed to support relation-centric computation including interaction networks (Battaglia et al. 2016) and gated graph sequence networks (Li et al. 2016). The relation network (Santoro et al. 2017) is a representative method that has been successfully adapted to computer vision problems, including visual question answering (Santoro et al. 2017), object detection (Baradel et al. 2018), action recognition (Sun et al. 2018a), and few-shot learning (Sung et al. 2018). The basic idea behind the relation network is to consider all pairs of entities and to integrate all these relations *e.g.*, to answer the question (Santoro et al. 2017) or to localize the objects of interest (Baradel et al. 2018). Motivated by this work, we leverage relations of body parts to obtain better person representations for part-based person reID. Differently, we exploit the relations between individual body parts and the rest of them, rather than considering all pairs of the parts. This encourages each part-level feature to incorporate information of other body parts as well, making it more discriminative, while retaining compact feature representations for person reID.

Our Approach

We show in Fig. 1 an overview of our framework. We extract a feature map of size $H \times W \times C$ from a person image, where H, W, C are height, width, and the number of channels, respectively. The resulting feature map is divided equally into six horizontal grids. We then apply GMP to each feature map, and obtain part-level features of size $1 \times 1 \times C$. We feed these features through two modules in order to extract novel local and global person representations: One-vs.-rest relation module and GCP. The first module makes each part-level feature more discriminative by considering the relations between individual body parts and the rest of them, and outputs local relational features of size $1 \times 1 \times c$ where $c < C$. The second module provides a global contrastive feature of size $1 \times 1 \times c$ representing the person image itself. We concatenate global contrastive and local relational features along the channel dimension, and use the feature of
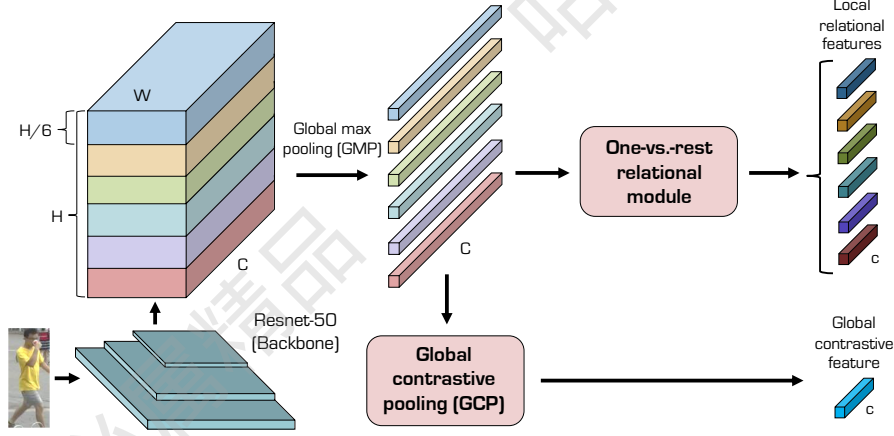
Figure 1: Overview of our framework. The proposed reID model mainly consists of three parts: We first extract part-level features by applying GMP to individual horizontal slices of the feature map from the backbone network. We then input the local features into separate modules, a one-vs.-rest relation module and GCP, that give local relational and global contrastive features, respectively.

size $1 \times 1 \times 7c$ as a person representation for reID. We train our model end-to-end using cross-entropy and triplet losses, with triplets of anchor, positive and negative person images, where the anchor image has the same identity as a positive one while having a different identity from a negative one. At test time, we extract features of person images, and compute the Euclidean distance between them to determine the identities of persons.

Relation networks for part-based reID

Part-level features. We exploit a ResNet-50 (He et al. 2016) trained for ImageNet classification (Deng et al. 2009) as our backbone network to extract an initial feature map from an input person image. Specifically, following the work of (Sun et al. 2018b), we remove the GAP and fully connected layers from the ResNet-50 architecture, and set stride of the last convolutional layer to 1. Similar to other part-based reID methods (Sun et al. 2018b; Fu et al. 2019), we split the initial feature map into multiple horizontal grids of size $H/6 \times W \times C$. We apply GMP to each of them, and obtain part-level features of size $1 \times 1 \times C$.

One-vs.-rest relational module. Extracting part-level features from the horizontal grids allows to leverage body parts implicitly for diverse person representations. Existing reID methods (Sun et al. 2018b; Fu et al. 2019; Wang et al. 2018) use these local features independently for person retrieval. They concatenate all local features in a particular order, considering rough geometric correspondences between person images. Although this gives a structural person representation robust to geometric variations and occlusion, the local features cover small parts of an image only, and more importantly they do not account for the relations between body parts. That is, individual parts are isolated, and do not communicate with other ones, which distracts computing the similarity between different persons with similar

attributes in corresponding parts. To alleviate this problem, we propose to leverage the relations between body parts for person representations. Specifically, we introduce a new relation network (Fig. 2(a)) that exploits one-vs.-rest relation of body parts, making it possible for each part-level feature to contain information of the corresponding part itself and other body parts.

Concretely, we denote by $\mathbf{p}_i$ $(i = 1, \ldots, 6)$ each part-level feature of size $1 \times 1 \times C$. We apply an average pooling to all part-level features, except the one of the particular part $\mathbf{p}_i$, aggregating the information from other body parts as follows: $\mathbf{r}_i = \frac{1}{5} \sum_{j \neq i} \mathbf{p}_j$. We then add a 1×1 convolutional layer separately for each $\mathbf{p}_i$ and $\mathbf{r}_i$, giving feature maps $\bar{\mathbf{p}}_i$ and $\bar{\mathbf{r}}_i$ of size $1 \times 1 \times c$, respectively. The relation network concatenates the features $\bar{\mathbf{p}}_i$ and $\bar{\mathbf{r}}_i$, and outputs a local relational feature $\mathbf{q}_i$ for each $\mathbf{p}_i$. We depict in Fig. 2(a) an example of extracting the local relational feature $\mathbf{q}_1$. Here, we assume that the feature $\mathbf{q}_i$ contains information of the original one $\bar{\mathbf{p}}_i$ itself and other body parts. We thus use a skip-connection (He et al. 2016) to transfer the relational information of $\bar{\mathbf{p}}_i$ and $\bar{\mathbf{r}}_i$ to $\bar{\mathbf{p}}_i$, as follows:

$$\mathbf{q}_i = \bar{\mathbf{p}}_i + R_{\mathrm{p}}(T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i)), \quad (i = 1, \ldots, 6), \quad (1)$$

where R_{p} is a sub-network consisting of a 1×1 convolution, batch normalization (Ioffe and Szegedy 2015), and ReLU (Krizhevsky, Sutskever, and Hinton 2012) layers. We denote by T a concatenation of features. The residual $R_{\mathrm{p}}(T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i))$ supports the part-level feature $\bar{\mathbf{p}}_i$, making it more discriminative and robust to occlusion. We may leverage all pairwise relations between the features $\bar{\mathbf{p}}_i$ similar to (Battaglia et al. 2016), but this requires a large computational cost and increases the dimension of features drastically. In contrast, our one-vs.-rest relation module computes the feature $\mathbf{q}_i$ in linear time, and also retains a compact feature representation.
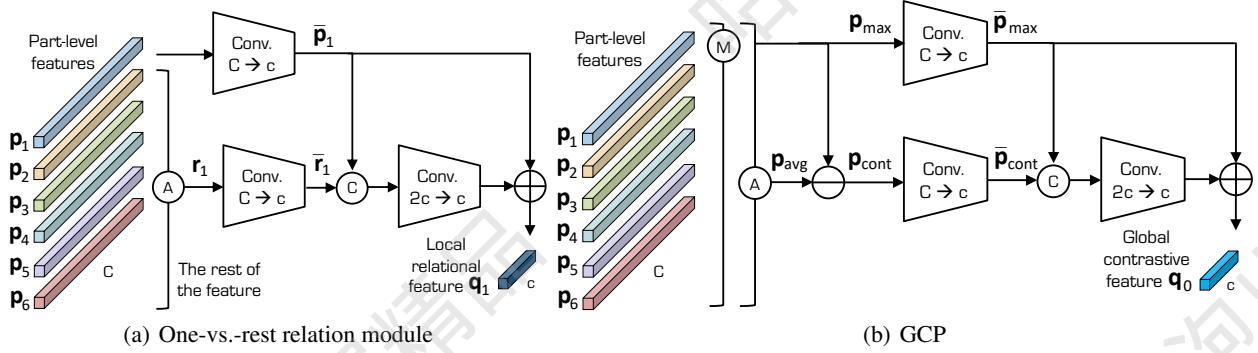
Figure 2: Illustration of (a) a one-vs.rest relational module and (b) GCP. The relation module gives a local relational feature $\mathbf{q}_i$ for each $\mathbf{p}_i$. Here, we show a process of extracting the feature $\mathbf{q}_1$. Other relational local features are similarly computed. The GCP outputs a global contrastive feature $\mathbf{q}_0$ considering all part-level features simultaneously. We do not share weight parameters of convolutional layers for all part-level features. See text for details.
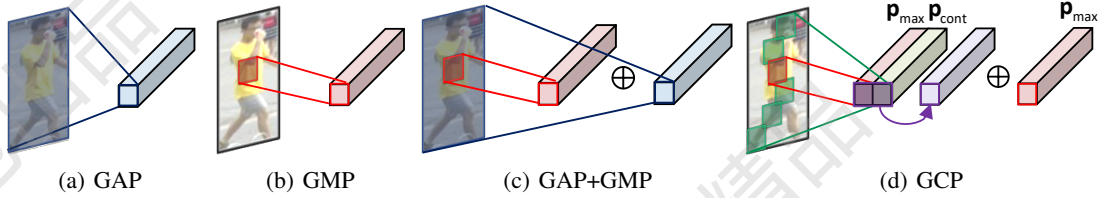

Figure 3: Illustration of various pooling methods: (a) GAP; (b) GMP; (c) GAP + GMP; (d) GCP.

GCP. To represent an entire person image, previous reID methods use GAP (Sun et al. 2018b), GMP (Wang et al. 2018), or both (Fu et al. 2019). GAP covers the whole body parts of the person image (Fig. 3(a)), but it is easily distracted by background clutter and occlusion. GMP overcomes this problem by aggregating the feature from the most discriminative part useful for reID while discarding background clutter (Fig. 3(b)). This, however, does not contain information from the whole body parts. A hybrid approach to exploiting both GAP and GMP (Fu et al. 2019) may perform better, but it is also influenced by background clutter (Fig. 3(c)). It has been proven that GMP is more effective than GAP (Fu et al. 2019), which will be also verified once more in our experiment. Motivated by this, we propose a novel GCP method based on GMP to extract a global feature map from the whole body parts (Fig. 2(b)). Rather than applying GAP or GMP to the initial feature map from the input person image (Sun et al. 2018b; Fu et al. 2019), we first perform average and max pooling with all part-level features. We denote by $\mathbf{p}_{avg}$ and $\mathbf{p}_{max}$ resulting feature maps obtained by average and max pooling, respectively. Note that $\mathbf{p}_{avg}$ and $\mathbf{p}_{max}$ are robust to background clutter, as we use a GMP method to obtain the initial part-level features (Fig. 1). That is, we aggregate the features from the most discriminative parts for every horizontal regions. In particular, $\mathbf{p}_{max}$ corresponds to the result of GMP with respect to the initial feature map from the backbone network. We then compute a contrastive feature $\mathbf{p}_{cont}$ by subtracting $\mathbf{p}_{max}$ from $\mathbf{p}_{avg}$, namely, the discrepancy between them. It aggregates most discriminative information from in-

dividual body parts (*e.g.*, green boxes in Fig. 3(d)) except the one for $\mathbf{p}_{max}$ (*e.g.*, the red box in Fig. 3(d)). We add bottleneck layers to reduce the number of channels of $\mathbf{p}_{cont}$ and $\mathbf{p}_{max}$ from C to c, denoted by $\bar{\mathbf{p}}_{cont}$ and $\bar{\mathbf{p}}_{max}$, respectively, and finally transfer the complementary information of the contrastive feature $\bar{\mathbf{p}}_{cont}$ to $\bar{\mathbf{p}}_{max}$ (Fig. 3(d)). Formally, we obtain a global contrastive feature $\mathbf{q}_0$ of the input image as follows:

$$\mathbf{q}_0 = \bar{\mathbf{p}}_{max} + R_g(T(\bar{\mathbf{p}}_{max}, \bar{\mathbf{p}}_{cont})), \qquad (2)$$

where R_g is a sub-network that consists of a 1×1 convolutional, batch normalization (Ioffe and Szegedy 2015), and ReLU (Krizhevsky, Sutskever, and Hinton 2012) layers. The global feature $\mathbf{q}_0$ is based on $\bar{\mathbf{p}}_{max}$, and aggregates the complementary information from the contrastive feature $\bar{\mathbf{p}}_{cont}$ with reference to $\bar{\mathbf{p}}_{max}$. It thus inherits the advantages of GMP such as the robustness to background clutter while covering the whole body parts. We concatenate the global contrastive feature $\mathbf{q}_0$ in (2) and local relational ones $\mathbf{q}_i$ ($i = 1, \ldots, 6$) in (1), and use it as a person representation for reID.

Training loss. We exploit ground-truth identification labels of person images to learn the person representation. To train our model, we use cross-entropy and triplet losses, balanced by the parameter λ as follows:

$$\mathcal{L} = \mathcal{L}_{triplet} + \lambda \mathcal{L}_{ce}, \qquad (3)$$

where we denote by $\mathcal{L}_{triplet}$ and $\mathcal{L}_{ce}$ triplet and cross-entropy losses, respectively. The cross-entropy loss is de-

| Methods | F-dim. | Market1501 | | CUHK03 | | | | DukeMTMC-reID | |
| | | | | Labeled | | Detected | | | |
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
SVDNet (Sun et al. 2017)	2,048	62.1	82.3	37.8	40.9	37.3	41.5	56.8	76.7
Triplet (Hermans, Beyer, and Leibe 2017)	128	69.1	84.9	-	-	-	-	-	-
HA-CNN (Li, Zhu, and Gong 2018)	1,024	75.7	91.2	41.0	44.4	38.6	41.7	63.8	80.5
Deep-Person (Bai et al. 2017)	2,048	79.5	92.3	-	-	-	-	64.8	80.9
AlignedReID (Luo et al. 2019)	2,048	79.1	91.8	-	-	59.6	61.5	69.7	82.1
PCB (Sun et al. 2018b)	1,536	77.3	92.4	-	-	54.2	61.3	65.3	81.9
PCB+RPP (Sun et al. 2018b)	1,536	81.0	93.1	-	-	57.5	63.7	68.5	82.9
HPM (Fu et al. 2019)	3,840	82.7	94.2	-	-	57.5	63.9	74.3	86.6
MGN (Wang et al. 2018)	2,048	86.9	**95.7**	67.4	68.0	66.0	66.8	<u>78.4</u>	88.7
Ours-S	1,792	<u>88.0</u>	94.8	<u>73.5</u>	<u>76.6</u>	<u>69.5</u>	<u>72.5</u>	77.1	<u>89.3</u>
Ours-F	3,840	**88.9**	<u>95.2</u>	**75.6**	**77.9**	**69.6**	**74.4**	**78.6**	**89.7**

Table 1: Quantitative comparison with the state of the art in person reID. We measure mAP(%) and rank-1 accuracy(%) on the Market1501 (Zheng et al. 2015), CUHK03 (Li et al. 2014) and DukeMTMC-reID (Ristani et al. 2016) datasets. We denote suffixes "-S" and "-F" by our models using $\mathbf{q}^{\mathrm{P6}}$ and $T(\mathbf{q}^{\mathrm{P2}}, \mathbf{q}^{\mathrm{P4}}, \mathbf{q}^{\mathrm{P6}})$, respectively.

fines as

$$\mathcal{L}_{\mathrm{ce}} = -\sum_{n=1}^{N} \sum_{i} y^n \log \hat{y}_i^n, \tag{4}$$

where we denote by N and y^n the number of images in mini-batch and a ground-truth identification label, respectively. $\hat{y}_i^n$ is a predicted identification label for each feature $\mathbf{q}_i$ in the person representation, defined as

$$\hat{y}_i^n = \underset{c \in K}{\mathrm{argmax}} \frac{\exp((\mathbf{w}_i^c)^{\mathrm{T}} \mathbf{q}_i)}{\sum_{k=1}^{K} \exp((\mathbf{w}_i^k)^{\mathrm{T}} \mathbf{q}_i)}. \tag{5}$$

K is the number of identification labels, and $\mathbf{w}_i^k$ is classifier for the feature $\mathbf{q}_i$ and the label k. We use a fully connected layer for the classifier. To enhance the ranking performance, we use the batch-hard triplet loss (Hermans, Beyer, and Leibe 2017), formulated as follows:

$$\mathcal{L}_{\mathrm{triplet}} = \sum_{k=1}^{N_K} \sum_{m=1}^{N_M} [\alpha + \max_{n=1\ldots M} \|\mathbf{q}_{k,m}^{\mathrm{A}} - \mathbf{q}_{k,n}^{\mathrm{P}}\|_2 \\ - \min_{\substack{l=1\ldots K \\ n=1\ldots N \\ l \neq k}} \|\mathbf{q}_{k,m}^{\mathrm{A}} - \mathbf{q}_{l,n}^{\mathrm{N}}\|_2]_+, \tag{6}$$

where N_K is the number of identities in mini-batch, and N_M is the number of images for each identification label in mini-batch ($N = N_K N_M$). α is a margin parameter to control the distances between positive and negative pairs in feature space. we denote by $\mathbf{q}_{i,j}^{\mathrm{A}}, \mathbf{q}_{i,j}^{\mathrm{P}}, \mathbf{q}_{i,j}^{\mathrm{N}}$ person representations of anchor, positive, and negative images, respectively, where i, j correspond to identification and image indexes.

Extension to different numbers of grids. We so far describe our model using global and local features, i.e., $T(\mathbf{q}_0 \ldots \mathbf{q}_6)$, for a person representation, which is denoted by $\mathbf{q}^{\mathrm{P6}}$ hereafter. Without loss of generality, we can use different numbers of horizontal grids for the person representation, to consider various parts of multiple scales (Fu et al. 2019), such as $\mathbf{q}^{\mathrm{P2}}$ and $\mathbf{q}^{\mathrm{P4}}$ that splits the initial feature

map into two and four horizontal regions, respectively. Accordingly, we concatenate the features of $\mathbf{q}^{\mathrm{P2}}, \mathbf{q}^{\mathrm{P4}}$ and $\mathbf{q}^{\mathrm{P6}}$, i.e., $T(\mathbf{q}^{\mathrm{P2}}, \mathbf{q}^{\mathrm{P4}}, \mathbf{q}^{\mathrm{P6}})$, and use it as a final person representation for an effective reID. Note that $\mathbf{q}^{\mathrm{P2}}, \mathbf{q}^{\mathrm{P4}}$, and $\mathbf{q}^{\mathrm{P6}}$ contain different local relational features, and thus have different global contrastive features. Note also that these features share the same backbone network with the same parameters.

Experimental Results

Implementation details

Dataset. We test our method on the following datasets and compare its performance with the state of the art. 1) The Market1501 dataset (Zheng et al. 2015) contains 32,668 person images of 1,501 identities captured by six cameras. We use the training/test split provided by (Zheng et al. 2015) where it consists of 12,936 images of 751 identities for training and 3,368 query and 19,732 gallery images of 750 identities for testing. 2) The CUHK03 dataset (Li et al. 2014) provides 14,097 images of 1,467 identities observed by two cameras, and it offers two types of person images: Manually labeled and detected ones by the DPM method (Felzenszwalb et al. 2008). Following the training/test split of (Zhong et al. 2017a), we divide it into 7,365 images of 767 identities for training, and 1,400 query and 5,332 gallery images of 700 identities. 3) The DukeMTMC-reID (Ristani et al. 2016) offers 16,522 training images of 702 identities, 2,228 query and 17,661 gallery images of 702 identities.

Training. We resize all images into 384×128 for training. We set the numbers of feature channels C to 2,048 and c to 256. This results in 1,792- and 3,840-dimensional features for $\mathbf{q}^{\mathrm{P6}}$ and $T(\mathbf{q}^{\mathrm{P2}}, \mathbf{q}^{\mathrm{P4}}, \mathbf{q}^{\mathrm{P6}})$, respectively. We augment the training datasets with horizontal flipping and random erasing (Zhong et al. 2017b). We use the stochastic gradient descent (SGD) as the optimizer with momentum of 0.9 and weight decay of 5e-4. We train our model with a batch size N of 64 for 80 epochs, where we randomly
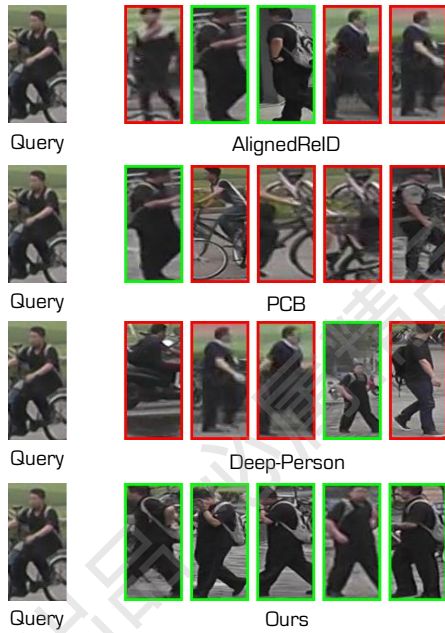
Figure 4: Qualitative comparison of person reID results on the Market1501 dataset (Zheng et al. 2015). We show the top-5 retrieval results (left: rank-1, right: rank-5) for AlignedReID (Luo et al. 2019), Deep-Person (Bai et al. 2017), PCB (Sun et al. 2018b) and Ours-F. Retrieved images with green and red boxes are correct and incorrect results, respectively.

choose 16 identities and sample 4 person images for each identity ($N_K = 16$, $N_M = 4$). A learning rate initially set to 1e-3 and 1e-2 for the backbone network and other parts, respectively, until 40 epochs is divided by 10 every 20 epochs. We empirically set the weight parameter λ to 2, and fix it to all experiments. All networks are trained end-to-end using `PyTorch` (Paszke et al. 2017). Training our model takes about six, three and eight hours with two NVIDIA Titan Xps for the Market1501, CUHK03, and DukeMTMC-reID datasets, respectively.

Comparison with the state of the art

Quantitative results. We compare in Table our models with the state of the arts including part-based person reID methods. We measure mean average precision (mAP) (%) and rank-1 accuracy (%) on the Market1501 (Zheng et al. 2015), CUHK03 (Li et al. 2014), and DukeMTMC-reID (Ristani et al. 2016) datasets. We report reID results for a single query for a fair comparison. We denote suffixes "-S" and "-F" by our models using $\mathbf{q}^{P6}$ and $T(\mathbf{q}^{P2}, \mathbf{q}^{P4}, \mathbf{q}^{P6})$, respectively, for final person representations. Table shows that Ours-S outperforms state-of-the-art reID methods in terms of mAP and rank-1 accuracy for all datasets, except MGN (Wang et al. 2018). This demonstrates the effectiveness of our one-vs.-rest relation module and GCP. Moreover, Ours-S uses 1,792-dimensional features, allowing to an efficient person retrieval, while providing state-of-the-art re-

sults. Ours-F gives the best results again on all datasets in terms of mAP. We achieve mAP of 88.9% and rank-1 accuracy of 95.2% for the Market1501, mAP of 69.6%/75.6% and rank-1 accuracy of 74.4%/77.9% with detected/labeled images for the CUHK03, and mAP of 78.6% and rank-1 accuracy of 89.7% for the DukeMTMC-reID. The rank-1 accuracy of Ours-F is slightly lower than that of MGN (Wang et al. 2018) on the Market1501, but Ours-F outperforms MGN on other datasets by a significant margin. Note that person images in the CUHK03 and DukeMTMC-reID datasets are much more difficult to retrieve, as they are typically with large pose variations, background clutter, occlusion, and confusing attributes.

Qualitative results. Figure 4 shows a visual comparison of person retrieval results with the state of the art (Luo et al. 2019; Bai et al. 2017; Sun et al. 2018b) on the Market1501 dataset (Zheng et al. 2015). We show the top-5 retrieval results for the query image. The results for all comparisons have been obtained from the official models provided by the authors. We can see that our method retrieves correct person images, and in particular it is robust to attribute variations (*e.g.*, bicycles) and background clutter (*e.g.*, grass). Other part-based methods including AlignedReID and PCB try to match local features between images for retrieval. For example, they focus on finding the correspondences for bicycles or grass in the query image, giving many false-positive results. Note that the gallery set contain many images of persons riding a bicycle, but they have different identities from the person in the query.

Discussion

Ablation study. We show an ablation analysis on different components of our model in Table . We compare the performance for several variants of our model in terms of mAP and rank-1 accuracy. From the first three rows, we can see that using both global and local features improves the retrieval performance, which confirms the finding in part-based reID methods. The fourth row shows that GMP gives better results than GAP, as GMP avoids background clutter. The results in the next row demonstrates the effect of local features obtained using our relation module. For example, this gives the performance gains of 3% and 1% for mAP and rank-1 accuracy, respectively, for the Market1501 dataset, which is quite significant. From the fifth and eighth rows, we show a comparison of the retrieval performance according to pooling methods for the global feature, and we can see that GCP performs better than GMP, GAP and GMP+GAP in terms of mAP and rank-1 accuracy. For example, compared with GMP+GAP, our GCP improves mAP from 86.6% to 88.0% on the Market1501 dataset. The last four rows suggest that exploiting part-level features of multiple scales is important, and using all components performs best.

We show in Fig. 5(a) retrieval results with/with a relational module. We can see that person representations obtained from the relation module successfully discriminate the same attribute (*e.g.*, violet shirts) for the person images of different identities (*e.g.*, gender), and they are robust to

| GF | LF | RM | Ext. | Pooling for | | F-dim. | Market1501 | | CUHK03 | | | | DukeMTMC-reID | |
| | | | | | | | | | Labeled | | Detected | | | |
				GF	LF		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
✓				GAP	-	256	74.5	88.2	57.0	60.6	54.3	58.4	62.9	79.6
	✓			-	GAP	1,536	79.0	92.3	65.1	67.9	62.4	65.6	70.0	84.0
✓	✓			GAP	GAP	1,792	82.9	92.9	68.1	71.4	63.6	66.2	73.5	85.5
✓	✓			GMP	GMP	1,792	83.7	93.2	70.7	73.9	64.3	66.8	74.8	86.1
✓	✓	✓		GMP	GMP	1,792	86.7	94.2	73.3	75.8	67.6	70.2	76.3	88.3
✓	✓	✓		GAP	GMP	1,792	85.8	94.1	72.6	75.0	67.6	69.6	75.6	88.2
✓	✓	✓		GAP+GMP	GMP	1,792	86.6	94.3	72.9	75.8	68.1	70.3	76.5	88.4
✓	✓	✓		GCP	GMP	1,792	_88.0_	_94.8_	_73.5_	_76.6_	_69.5_	_72.5_	77.1	_89.3_
✓	✓		✓	GMP	GMP	3,840	86.7	94.4	72.8	74.6	67.7	69.1	76.1	87.7
✓	✓		✓	GAP	GMP	3,840	86.5	94.2	72.5	74.7	67.3	69.9	76.5	87.3
✓	✓		✓	GAP+GMP	GMP	3,840	86.5	94.1	72.8	75.4	66.5	69.9	76.6	87.8
✓	✓		✓	GCP	GMP	3,840	87.3	94.5	73.0	75.9	69.0	71.6	_77.4_	88.3
✓	✓	✓	✓	GCP	GMP	3,840	**88.9**	**95.2**	**75.6**	**77.9**	**69.6**	**74.4**	**78.6**	**89.7**

Table 2: Quantitative comparison of different network architectures. We measure mAP(%) and rank-1 accuracy(%) on the Market1501 (Zheng et al. 2015), CUHK03 (Li et al. 2014) and DukeMTMC-reID (Ristani et al. 2016) datasets. Numbers in bold indicate the best performance and underscored ones are the second best. GF: Global features; LF: Local features; RM: One-vs.-rest relational module; Ext.: An extension to multiple scales.

Figure 5: Visual comparison of retrieval results: (a) a relational module and (b) pooling methods. We show top-1 results. The relation module discriminates the same attribute for the person images of different identities. GCP allows to aggregate features from discriminative regions, and provides a person representation robust to background clutter, overcoming the drawbacks of GAP and GMP.

Figure 6: Examples of activation maps for the models using $R_p(\bar{\mathbf{r}}_3)$ (top) and $R_p(T(\bar{\mathbf{p}}_3, \bar{\mathbf{r}}_3))$ (bottom). We also show top-3 retrieval results for each model. We can see that the rest regions (*e.g.*, the fifth and sixth horizontal grids) are highly activated by the person representation using $R_p(T(\bar{\mathbf{p}}_3, \bar{\mathbf{r}}_3))$, compared to that using $R_p(\bar{\mathbf{r}}_3)$, indicating that our one-vs.-rest relation allows each part-level feature to see the rest regions effectively.

occlusion (*e.g.*, the person occluded by a bag or a bicycle). We compare in Fig. 5(b) retrieval results for different pooling methods. We confirm once again that GAP is not robust to background clutter and GMP sees the most discriminative region (*e.g.*, bicycle) only rather than the person. We observe that GCP alleviates these problems while maintaining the advantage of GMP.

One-vs.-rest relation. We consider relations between each part-level feature $\bar{\mathbf{p}}_i$ and its rest feature $\bar{\mathbf{r}}_i$. To show how the relation module works, we train a model using the combined rest feature $\bar{\mathbf{r}}_i$ instead of $T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i)$ in (1). In this case, we do not use a multi-scale extension. We visualize activation maps of $R_p(\bar{\mathbf{r}}_i)$ and $R_p(T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i))$ in Fig. 6. We observe that $R_p(T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i))$ focuses more on the regions whose attributes are different from those of $\bar{\mathbf{p}}_i$, compared to $R_p(\bar{\mathbf{r}}_i)$.
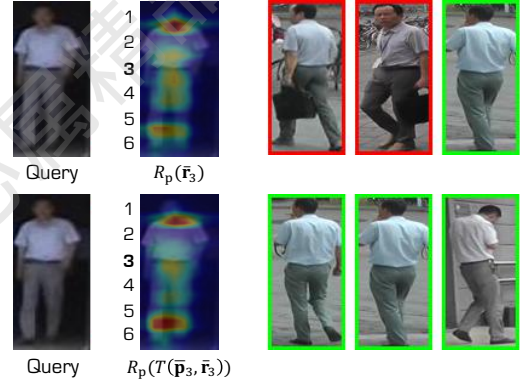
This demonstrates that $R_p(T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i))$ extracts the complementary features of $\bar{\mathbf{p}}_i$, that are helpful for person reID but not contained in $\bar{\mathbf{p}}_i$, from the rest of the parts. This also verifies that using the feature of $\bar{\mathbf{p}}_i$ only is not enough to discriminate the identities of different persons having similar attributes in corresponding parts between images. The mAP/rank-1 for $\bar{\mathbf{r}}_i$ on Market1501 are 84.5/93.4 which are lower than 86.7/94.2 obtained using $T(\bar{\mathbf{p}}_i, \bar{\mathbf{r}}_i)$ in Table .

Performance comparison of local features. We demonstrate the capabilities of providing discriminative features of

Horizontal region	w/o RM		RM	
	mAP	rank-1	mAP	rank-1
1	29.6	49.4	58.7	82.2
2	46.4	69.2	68.5	87.1
3	**55.3**	**77.2**	**73.5**	**89.1**
4	51.8	72.7	71.9	87.6
5	49.7	69.2	69.7	85.4
6	35.5	53.4	55.7	76.9

Table 3: Quantitative comparison of single local features on the Market1501 dataset (Zheng et al. 2015). RM: One-vs.-rest relational module.

the one-vs.-rest relation module. We extract a local feature of size $1 \times 1 \times 256$ for each horizontal region with and without using the relation module. Given a query image, we then retrieve person images using a single local feature in the person representation of $\mathbf{q}^{P6}$. We report mAP and rank-1 accuracy for individual local features extracted from different horizontal regions (1: top, 6: bottom), on the Market1501 dataset (Zheng et al. 2015) in Table . From this table, we can observe two things: (1) The relation module improves mAP and rank-1 accuracy drastically. The improvements in mAP and rank-1 accuracy are 21.7% and 19.6%, respectively, on average. The rank-1 accuracy measures the performance of retrieval results for the easiest match, while mAP characterizes the ability to retrieve all person images of the same identity, indicating that the relation module is beneficial especially for retrieving challenging person images. (2) The local features from the third and last horizontal regions give the best and the worst results, respectively. This suggests that the middle part typically corresponding to the torso of a person provides the most discriminative feature for specifying a particular person, and the bottom part (*e.g.*, legs or sometimes background due to the incorrect localization of the person detector) gives the least discriminative feature for person reID.

Performance comparison of global features. Table compares the reID performance of single global features in terms of mAP and rank-1 accuracy on the Market1501 dataset (Zheng et al. 2015). We use only the 256-dimensional global feature in the person representation of $\mathbf{q}^{P6}$ for person retrieval but obtained by different pooling methods. Note that the size of the global feature is much smaller than typical person representations in Table . We can see from Table that GCP gives the best retrieval results in terms of both mAP and rank-1 accuracy, outperforming GAP, GMP, and GAP+GMP by a large margin. Compared with other reID methods in Table , GCP offers a good compromise in terms of the accuracy and the size of features. For example, our global contrastive feature of size $1 \times 1 \times 256$ achieves rank-1 accuracy of 93.4%, which is comparable with 93.1% for PCB+RPP (Sun et al. 2018b) using 1,536-dimensional features.

Methods	mAP	rank-1
GAP	81.0	91.3
GMP	81.9	92.0
GAP+GMP	82.5	92.6
GCP	**84.6**	**93.4**

Table 4: Quantitative comparison of global features on the Market1501 dataset (Zheng et al. 2015).

Conclusion

We have presented a relation network for person reID considering the relations between individual body parts and the rest of them, making each part-level feature more discriminative. We have also proposed to use contrastive features for a global person representation. We set a new state of the art on person reID, outperforming other reID methods by a significant margin. The ablation analysis clearly demonstrates the effectiveness of each component in our model.

References

Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; and Xu, Y. 2017. Deep-Person: Learning discriminative deep features for person re-identification. *arXiv preprint arXiv:1711.10658*.

Baradel, F.; Neverova, N.; Wolf, C.; Mille, J.; and Mori, G. 2018. Object level visual reasoning in videos. In *ECCV*.

Battaglia, P.; Pascanu, R.; Lai, M.; Rezende, D. J.; et al. 2016. Interaction networks for learning about objects, relations and physics. In *NIPS*.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Felzenszwalb, P. F.; McAllester, D. A.; Ramanan, D.; et al. 2008. A discriminatively trained, multiscale, deformable part model. In *CVPR*.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2019. Horizontal pyramid matching for person re-identification. In *AAAI*.

Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; et al. 2018. FD-GAN: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

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

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*.

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

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2016. Gated graph sequence neural networks. In *ICLR*.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *CVPR*.

Lin, J.; Ren, L.; Lu, J.; Feng, J.; and Zhou, J. 2017a. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; and Yang, Y. 2017b. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*.

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017. HydraPlus-Net: Attentive deep features for pedestrian analysis. In *ICCV*.

Luo, H.; Jiang, W.; Zhang, X.; Fan, X.; Qian, J.; and Zhang, C. 2019. AlignedReID++: Dynamically matching local information for person re-identification. *PR*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; De-Vito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NIPS*.

Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Sun, Y.; Zheng, L.; Deng, W.; and Wang, S. 2017. SVDNet for pedestrian retrieval. In *ICCV*.

Sun, C.; Shrivastava, A.; Vondrick, C.; Murphy, K.; Sukthankar, R.; and Schmid, C. 2018a. Actor-centric relation network. In *ECCV*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018b. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*.

Tang, S.; Andriluka, M.; Andres, B.; and Schiele, B. 2017. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*.

Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; and Tian, Q. 2019. Deep representation learning with part loss for person re-identification. *IEEE TIP*.

Zhang, L.; Xiang, T.; and Gong, S. 2016. Learning a discriminative null space for person re-identification. In *CVPR*.

Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; and Tang, X. 2017a. Spindle Net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*.

Zhao, L.; Li, X.; Zhuang, Y.; and Wang, J. 2017b. Deeply-learned part-aligned representations for person re-identification. In *ICCV*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zheng, L.; Huang, Y.; Lu, H.; and Yang, Y. 2017a. Pose invariant embedding for deep person re-identification. *IEEE TIP*.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017b. Person re-identification in the wild. In *CVPR*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2018. A discriminatively learned CNN embedding for person re-identification. In *ACM TOMM*.

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

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2017b. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*.