Weighted Bilinear Coding over Salient Body Parts for Person Re-identification

Qin Zhou[1,2], Heng Fan[2], Hang Su[3], Hua Yang[1], Shibao Zheng[1] and Haibin Ling[2]
[1]Institute of Image Processing and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Computer & Information Sciences, Temple University, Philadelphia 19122, USA
[3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract

Deep convolutional neural networks (CNNs) have demonstrated dominant performance in person re-identification (Re-ID). Existing CNN based methods utilize global average pooling (GAP) to aggregate intermediate convolutional features for Re-ID. However, this strategy only considers the first-order statistics of local features and treats local features at different locations equally important, leading to sub-optimal feature representation. To deal with these issues, we propose a novel *weighted bilinear coding* (WBC) model for local feature aggregation in CNN networks to pursue more representative and discriminative feature representations. In specific, bilinear coding is used to encode the channel-wise feature correlations to capture richer feature interactions. Meanwhile, a weighting scheme is applied on the bilinear coding to adaptively adjust the weights of local features at different locations based on their importance in recognition, further improving the discriminability of feature aggregation. To handle the spatial misalignment issue, we use a salient part net to derive salient body parts, and apply the WBC model on each part. The final representation, formed by concatenating the WBC eoncoded features of each part, is both discriminative and resistant to spatial misalignment. Experiments on three benchmarks including Market-1501, DukeMTMC-reID and CUHK03 evidence the favorable performance of our method against other state-of-the-art methods.

1 Introduction

Person re-identification (Re-ID) aims at associating a probe image with images of the same identity in the gallery set (usually across different non-overlapping camera views). It is attracting increasing attentions due to its importance for various applications including video surveillance, human-machine interaction, robotics, etc. Despite years of efforts, accurate Re-ID remains largely unsolved because of great challenges posed by illumination changes, pose variations or viewpoint changes, and other factors like background clutters and occlusions. Various techniques have been proposed to improve the recognition performance against the above-mentioned challenges.

Recently deep convolutional neural networks (CNNs) [LeCun *et al.*, 1989] have been widely utilized for Re-ID because of the power in learning discriminative and representative features. Being an end-to-end architecture, CNNs directly take as input the raw images, and hierarchically aggregate local features into a final vectorized representation for further processing. In such Re-ID solutions, one crucial problem is how to aggregate the intermediate convolutional features to build more discriminative appearance representation for better recognition performance. For the sake of efficiency and simplicity, most CNN based approaches use global average pooling to aggregate the convolutional features to represent human appearance [Zhao *et al.*, 2017b; Su *et al.*, 2017]. However, discarding the information of feature correlations as well as the various feature importance across different locations, GAP leads to suboptimal aggregated appearance representation.

To deal with this issue, in this paper, we propose a novel *weighted bilinear coding* (WBC) model for discriminative feature aggregation in CNNs, which is able to model richer higher-order feature interactions as well as the various feature impacts for Re-ID. In our WBC model, the bilinear coding takes into consideration the channel-wise correlations for each local feature. In comparison to global average pooling, bilinear coding captures richer feature information. Meanwhile, considering that the features at different locations have different impacts on the recognition performance, we further introduce a weighting scheme into bilinear coding, which adaptively weighs different features according to their relative importance in recognition. The proposed WBC model is flexible and can be embedded into arbitrary networks.

To deal with the problem of spatial misalignment in Re-ID, we integrate the proposed WBC model with a salient part net to pursue part-aligned discriminative representation for Re-ID. In specific, the salient part net is used to derive salient human body parts, then we apply the proposed WBC on each part to obtain corresponding discriminative feature representation. The final representation for each human image, formed by concatenating the features of each part, bears the properties of both discriminability and resistance to spatial misalignment. The proposed Re-ID system, which is end-to-end trainable, is illustrated in Figure 1.

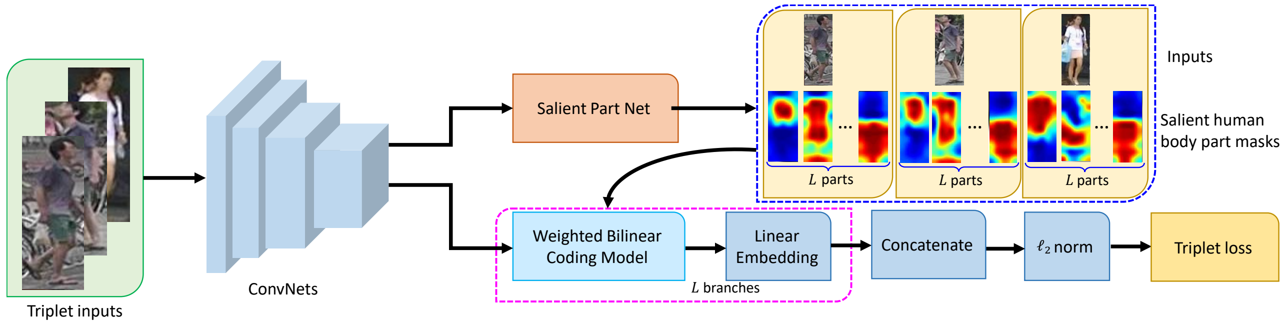


Figure 1: Illustration of the proposed person Re-ID framework. Based on the feature maps extracted from base ConvNets, we first adopt a salient part net to obtain salient human body parts, then the proposed WBC model is applied on each part for discriminative feature aggregation. The final representation of each person is formed by concatenating the features of all parts, followed by ℓ_2 normalization. Triplet loss calculated on the final representations is adopted for parameter learning of the Re-ID network.

In summary, we make the following contributions:

- We propose a novel WBC model for representative and discriminative feature aggregation, which can be flexibly plugged into existing deep architectures.
- To alleviate the spatial misalignment problem, we integrate the WBC model with a salient part generation network to pursue part-aligned discriminative features in an end-to-end trainable network for Re-ID.
- Extensive experiments on three large-scale datasets including Market-1501 [Zheng *et al.*, 2015], DukeMTMC-reID [Ristani *et al.*, 2016] and CUHK03 [Li *et al.*, 2014] demonstrate the favorable performance of our algorithm against state-of-the-art approaches.

2 Related Work

Being extensively studied, numerous approaches have been proposed for Re-ID in recent years. In this section, we briefly review some related works of this paper. For a detailed survey of Re-ID, please refer to [Zheng *et al.*, 2016].

The literature of Re-ID can be generally summarized into two categories: global based models which take the whole human body into consideration during feature design or metric learning and local based ones that extract features from local body parts and then aggregate these local features for final ranking. Earlier works mainly focus on global models for Re-ID [Zheng *et al.*, 2011; Köstinger *et al.*, 2012; Xiong *et al.*, 2014]. These methods, nevertheless, degrade in presence of spatial misalignment caused by large variations in view angles and human poses.

To alleviate the problem of spatial misalignment, many local based algorithms have been proposed. Given that the human body is usually centered in a manually cropped bounding boxes, some researchers argue that the body parts are vertically roughly aligned. Therefore a possible solution is to decompose human body into uniform stripes, and pool features extracted from these stripes into a robust representation. In [Chen *et al.*, 2016], the authors propose to learn a sub-similarity function for each stripe, and then fuse all sub-similarity scores for final recognition. The work of [Cheng *et al.*, 2016] proposes to learn deep features from both global body and local stripes to pursue better representation of human appearance. Whereas the images at hand may be not perfectly cropped (e.g., the bounding boxes are obtained by existing detection algorithms). In such a case, the fixed stripes based partition may fail. On the other hand, some other algorithms [Zhao *et al.*, 2017c; Zhou *et al.*, 2018] directly perform patch-level matching, which is more flexible than stripe-based ones, and can well address the spatial misalignment problem if patch-wise correspondences are accurately established. However, how to establish dense pair-wise correspondences still remains a challenging problem.

Recently, part based approaches are introduced into CNNs to automatically generate semantically aligned body parts to guide feature learning. In [Yao *et al.*, 2017], salient body parts are generated by performing clustering on intermediate features, and an identity classification loss is imposed on both the whole body and body parts. During testing, the generated part features are concatenated with the global feature to enhance the representative ability. Inspired by the attention model, [Zhao *et al.*, 2017b] proposes a part net composed by convolutional layers to automatically detect salient body parts, and aggregate features over these parts into a global representation. Despite promising performance in handling spatial misalignment, the usage of global average pooling for feature aggregation in these approaches [Yao *et al.*, 2017; Zhao *et al.*, 2017b] leads to sub-optimal results due to the ignorance of richer feature interactions and various feature significance. Unlike the aforementioned algorithms, we present a novel WBC model for discriminative feature aggregation, demonstrating obvious advantages compared to global average pooling.

3 The Proposed Approach

3.1 Problem formulation

In this paper, we formulate the task of Re-ID as a ranking problem, where the goal is to minimize the intra-person divergence while maximize the inter-person divergence. Specifically, given an image set $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \cdots, \mathbf{I}_N\}$ with N images, we form the training set into a set of triplets $\mathcal{T} = \{(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k)\}$, where $\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k$ are images with identity labels y_i, y_j and y_k respectively. In a triplet unit, $(\mathbf{I}_i, \mathbf{I}_j)$ is a positive image pair of the same person (i.e., $y_i = y_j$), while $(\mathbf{I}_i, \mathbf{I}_k)$ is a negative image pair (i.e., $y_i \neq y_k$). Then the purpose of Re-ID is to rank $\mathbf{I}_j$ before $\mathbf{I}_k$ for all triplets, which can be mathematically expressed as

$$d(\phi(\mathbf{I}_i), \phi(\mathbf{I}_j)) + \alpha \leq d(\phi(\mathbf{I}_i), \phi(\mathbf{I}_k)) \tag{1}$$

where $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ represents the Euclidean distance, $\phi(\cdot)$ denotes the feature transformation using deep neural networks as described later, and $\alpha > 0$ is the margin by which the distance between a negative image pair is greater than that between a positive image pair. To enforce this constraint, a common relaxation of Eq. (1) is the minimization of the triplet hinge loss as

$$\ell_{tri}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) = \Big[d(\phi(\mathbf{I}_i), \phi(\mathbf{I}_j)) - d(\phi(\mathbf{I}_i), \phi(\mathbf{I}_k)) + \alpha\Big]_+ \tag{2}$$

where the operator $[\cdot]_+ = \max(0, \cdot)$ represents the hinge loss. The whole loss function for all triplets in training set is then expressed as

$$\mathcal{L}(\phi) = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) \in \mathcal{T}} \ell_{tri}(\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k) \tag{3}$$

where $|\mathcal{T}|$ denotes the number of triplets in $\mathcal{T}$.

3.2 Salient part-based representation

For Re-ID, one of the most challenging problems is spatial misalignment caused by variations in views and human poses. To deal with this issue, we adopt a salient part-based representation for human appearance (similar to [Zhao *et al.*, 2017b]). Different from previous approaches using spatially fixed vertical or horizontal stripe-based representation, our method aims at partitioning human body into several salient regions and each salient region across humans are semantically aligned, alleviating the problem of spatial misalignment.

In this work, the salient part-based representation is derived by a sub-network (referred to as salient part net). In specific, the salient part net consists of L branches, and each branch is composed of a 1×1 convolutional layer followed by a nonlinear sigmoid layer. Figure 2 illustrates the architecture of the salient part net. The input to the salient part net is the 3-dimension intermediate convolutional feature maps, and its outputs are L 2-dimension salient part masks. Specifically, let $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ represent the input feature maps to the salient part net, then we can estimate the part masks $\mathbf{M}_l \in \mathbb{R}^{H \times W}, l \in \{1, \cdots, L\}$ as

$$\mathbf{M}_l = \Phi_{\mathrm{SalientMask}_l}(\mathbf{F}) \tag{4}$$

where $\Phi_{\mathrm{SalientMask}_l}(\cdot)$ represents the l^{th} salient part mask generator. In our end-to-end formulation, the values of the obtained salient part masks reflect the relative importance of their corresponding local features. Taking $\mathbf{M}_l$ as the automatically learned weights, we can compute the part-based feature $\mathbf{F}_l$ using the proposed WBC as

$$\mathbf{F}_l = \Psi_{\mathrm{WBC}}(\mathbf{M}_l, \mathbf{F}) \tag{5}$$

where $\Psi_{\mathrm{WBC}}(\cdot, \cdot)$ represents the proposed feature coding algorithm and will be discussed later. It is worth noting that different from existing part-aligned Re-ID method [Zhao *et al.*, 2017b] which uses global average pooling for feature aggregation, our WBC is able to fully explore richer higher-order

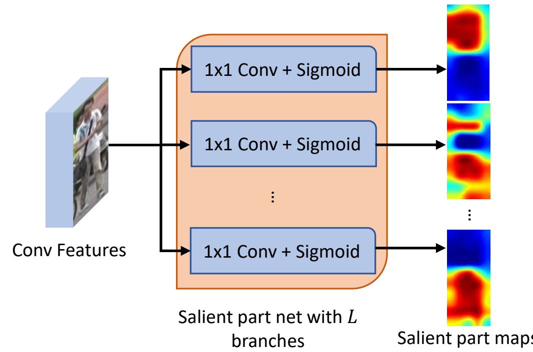


Figure 2: Illustration of the architecture of salient part net.

channel-wise feature interactions, improving the representative ability and discriminability of feature aggregation.

Afterwards, the encoded feature of each part $\mathbf{F}_l$ is passed into a linear embedding for dimension reduction. Let $\mathbf{F}'_l$ denote the dimension-reduced feature of $\mathbf{F}_l$, then the discriminative part-aligned feature representation is formed by concatenating $\mathbf{F}'_l$ for each part, followed by ℓ_2 normalization,

$$\mathbf{f} = \left\| [(\mathbf{F}'_1)^\top, (\mathbf{F}'_2)^\top, \cdots, (\mathbf{F}'_L)^\top]^\top \right\|_2 \tag{6}$$

The obtained feature representation $\mathbf{f}$ is then utilized as the feature transformation $\phi(\mathbf{I})$ in Eq. (1).

3.3 Weighted bilinear coding

Given the input feature maps $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, much identity-aware discriminative information of the input image $\mathbf{I}$ is implicitly captured. However, how to aggregate the local features of $\mathbf{F}$ to fully explore its representative and discriminative potential for Re-ID remains a problem. Most of the existing algorithms [Zhao *et al.*, 2017b; Yao *et al.*, 2017] adopt global average pooling (GAP) for the sake of efficiency and simplicity. However, GAP only captures the first-order statistics of local features and considers all the units inside the feature maps equally important. This may undermine both the representative and discriminative ability of the final representation. Bilinear coding [Lin *et al.*, 2015] is recently introduced into the CNN network to model the higher-order channel-wise feature interactions, enhancing the representative ability of the learned deep features. Originally, the bilinear coding takes all the local features as input and outputs a representation $\mathbf{B}$ as follows:

$$\mathbf{B} = \sum_{p=1}^{H} \sum_{q=1}^{W} \mathbf{F}(p, q)^T \mathbf{F}(p, q) \tag{7}$$

where $\mathbf{F}(p, q) \in \mathbb{R}^{1 \times C}$ is the local feature at the (p, q)-th location. Nevertheless, it is suboptimal for Re-ID without considering the various impacts of different local features.

To address the above-mentioned issue, we introduce a novel *weighted bilinear coding* model to adaptively weigh local features at different locations according to their relative importance. In our approach, the relative importance is automatically captured in the salient part masks generated by the salient part net. And the weighted bilinear coded feature

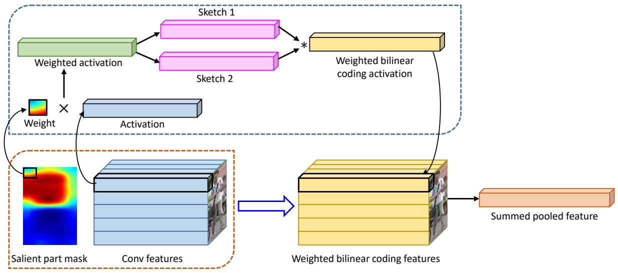


Figure 3: Illustration of the WBC model applied on one salient part mask.

is calculated as

$$\mathbf{F}_l = \sum_{p=1}^{H}\sum_{q=1}^{W}(\mathbf{M}_l(p,q)\mathbf{F}(p,q))^T(\mathbf{M}_l(p,q)\mathbf{F}(p,q)) \quad (8)$$

where $\mathbf{M}_l$ is the l-th part mask generated from Eq. (4). In this way, local features are weighed adaptively such that more critical units can play a more important role in the subsequent recognition process. The part feature $\mathbf{F}_l$ is then reshaped to a C^2 length vector and further passed through a signed square-root step ($\mathbf{F}_l = sign(\mathbf{F}_l)\sqrt{|\mathbf{F}_l|}$) before fed into the linear embedding layer to perform feature dimension reduction. In the WBC model, the outer product in Eq. (8) helps to capture richer local feature interactions, enhancing the representative ability of the final representation, and the weighting scheme encodes the relative importance of different local features, leading to more discriminative representation. Figure 3 illustrates the WBC model.

4 Experiments

In this section, we describe our experimental evaluation and provide a detailed ablation study of our proposed architecture. Extensive experiments on three challenging benchmarks including Market-1501 [Zheng *et al.*, 2015], DukeMTMC-reID [Ristani *et al.*, 2016] and CUHK03 [Li *et al.*, 2014] show that the proposed algorithm performs favorably against other state-of-the-art approaches. Some example ranking results are demonstrated in Figure 4.

4.1 Datasets and evaluation metric

Market-1501 is one of the most challenging datasets for Re-ID. It is collected in front of a supermarket using five high-resolution and one low-resolution cameras. In total, this dataset contains 32,768 annotated bounding boxes belonging to 1,501 identities obtained from existing pedestrian detection algorithm [Felzenszwalb *et al.*, 2010]. Among the 1,501 identities, 750 individuals are set for training and the rest for testing.

DukeMTMC-reID consists of 36,411 bounding boxes with labeled IDs, among which 1,404 identities appear in more than two cameras and 408 identities (distractor ID appears in only one camera). This dataset is further divided into training subset with 16,522 images of 702 identities, and testing subset with 2228 query images of the other 702 identities and 17,661 gallery images (images of the remaining 702 IDs and 408 distractor IDs).

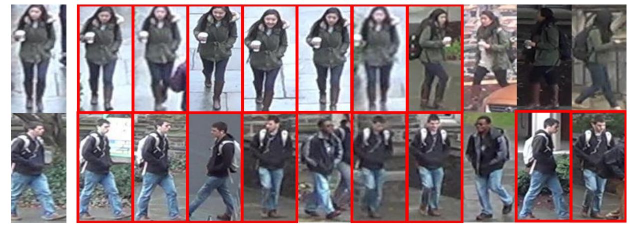


Figure 4: Example ranking results generated by our approach. The first column are two probe images, followed by ten top-ranked gallery images. Images marked by red bounding boxes are correct matches with the same identity as the probes.

CUHK03 contains 13,164 images of total 1,360 persons captured under six cameras. In this dataset, each individual appears in two disjoint camera views, and on average 4.8 images of each view are collected for each person. The performance is originally evaluated on 20 random splits of 1276 persons for training and 100 individuals for testing, which is time-consuming. Instead, we follow the evaluation protocol in [Zhong *et al.*, 2017] to split the dataset into training set with 767 identities and testing set with the rest identities. The CUHK03 benchmark provides both hand-labeled and DPM-detected [Felzenszwalb *et al.*, 2010] bounding boxes, we conduct experiments on both of them to validate the effectiveness of the proposed algorithm.

Following recent literature, all experiments are evaluated under the single-shot setting, where a ranking score is generated for each query image and all the scores are averaged to get the final recognition accuracy. The recognition performance are evaluated by the cumulative matching characteristic (CMC) curve and the mean average precision (mAP) criterion. The CMC curve represents the expected probability of finding the first correct match for a probe image in the top r match in the gallery list. And as supplementary, mean average precision summarizes the ranking results for all the correct matches in the gallery list.

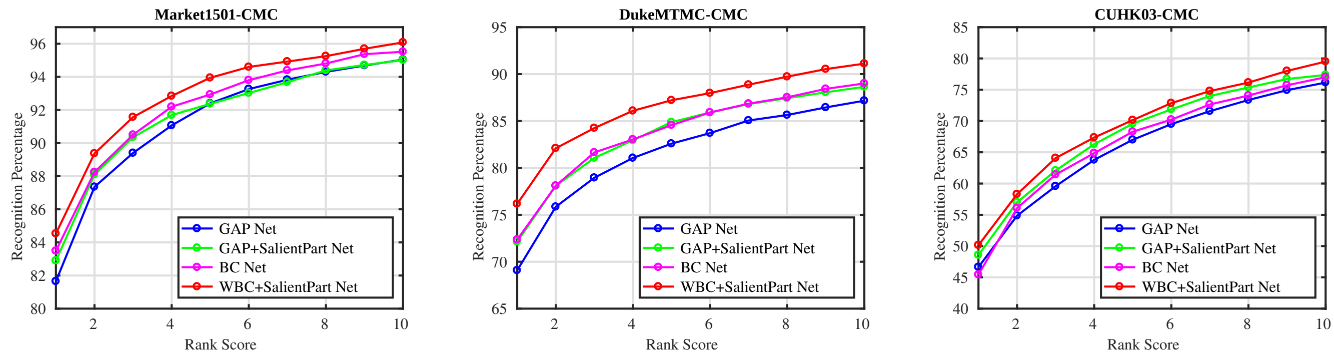
4.2 Implementation details

The proposed algorithm is implemented using Caffe [Jia *et al.*, 2014] on a NVIDIA GTX 1080 GPU with 8GB memory. We adopt the GoogLeNet [Szegedy *et al.*, 2015] as the base CNN network, and feature maps are extracted from the $inception\_4e$ layer, followed by a 1×1 convolutional layer with 512 feature channels. The input images are resized to 160×80. The number L of parts generated by the salient part net is discussed later, and distance margin α in Eq. (2) is set to 0.2 throughout the experiments. The whole network is optimized using stochastic gradient descent (SGD) method on mini-batches. The mini-batch size is set to 300, with on average 0.4 million triplets in each iteration. The initial learning rate is set to 0.008, and it is divided by 2 every 4,000 iterations. The weight decay and the momentum are set to 0.0005 and 0.9, respectively. The source code will be made publicly available upon acceptance.

4.3 Ablation study

4.3.1 Baseline comparisons

To further validate the proposed Re-ID algorithm, we conduct experiments on several baselines and compare with them.



(a) Market1501 (b) DukeMTMC-reID (c) CUHK03

Figure 5: Comparison results of the proposed algorithm with three baselines in terms of the top r matching rate using CMC (%).

In specific, we develop three baselines including GAP Net, GAP+SalientPart Net and BC Net as follows.

GAP Net is implemented by removing the salient part net from our method and replacing the proposed WBC with global average pooling, and other settings are kept exactly the same.

GAP+SalientPart Net is implemented by substituting the proposed WBC with global average pooling, and other settings are kept exactly the same.

BC Net is implemented by removing the salient part net from our method and replacing the proposed WBC with original bilinear coding (BC), and other settings are kept exactly the same.

Our method is referred to as **WBC+SalientPart Net**. The recognition performance of each network on three challenging large-scale person re-identification benchmarks are shown in Figure 5.

As demonstrated in Figure 5, our **WBC+SalientPart Net** performs consistently better than the other three baselines on all the three benchmarks. More specifically, in comparison with GAP, our approach has 2.8 %, 4.1 % and 4.7 % rank-1 performance gain on the Market1501 dataset, DukeMTMC-reID dataset and CUHK03 dataset respectively. Please note here the **GAP+SalientPart Net** baseline also aggregates local features over each salient part and concatenate them to form the final representation, but our approach achieves better performance, validating the more powerful representative ability of our *weighted bilinear coding* (WBC) than GAP. Furthermore, the comparison results with the **BC+SalientPart Net** baseline demonstrate the effectiveness of the weighting scheme introduced in our WBC model.

4.3.2 Analysis on different number of salient parts

We empirically study the optimal number L of salient parts on each dataset. In specific, we record the recognition performance of $L = 1, 3, 5, 8$, respectively. As shown in Table 1, on the Market1501, the best result is obtained with $L = 8$, which outperforms $L = 1$ by 1% in CMC recognition rate ($r = 1$) and 2% in term of mean average precision (mAP). On the DukeMTMC-reID dataset, $L = 3$ and $L = 5$ achieve better results than $L = 1, 8$, where $L = 3$ achieves the best CMC performance (76.2% for $r = 1$), while $L = 5$ has the highest mAP (56.94%). On the CUHK03 (with human-labeled bounding boxes), $L = 3$ outperforms all other settings ($L = 1, 5, 8$) in terms of both CMC and mAP. Overall, on all three datasets, multiple salient parts (with L bigger than 1) indeed bring performance gain to the proposed model. Since using multiple parts requires more computational resources, L is suggested to be set to 3 or 5 for the balance between effectiveness and efficiency. In the following part, the best result reported on each dataset is used for comparison with other state-of-the-art algorithms without special clarification.

Table 1: Analysis on the influence of different number of salient parts, both the CMC (%) top r ranking rates and the mAP (%) are reported.

Datasets	# parts	r=1	r=5	r=10	r=20	mAP
Market1501	L = 1	83.5	93.3	95.8	97.4	66.67
	L = 3	84.4	93.9	96.1	97.6	67.61
	L = 5	84.5	93.2	95.6	97.2	67.07
	L = 8	84.5	93.9	96.1	97.7	68.69
DukeMTMC	L = 1	74.9	85.6	89.4	92.0	54.75
	L = 3	75.8	87.4	90.7	93.4	56.94
	L = 5	76.2	87.2	91.1	93.5	56.85
	L = 8	75.9	86.2	90.6	93.4	56.40
CUHK03	L = 1	46.6	67.2	76.5	83.5	44.3
	L = 3	50.1	70.1	79.5	86.6	47.72
	L = 5	48.4	70.5	78.2	85.1	46.51
	L = 8	46.9	68.7	77.9	85.8	44.65

4.4 Comparison to state-of-the arts

In this section, we present the comparison results with state-of-the-art algorithms on Market-1501, DukeMTMC-reID and CUHK03 benchmarks.

4.4.1 Results on Market1501

On the Market1501 dataset, we compare the proposed Re-ID algorithm with many state-of-the-art algorithms, including feature designing based algorithms: LOMO+XQDA [Liao *et al.*, 2015] and BoW [Zheng *et al.*, 2015]; metric learning based algorithms: weighted approximate rank component analysis (WARCA) [Jose and Fleuret, 2016], SCSP [Chen *et al.*, 2016], Re-ranking [Zhong *et al.*, 2017] and DNS [Zhang

Table 2: Comparison of top r matching rate using CMC (%) and mean average precision (mAP %) on the Market1501 dataset. The best and second best results are marked in red and blue, respectively.

Methods	r=1	r=5	r=10	r=20	mAP
LOMO+XQDA	43.8	–	–	–	22.2
BoW	44.4	63.9	72.2	79.0	20.8
WARCA	45.2	68.2	76.0	–	–
SCSP	51.9	–	–	–	26.4
Re-ranking	77.1	–	–	–	63.6
DNS	55.4	–	–	–	29.9
Gated S-CNN	65.9	–	–	–	39.6
P2S	70.7	–	–	–	44.2
CADA	73.8	–	–	–	47.1
Spindle Net	76.9	91.5	94.6	96.7	
LSRO	79.3	–	–	–	56.0
MSCAN	80.3	–	–	–	57.5
PADF	81.0	92.0	94.7	–	63.4
SSM	82.2	–	–	–	**68.8**
SVDNet	82.3	92.3	95.2	–	62.1
ACRN	83.6	92.6	**95.3**	**97.0**	62.6
PDC	84.1	**92.7**	94.9	96.8	63.4
JLML	**85.1**	–	–	–	65.5
Ours	**84.5**	**93.9**	**96.1**	**97.7**	**68.7**

et al., 2016]; and deep learning based algorithms: Gated S-CNN [Varior *et al.*, 2016], set similarity learning (P2S) [Zhou *et al.*, 2017], consistent aware deep network (CADL) [Lin *et al.*, 2017], Spindle Net [Zhao *et al.*, 2017a], LSRO [Zheng *et al.*, 2017b], multi-scale context aware network (MSCAN) [Li *et al.*, 2017a], part aligned deep features (PADF) [Zhao *et al.*, 2017b], SSM [Bai *et al.*, 2017], JLML [Li *et al.*, 2017b] and pose-driven deep convolutional model (PDC) [Su *et al.*, 2017].

The detailed comparison results are reported in Table 2, from which we can see that in general our approach outperforms other state-of-the-art algorithms except a slightly lower rank-1 recognition rate than JLML [Li *et al.*, 2017b], and achieves very competitive recalls (the second best mAP). It is worth noting that PADF [Zhao *et al.*, 2017b] and PDC [Su *et al.*, 2017] are two deep learning based methods which utilize part-based strategy and adopt global average pooling for feature aggregation. In comparison to PADF and PDC, the proposed model consistently generates better performance, demonstrating the superiority of our WBC model over global average pooling.

4.4.2 Results on DukeMTMC-reID

On the DukeMTMC-reID dataset, we compare our method with LOMO+XQDA, BoW, LSRO, ACRN [Schumann and Stiefelhagen, 2017], PAN [Zheng *et al.*, 2017a], OIM [Xiao *et al.*, 2017], and SVDNet [Sun *et al.*, 2017], and the comparison results are listed in Table 3. As shown in Table 3, our algorithm performs better in general except a slightly lower rank-1 recognition rate compared to SVDNet [Sun *et al.*, 2017]. Considering that SVDNet adopts ResNet-50 for feature extraction, which is more powerful than GoogLeNet in our approach, we still achieve competitive performance.

Table 3: Comparison of top r matching rate using CMC (%) and mean average precision (mAP %) on the DukeMTMC-reID dataset. The best and second best results are marked in red and blue, respectively.

Methods	r=1	r=5	r=10	r=20	mAP
LOMO+XQDA	52.4	74.5	83.7	89.9	
BoW	25.1	–	–	–	12.2
LSRO	67.7	–	–	–	47.1
ACRN	72.6	84.8	88.9	91.5	52.0
PAN	71.6	83.9	–	90.6	51.5
OIM	68.1	–	–	–	–
SVDNet	**76.7**	**86.4**	**89.9**	–	**56.8**
Ours	**76.2**	**87.2**	**91.1**	**93.5**	**56.9**

4.4.3 Results on CUHK03

On the CUHK03 dataset, we record the performance on both settings of human-labeled and auto-detected bounding boxes and compare it with LOMO+XQDA, BOW+XQDA [Zheng *et al.*, 2015], IDE+DaF [Yu *et al.*, 2017], PAN, DPFL [Chen *et al.*, 2017], Re-ranking, and SVDNet. Table 4 demonstrates the detailed comparison results. As shown, our approach obtains the best rank-1 recognition rate on both human-labeled and detected datasets, outperforming the second best by 7.1% and 2.4% respectively. Besides, the mAP of our algorithm also has large performance gain compared with the state-of-the-arts (improved by 7.2% and 4.7% respectively).

Table 4: Comparison of top r matching rate using CMC (%) and mean average precision (mAP %) on the CUHK03 dataset. The best and second best results are marked in red and blue, respectively.

	Labeled		Detected	
Methods	r=1	mAP	r=1	mAP
LOMO+XQDA	14.8	13.6	12.8	11.5
BOW+XQDA	7.9	7.3	6.4	6.4
Re-ranking	38.1	40.3	34.7	**37.4**
IDE+DaF	27.5	31.5	26.4	30.0
PAN	36.9	35.0	36.3	34.0
DPFL	**43.0**	**40.5**	40.7	37.0
SVDNet	40.9	37.8	**41.5**	37.3
Ours	**50.1**	**47.7**	**43.9**	**42.1**

5 Conclusions

This paper proposes a novel *weighted bilinear coding* (WBC) model to pursue more representative and discriminative aggregation for the intermediate convolutional features in CNN networks. In specific, channel-wise feature correlations are encoded to model higher-order feature interactions, improving the representative ability. Moreover, a weighting scheme is adopted to adaptively weigh local features to reflect local feature importance. Besides, to deal with spatial misalignment, a salient part net is introduced to automatically derive salient body parts. By integrating the WBC model and the salient part net, the final human appearance representation is both discriminative and resistant to spatial misalignment. Extensive experiments on three large-scale benchmarks demonstrate the effectiveness of the proposed approach.

References

[Bai *et al.*, 2017] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017.

[Chen *et al.*, 2016] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016.

[Chen *et al.*, 2017] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2017.

[Cheng *et al.*, 2016] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *CVPR*, 2016.

[Felzenszwalb *et al.*, 2010] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[Jia *et al.*, 2014] Y. Jia, E Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[Jose and Fleuret, 2016] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, 2016.

[Köstinger *et al.*, 2012] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[LeCun *et al.*, 1989] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

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

[Li *et al.*, 2017a] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[Li *et al.*, 2017b] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017.

[Liao *et al.*, 2015] S. Liao, Y. Hu, X. Zhu, and S. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[Lin *et al.*, 2015] T.-Y. Lin, A. Chowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, 2015.

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

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

[Schumann and Stiefelhagen, 2017] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPR Workshop*, 2017.

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

[Sun *et al.*, 2017] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[Szegedy *et al.*, 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Varior *et al.*, 2016] R. Rama Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[Xiao *et al.*, 2017] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[Xiong *et al.*, 2014] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[Yao *et al.*, 2017] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv*, 2017.

[Yu *et al.*, 2017] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. In *BMVC*, 2017.

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

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

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

[Zhao *et al.*, 2017c] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by saliency learning. *TPAMI*, 39(2):356–370, 2017.

[Zheng *et al.*, 2011] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

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

[Zheng *et al.*, 2016] L. Zheng, Y. Yang, and A. Hauptmann. Person re-identification: Past, present and future. *arXiv*, 2016.

[Zheng *et al.*, 2017a] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv*, 2017.

[Zheng *et al.*, 2017b] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, 2017.

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

[Zhou *et al.*, 2017] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, 2017.

[Zhou *et al.*, 2018] Q. Zhou, H. Fan, S. Zheng, H. Su, X. Li, S. Wu, and H. Ling. Graph correspondence transfer for person re-identification. In *AAAI*, 2018.