Neurocomputing 407 (2020) 454–464

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weighted bilinear coding over salient body parts for person re-identification

Zhigang Chang [a], Zhou Qin [a,e], Heng Fan [b], Hang Su [d], Hua Yang [a], Shibao Zheng [a,*], Haibin Ling [b,c]

[a] Institute of Image Processing and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Department of Computer & Information Sciences, Temple University, Philadelphia 19122, USA
[c] Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[d] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[e] Artificial Intelligence Center-City Brain, Alibaba Cloud, Hangzhou 311100, China

ARTICLE INFO

Article history:
Received 3 January 2019
Revised 9 January 2020
Accepted 1 May 2020
Available online 12 May 2020
Communicated by D. Tao

Keywords:
Person re-identification
Feature aggregation
Bilinear coding

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) framework for local feature aggregation in CNN networks to pursue more representative and discriminative feature representations, which can adapt to other advanced methods and improve their performance. 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 encoded 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 outstanding methods.

1. Introduction

Person re-identification (Re-ID) [36,43–47,58] 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.

Motivated by the success in image classification [25,8,22], semantic segmentation [19] and tracking [5,6], convolutional neural networks (CNNs) [12] have been widely utilized for Re-ID because of its 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 (GAP) to aggregate the convolutional features to represent human appearance [33,23]. 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) framework 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. The superiority of our WBC framework comes from two aspects: Firstly, the bilinear coding takes into consideration

[*] Corresponding author.
E-mail address: sbzh@sjtu.edu.cn (S. Zheng).

https://doi.org/10.1016/j.neucom.2020.05.009
0925-2312/© 2020 Elsevier B.V. All rights reserved.

the channel-wise correlations of each local feature. In comparison to global average pooling, bilinear coding captures richer feature information. More importantly, 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 framework is flexible and can be embedded into arbitrary networks as a feature aggregation part. The WBC framework provides a new template for higher-order feature aggregation and the popular Re-ID models such as Refined Part Pooling (PCB-RPP) [46] assembled with WBC get a significant improvement in performance compared to the original models.

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 several 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. So the representations over the parts are learned end to end and the similarities between the corresponding parts are aggregated. Therefore, each branch of the salient part net can learn attention mask of a local area of original feature map due to feature concatenation and triplet loss learning. The proposed Re-ID framework with salient part net and WBC, is illustrated in Fig. 1.

In summary, we make the following contributions:

- We propose a novel framework for representative and discriminative feature aggregation considering channel-wise correlations of aligned local features, which can be flexibly plugged into existing deep architectures.
- To alleviate the spatial misalignment problem, we integrate the WBC model with a salient part network to pursue part-aligned higher-order interacted features in an end-to-end trainable network for Re-ID.
- Extensive experiments on three large-scale datasets including Market-1501 [35], DukeMTMC-reID [20] and CUHK03 [14] demonstrate the favorable performance of our framework against state-of-the-art approaches. Some experiments demonstrate that the popular Re-ID models with WBC outperforms the original models, which has proven its flexibility and generalization.

The rest of this paper is organized as follows. Section 2 reviews the relevant works of this paper. Section 3 illustrates the details of the proposed Re-ID method. Section 4 demonstrates the experimental results, followed by conclusion in Section 5.

2. Related work

Being extensively studied, numerous approaches have been proposed for Re-ID in recent years [36]. Sorted from the perceived scale of the extracted features, the Re-ID methods 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 part-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 [37,11,28]. These methods, nevertheless, degrade in presence of spatial misalignment caused by large variations in view angles and human poses.

Due to the inaccuracy of the foreground boxes of the pedestrian, such as scale problems, occlusion, environmental noise, etc., the aggregated parts features provide higher discriminative ability than global features.

2.1. Part-based re-ID

To alleviate the problem of spatial misalignment, many part-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 [2], 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 [4] 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 [34,41] 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, establishing 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 [29], 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, [33] 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,

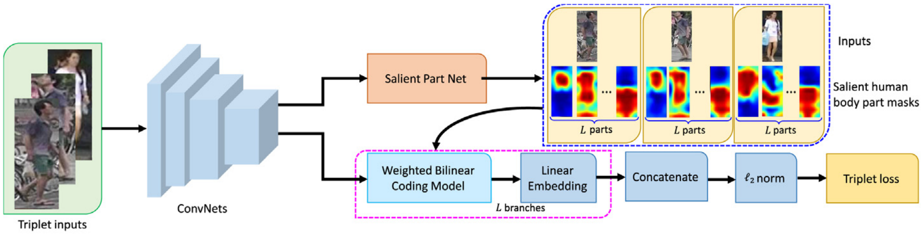


Fig. 1. Illustration of the proposed person Re-ID framework. Based on the feature maps extracted from the backbone ConvNets, we first adopt a salient part net to obtain salient human body parts, then the proposed WBC model is applied on each part for discriminative feature aggregation. The final representation 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.

the usage of global average pooling for feature aggregation in these approaches [29,33] leads to sub-optimal results due to the ignorance of richer feature interactions and various feature significance. Part-based Convolutional Baseline (PCB) and Refined Part Pooling (RPP) have been proposed by [46] to raise a new baseline and many other improved versions [46] have been released.

2.2. Second order feature aggregation

Feature aggregation is an essential part of visual tasks, which means encoding and pooling of visual features to make them both efficient and discriminative, like fisher encoding [53] and spatial pyramid [54]. In CNNs, fully connected layer, average pooling and max pooling are popular operations for feature aggregation. Bilinear models are designed to separate style and content by [48]. Then the second order pooling have been explored using deeply-learned features [52,51] and hand-crafted features [49]. However, bilinear features [18] always involve high dimensional operations of the matrix with high computational complexity, typically on the order of a few million. Compact Bilinear Pooling [50] provides an efficient kernelized solution with same discriminative power but with only a few thousand dimensions.

Despite being related to [33], our approach is significantly different. In this paper, we focus on improving the performance of person Re-ID by learning discriminative fine-grained feature aggregation. To this end, we present a novel WBC model. Different from [33] using GAP for feature aggregation, our WBC model takes into account higher-order channel-wise feature interactions, enriching the representative ability and discriminability of the learned feature embedding. Experimental results evidence the advantages of our WBC model compared to GAP in [33] for Re-ID.

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

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

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

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

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

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

where $|\mathcal{T}|$ denotes the number of triplets in $\mathcal{T}$. The loss function will be presented in Section 3.4.

3.2. Salient part-based representation

For Re-ID, one of the most technical bottleneck is spatial misalignment caused by variations in views and human poses. Local representations are computed typically by partitioning full images into fixed horizontal stripes or grids, assuming the bounding boxes are well detected and spatial distributions of human body are similar. To deal with this issue, we adopt a salient part-based representation (spatial attention module) to measure the importance of human appearance parts in spatial domain. We call it salient part net. Images are passed into the backbone to get the intermediate feature maps. The salient part masks are generated by feeding the obtained feature maps into the salient part net. Specifically, the salient part nets consist of L branches, each corresponding to a certain part of the human body. For simplicity, each branch is composed of a 1×1 conv, followed by a Sigmoid layer to map the values to $(0, 1)$. So the representations over the parts are learned end-to-end and the similarities between the corresponding parts are aggregated. Therefore, each convolutional branch can learn attention mask of a local area of original feature map due to feature concatenation and triplet loss learning.

Fig. 2 illustrates the architecture of the 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. The input to the salient part net is the 3-dimension intermediate convolutional feature maps, and its outputs are L2-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 $\mathbf{M}_l \in \mathbb{R}^{H \times W \times 1}, l \in \{1, \cdots, L\}$ as

$$\mathbf{M}_l = \Phi_{\text{SalientMask}_l}(\mathbf{F}) \quad (4)$$

where $\Phi_{\text{SalientMask}_l}(\cdot)$ represents the l^{th} salient part mask generator. In Eq. (4), the values of elements in each $\mathbf{M}_l$ are within the range $(0, 1)$, reflecting 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

$$\mathbf{F}_l = \Psi_{\text{WBC}}(\mathbf{M}_l, \mathbf{F}) \quad (5)$$

where $\Psi_{\text{WBC}}(\cdot, \cdot)$ represents the proposed feature coding algorithm and will be discussed later. It is worth noting that different from existing part-aligned Re-ID method [33] which uses global average pooling for feature aggregation, our WBC is able to fully explore richer higher-order 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,

$$\mathbf{f} = \phi(\mathbf{I}) = \left\| \left[(\mathbf{F}_{1'})^\top, (\mathbf{F}_{2'})^\top, \cdots, (\mathbf{F}_{L'})^\top \right]^\top \right\|_2 \quad (6)$$

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 [33,29] adopt global average pooling (GAP) for the sake of efficiency and simplic-

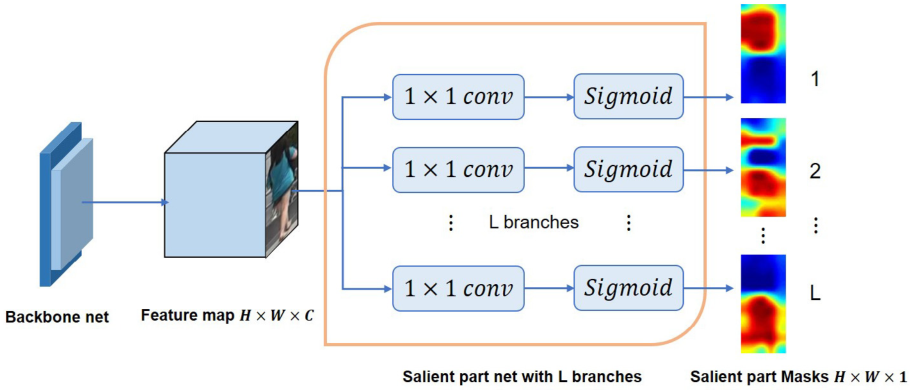

Backbone net
Feature map $H \times W \times C$
1×1 conv
Sigmoid
1×1 conv
Sigmoid
L branches
1×1 conv
Sigmoid
1
2
L
Salient part net with L branches
Salient part Masks $H \times W \times 1$

Fig. 2. Illustration of the architecture of salient part net.

ity. 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 [52,51] 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 **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. The process is fully shown in Fig. 3. 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 is calculated as

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

where $\mathbf{M}_l$ is the l-th part mask generated from Eq. (4).

The details of bilinear operations on each position of activation map are illustrated in Fig. 3. The salient part masks are generated by salient part net. For each salient mask,the weighted activation $1 \times 1 \times C$ are resized to $C \times 1$ and $1 \times C$, then they do the outer product operation to produce a $C \times C$ matrix in Eq. (8). In this way, local features are weighted adaptively such that more critical units can play a more important role. The activation is then reshaped to a C^2 length vector, accumulate and recover to the weighted bilinear coding feature map $H \times W \times (C * C)$. It further passed through a sum-pooling layer, a signed square-root step ($\mathbf{F}_l = sign(\mathbf{F}_l)\sqrt{|\mathbf{F}_l|}$) and a $L2$ normalization layer before fed into the linear embedding layer to perform feature dimension reduction. In the WBC model, the outer product helps to capture richer local feature interactions, enhancing the representative ability of the deep features. Meanwhile, the weighting scheme encodes the relative importance of different local features, leading to more discriminative representation.

3.4. Loss function

We use the batch hard triplet loss function which was originally proposed in [43]. To form a batch, we randomly sample P identities and randomly sample K clips for each identity (each clip contains T frames); Totally there are PK clips in a batch. For each sample a in the batch, the hardest positive and the hardest negative samples within the batch are selected when forming the triplets for computing the loss $L_{triplet}$.

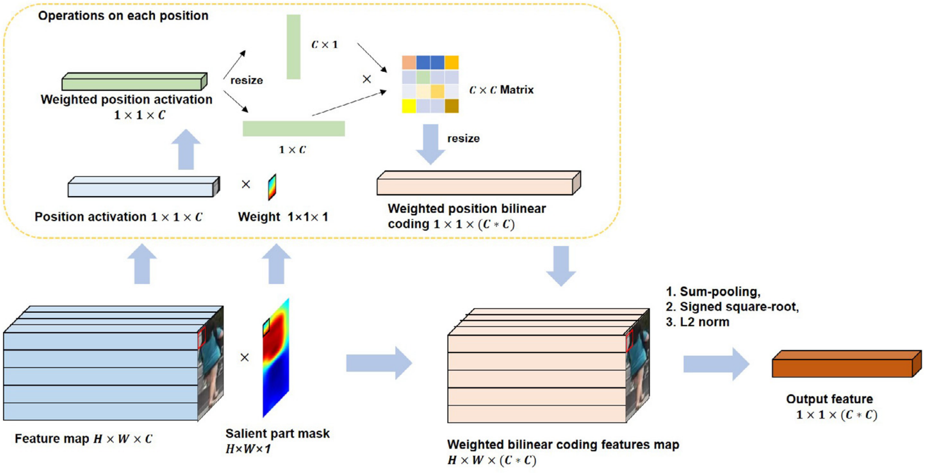

Operations on each position
$C \times 1$
resize
×
$C \times C$ Matrix
Weighted position activation $1 \times 1 \times C$
$1 \times C$
resize
Position activation $1 \times 1 \times C$
×
Weight $1\times1\times1$
Weighted position bilinear coding $1 \times 1 \times (C * C)$
Feature map $H \times W \times C$
×
Salient part mask $H\times W\times 1$
Weighted bilinear coding features map $H \times W \times (C * C)$
1. Sum-pooling,
2. Signed square-root,
3. L2 norm
Output feature $1 \times 1 \times (C * C)$

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

$$L_{tripletloss} = \sum_{i=1}^{P}\sum_{a=1}^{K}\left[m + \overbrace{\max_{p=1\dots k} D\left(\phi\left(\mathbf{I}_a^i\right), \phi\left(\mathbf{I}_p^i\right)\right)}^{hardest\ \ positive} - \underbrace{\min_{\substack{j=1\dots P \\ j \neq i \\ n=1\dots K}} D\left(\phi\left(\mathbf{I}_a^i\right), \phi\left(\mathbf{I}_n^j\right)\right)}_{hardest\ \ negative} \right] \tag{9}$$

3.5. Flexibility and portability

Functioning as a feature aggregation part, the proposed WBC module can be readily combined with state-of-the-art feature networks to enhance their recognition performance. In this section, we combine WBC with the popular PCB-RPP [46] method to further boost its performance. In PCB-RPP [46], it used a learned classifiers containing a linear layer followed by Softmax activation to relocate and weigh the predicted probability of position activation in feature map belonging to local part. In specific, we take the output after the Softmax operation in RPP module in [46] as the salient mask, then features of each part are encoded by the proposed WBC module to form the final part features. For fair comparison, other parts of the PCB-RPP algorithm keep exactly the same. This part of experiments will be discussed in detail in Section 4.

4. Experiments

In this section, we describe our evaluation protocols and provide a detailed ablation study of the proposed architecture. Extensive experiments on three challenging benchmarks including Market-1501 [35], DukeMTMC-reID [20] and CUHK03 [14] show that the proposed algorithm performs favorably against other approaches.

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 [7]. 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).

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 [40] to split the dataset into training set composed of 767 identities and testing set with the rest identities. The CUHK03 benchmark provides both hand-labeled and DPM-detected [7] 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

We adopt the GoogLeNet [25] and Resnet-50 [8] as the backbones CNN network.

4.2.1. GoogLeNet backbone

Feature maps are extracted from the *inception_4e* layer when GoogLeNet, followed by a 1 × 1 convolutional layer with 512 feature channels. The input images are resized to 160 × 80. The number of parts (L) generated by the salient part net is discussed later, and distance margin α in Eq. (9) is set to 0.3 throughout the experiments. The whole network is optimized using stochastic gradient descent (SGD) method on mini-batches. Each mini-batch is sampled with randomly selected P identities and randomly sampled K images for each identity from the training set. $P = 32, K = 4$ when Market-1501 and DukeMTMC-reID, and $P = 16, K = 4$ when CUHK03. 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. The channel dimension is set to 1024 after the linear embedding in Weighted Bilinear Coding Module.

4.2.2. Resnet-50 backbone

We also implement a Resnet-50 backbone as comparison. The Resnet-50 baseli Weights of pretrained ResNet-50 [8] initializes the backbone. We removed the last spatial downsampling in the last residual block in ResNet50 to increase the size of feature map. Input images are resized to 384 × 128, so the final output feature map size before pooling layers is 24 × 8 and the channel dimension is 2048. ADAM optimizer is used with momentum 0.9. We set 1.2 for the margin parameter for triplet loss and 10^{-4} for initial learning rate. Learning rate decays to 10^{-5} and 10^{-6} after training for 120 and 160 epochs. Random horizontal flipping for data augmentation and features averaged from original images and the horizontally flipped versions for evaluation are deployed. The channel dimension is set to 1024 after the linear embedding in Weighted Bilinear Coding Module.

4.3. Ablation study

4.3.1. Modules validity analysis

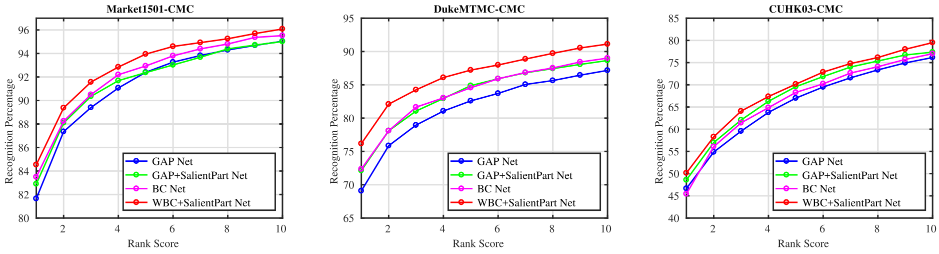
To further validate the proposed Re-ID algorithm, we conduct experiments on several baselines and compare with them on GoogLeNet Backbone. In specific, we develop three baselines including GAP Net, GAP + SalientPart Net and BC Net on GoogLeNet Baseline as follows.

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. L is set to 3 in this part.

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**. L is set to 3 in this part. The recognition performance of each network on three challenging large-scale person re-identification benchmarks are shown in Fig. 4.



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

Fig. 4. Comparison results of the proposed algorithm with three baselines in terms of the top r matching rate using CMC. Best viewed in color.

As demonstrated in Fig. 4, 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 Net** baseline demonstrate the effectiveness of the weighting scheme in our WBC framwork.

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 and Fig. 5, 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

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

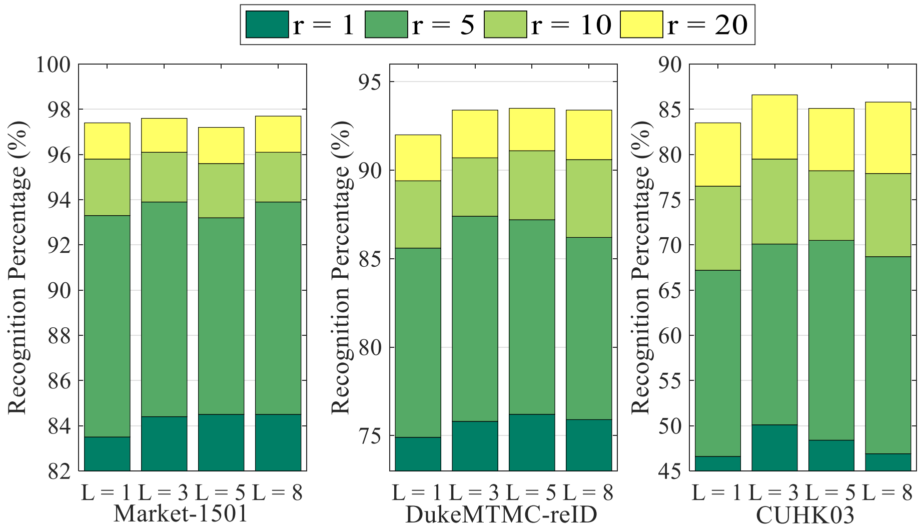


Fig. 5. Recognition performance of different number of salient parts. As shown, the number of parts set to $L = 3$ is a good compromise between efficiency and effectiveness.

$L = 5$ achieves the best CMC performance (76.2% for $r = 1$), while $L = 3$ 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 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.

4.3.3. Performance comparisons between different backbones along with Flexibility and Portability verification analysis

In this section, Firstly we compare the performances of two backbone networks, the purpose of comparison is to measure the performance gain when transferring to the Resnet-50. Because the Resnet-50 is also the backbone network of PCB-RPP [46] many other methods, we can more easily and directly compare with others algorithms. Secondly, we assemble our WBC framework with PCB + RPP method mentioned above in Section 3.5.

GAP-GoogLeNet is the GAP Net in Section 4.3.1 based on GoogLeNet backbone. **GAP-Resnet** is the same structure based on Resnet-50 backbone. **BC-GoogLeNet** is the BC Net in Section 4.3.1 based on GoogLeNet backbone removing the salient part net from our method and replacing the proposed WBC with original bilinear coding (BC). **BC-Resnet** is the same structure based on Resnet-50 backbone. **WBC-full-GoogLeNet** is the complete implementation as **WBC + SalientPart Net** in Section 4.3.1 based on GoogLeNet backbone. **WBC-full-Resnet** is the same structure based on Resnet-50 backbone. L is set to 3 in this part.

The comparison of different backbones performance is shown in the upper part of Table 2. After transferring to Resnet-50, our method **WBC + SalientPart Net** has significant performance gain on Market1501 and DukeMTMC datasets. Rank1 accuracy increased by 7.7 % and 5.9 % while mAP rises by 14.3 % and 12.2 % on both datasets. Incomplete implements such as GAP Net and BC Net can also confirm that considerable gain has been achieved after transferring to Resnet-50 backbone network.

The PCB-RPP [46] model uses Cross Entropy Loss as loss function, we completely follow the setting of the paper. The WBC module serves as the pooling operation upon the weighted part-based features. The part-based feature channel dimension is set to 2048 after the Weighted Bilinear Coding layer. The other settings of RPP remain the same. As shown in the lower part of Table 2, the mAP increased by 1.7 % and 3.5 % on both datasets. Apart from rank1 accuracy in Market1501, Other performance indicators listed have improved significantly, which effectively prove that our framework can be flexibly plugged into existing deep architectures like PCB-RPP.

4.4. Comparison with 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 based on two backbones with many Re-ID algorithms, including feature designing based algorithms: LOMO + XQDA [16] and BoW [35]; metric learning based algorithms: weighted approximate rank component analysis (WARCA) [10], SCSP [2], Re-ranking [40] and DNS [31]; and deep learning based algorithms: Gated S-CNN [26], set similarity learning (P2S) [42], consistent aware deep network (CADL) [17], Spindle Net [32], LSRO [39], multi-scale context aware network (MSCAN) [13], part aligned deep features (PADF) [33], SSM [1], SVDNet [24], ACRN [21], JLML [15], pose-driven deep convolutional model (PDC) [23],Harmonious Attention Network (HA-CNN) [57], AOS [56], Multi-Level Factorisation Net (MLFN) [55] and PCB-RPP [46].

The detailed comparison results are reported in Table 3, from which we can see that in general our approach based on Resnet backbones outperforms most of other state-of-the-art algorithms except a slightly lower performance than PCB-RPP [46]. But it has been proven that our framework with RPP in Table 2 has better performance than PCB-RPP.

Only using GoogLeNet as backbone, its performance can beat other methods listed above JLML [15]. Considering that JLML [15] utilizes the ResNet-50 [8] as the backbone network, which is more powerful than our adopted GoogLeNet, our performance is still competitive. Besides, our approach achieves very competitive recalls (the second best mAP). It is worth noting that PADF [33] and PDC [23] 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 [21], PAN [38], OIM [27], SVDNet [24],Harmonious Attention Network (HA-CNN) [57], AOS [56], Multi-Level Factorisation Net (MLFN) [55] and PCB-RPP [46]. and the comparison results are listed in Table 4.

As shown in Table 4, our approach based on Resnet backbones outperforms most of other state-of-the-art algorithms except a slightly lower performance than PCB-RPP [46]. But it has been proven that our framework with RPP in Table 2 has better performance than PCB-RPP.

Table 2
Performance comparisons between different backbone on datasets Market1501 and DukeMTMC-reID.

Method	Market1501			DukeMTMC		
	mAP	rank1	rank5	mAP	rank1	rank5
GAP-GoogLeNet	59.6	81.8	92.3	47.2	69.3	82.4
GAP-Resnet	66.0	84.6	91.1	55.3	72.3	86.5
BC-GoogLeNet	63.4	83.7	88.3	53.9	72.6	84.9
BC-Resnet	77.1	87.4	94.0	63.2	77.5	89.1
WBC-full-GoogLeNet	67.6	84.4	93.9	56.9	75.8	87.4
WBC-full-Resnet	**81.9**	**92.1**	**96.5**	**69.1**	**81.7**	**91.4**
PCB-triplet	74.9	88.9	94.7	64.1	77.6	88.9
PCB[46]	77.4	92.3	97.2	66.1	81.8	89.4
PCB + RPP[46]	81.6	**93.7**	97.5	69.2	83.3	90.5
WBC + RPP	**83.3**	93.6	**98.1**	**72.7**	**84.1**	**91.7**

Table 3
Comparison results of top *r* matching rate using CMC (%) and mean average precision (mAP %) on the Market1501 dataset.

Methods	r = 1	r = 5	r = 10	r = 20	mAP
LOMO + XQDA [16]	43.8	–	–	–	22.2
BoW [35]	44.4	63.9	72.2	79.0	20.8
WARCA [10]	45.2	68.2	76.0	–	–
SCSP [2]	51.9	–	–	–	26.4
Re-ranking [40]	77.1	–	–	–	63.6
DNS [31]	55.4	–	–	–	29.9
Gated S-CNN [26]	65.9	–	–	–	39.6
P2S [42]	70.7	–	–	–	44.2
CADL [17]	73.8	–	–	–	47.1
Spindle Net [32]	76.9	91.5	94.6	96.7	
LSRO [39]	79.3	–	–	–	56.0
MSCAN [13]	80.3	–	–	–	57.5
PADF [33]	81.0	92.0	94.7	–	63.4
SSM [1]	82.2	–	–	–	68.8
SVDNet [24]	82.3	92.3	95.2	–	62.1
ACRN [21]	83.6	92.6	95.3	97.0	62.6
PDC [23]	84.1	*f*92.7	94.9	96.8	63.4
JLML [15]	85.1	–	–	–	65.5
AOS [56]	86.5	–	–	–	70.4
MLFN [55]	90.0	–	–	–	74.3
HA-CNN [57]	91.2	–	–	–	75.7
PCB-RPP [46]	93.8	97.5	98.5	–	81.6
WBC-full-GoogLeNet	84.5	93.9	96.1	97.7	68.7
WBC-full-Resnet	92.1	96.5	98.6	99.7	81.9

Table 4
Comparison results of top *r* matching rate using CMC (%) and mean average precision (mAP %) on the DukeMTMC-reID dataset.

Methods	r = 1	r = 5	r = 10	r = 20	mAP
LOMO + XQDA [16]	52.4	74.5	83.7	89.9	
BoW [35]	25.1	–	–	–	12.2
LSRO [39]	67.7	–	–	–	47.1
ACRN [21]	72.6	84.8	88.9	91.5	52.0
PAN [38]	71.6	83.9	–	90.6	51.5
OIM [27]	68.1	–	–	–	–
SVDNet [24]	76.7	86.4	89.9	–	56.8
AOS [56]	79.2	–	–	–	62.1
MLFN [55]	81.0	–	–	–	62.8
HA-CNN [57]	80.5	–	–	–	63.8
PCB-RPP [46]	83.3	90.5	92.5	–	69.2
WBC-full-GoogLeNet	76.2	87.2	91.1	93.5	56.9
WBC-full-Resnet	81.7	91.4	93.1	95.7	69.1

Only using GoogLeNet as backbone, our model's performance can beat other methods listed above SVDNet [24] except a slightly lower rank-1 recognition rate compared to SVDNet [24]. In ACRN [21], the authors propose to separately learn a classifier to leverage the complementary information of attributes for better representation of the human appearance. The notable performance gain over ACRN [21] (3.6% in rank 1 recognition rate and 4.9% in mAP) clearly demonstrates that our approach can generate more representative appearance descriptors without the need for extra attribute annotations. Comparing Table 4 with Table 1, we can observe that our WBC model with one salient part ($L = 1$ in Table 1) demonstrates superior performance than PAN [38], reflecting the discriminative ability of the proposed weighted bilinear coding model. The performance gain with bigger L ($L = 5$ in Table 1) further validates the superiority of our salient part based representation over the global parameter aligned deep features utilized in PAN [38]. Our approach achieves slightly lower rank-1 recognition performance than SVDNet [24] because SVDNet adopts ResNet-50 for feature extraction, which is more powerful than GoogLeNet in our approach.

4.4.3. Results on CUHK03

On the CUHK03 dataset, we follow the new evaluation protocol in [40] to demonstrate the effectiveness of the proposed algorithm. We record the recognition performance on both settings of human-labeled and auto-detected bounding boxes and compare it with LOMO + XQDA, BOW + XQDA [35], IDE + DaF [30], PAN, DPFL [3], Re-ranking, and SVDNet. Table 5 demonstrates the detailed comparison results. In this section, we product experiments based on GoogLeNet backbone.

As shown in Table 5, under the 767/700 setting, our approach clearly outperforms the other state-of-the-arts. For example, on the labeled bounding-boxes, our approach outperforms the second best DPFL [3] by 7.1% in rank-1 recognition rate and 7.2% in mAP. Meanwhile on the detected bounding-boxes, we gain by 2.4% in rank-1 recognition rate and 4.7% in mAP. It is worth noting that PAN [38] is also an alignment net based deep feature learning model. Our approach reports to be superior than PAN [38] in both the labeled and detected settings. We attribute the obvious performance gain to two factors: (1) The richer higher-order information encoded in our weighted bilinear coding model brings more representative ability to the deep features, and (2) The salient part network adopted in our model achieves more flexibility and better alignment performance than the globally parameterized spatial transformer network in PAN [38]. Another highly expected phenomenon on the CUHK03 dataset is that recognition performance on the manually-labeled images is indeed better than that of the

Table 5
Comparison results of top *r* matching rate using CMC (%) and mean average precision (mAP %) on the CUHK03 dataset.

Methods	Labeled		Detected	
	r = 1	mAP	r = 1	mAP
LOMO + XQDA [16]	14.8	13.6	12.8	11.5
BOW + XQDA [35]	7.9	7.3	6.4	6.4
Re-ranking [40]	38.1	40.3	34.7	37.4
IDE + DaF [30]	27.5	31.5	26.4	30.0
PAN [38]	36.9	35.0	36.3	34.0
DPFL [3]	43.0	40.5	40.7	37.0
SVDNet [24]	40.9	37.8	41.5	37.3
WBC-full-GoogLeNet	50.1	47.7	43.9	42.1

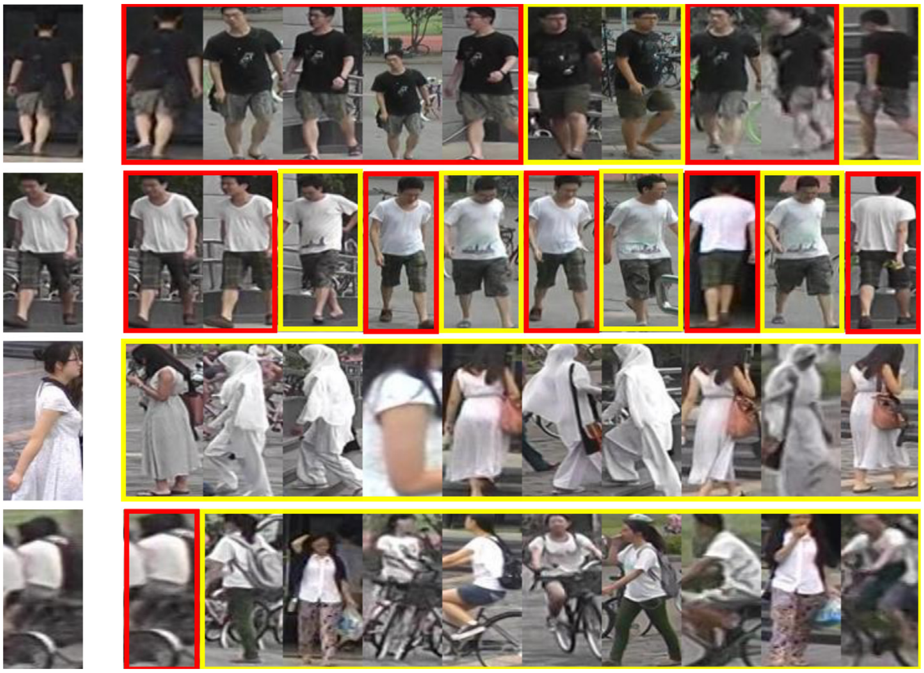


Fig. 6. Instances of some typical failure cases. The first column are four probe images, and the following are ten most similar images generated by the proposed algorithm. Images in the red bounding boxes are true matches with the same identity, and images in the yellow bounding boxes are false positives with different identities. Generally, the proposed algorithm can rank visually similar images ahead of others. Best viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

auto-detected images, which reflects that more severe misalignment and noisy background clutters present in the auto-detected images need further consideration during the model design.

4.5. Typical failure cases analysis

In order to explore the limitations of the proposed algorithm, we present some typical failure cases on the Market-1501 dataset. As shown in Fig. 6, there are mainly three factors that lead to undesirable matching results. The first two rows illustrate that when the negative instances are very similar with the probe image, false positives may be ranked ahead of true matches (e.g., the false positives in the first and second rows only slightly differ from the related probe images in the patterns on the T-shirts). In the third row, our algorithm does not find out the true matches within the top-10 rankings. This can be attributed to the partial detection result of the probe image, resulting in incomplete feature representation of the whole body. In the last row, blurring and serious background clutters misleads our algorithm into matching the probe with images in white upper clothes and on/beside a bicycle.

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. The flexibility and generalization of this framework also have been proven.

CRediT authorship contribution statement

Zhigang Chang: Methodology, Software, Validation, Writing - original draft. **Zhou Qin:** Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Heng Fan:** Methodology, Writing - review & editing. **Hang Su:** Supervision, Writing - review & editing. **Hua Yang:** Writing - review & editing. **Shibao Zheng:** Writing - review & editing. **Haibin Ling:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 61771303, 61671289), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 19DZ1209303, 18DZ1200102, 18DZ2270700), and SJTUYitu/Thinkforce Joint laboratory for visual computing and application.

References

[1] S. Bai, X. Bai, Q., Tian, Scalable person re-identification on supervised smoothed manifold, in: CVPR, 2017..
[2] D. Chen, Z. Yuan, B. Chen, N., Zheng, Similarity learning with spatial constraints for person re-identification, in: CVPR, 2016..
[3] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: ICCV, 2017..
[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: CVPR, 2016..
[5] H. Fan, H. Ling, Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking, in: ICCV, 2017a..
[6] H. Fan, H. Ling, Sanet: Structure-aware network for visual tracking, in: CVPRW, 2017b..
[7] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, TPAMI 32 (2010) 1627–1645.
[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016. pp. 770–778..
[10] C. Jose, F., Fleuret, Scalable metric learning via weighted approximate rank component analysis, in: ECCV, 2016..
[11] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: CVPR, 2012..
[12] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (1989) 541–551.
[13] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: CVPR, 2017a..
[14] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: CVPR, 2014..
[15] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: IJCAI, 2017b..
[16] S. Liao, Y. Hu, X. Zhu, S., Li, Person re-identification by local maximal occurrence representation and metric learning, in: CVPR, 2015..
[17] J. Lin, L. Ren, J. Lu, J. Feng, J. Zhou, Consistent-aware deep learning for person re-identification in a camera network, in: CVPR, 2017..
[18] T.Y. Lin, A. Chowdhury, S., Maji, Bilinear CNN models for fine-grained visual recognition, in: ICCV, 2015..
[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015..
[20] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: ECCV, 2016..
[21] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: CVPR Workshop, 2017..
[22] K. Simonyan, A., Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, 2015..
[23] C. Su, J., Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: ICCV, 2017..
[24] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: ICCV, 2017..
[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015..
[26] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: ECCV, 2016..
[27] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: CVPR, 2017..
[28] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: ECCV, 2014..
[29] H. Yao, S., Zhang, Y. Zhang, J. Li, Q. Tian, Deep representation learning with part loss for person re-identification. arXiv, 2017..
[30] R. Yu, Z. Zhou, S. Bai, X. Bai, Divide and fuse: A re-ranking approach for person re-identification, in: BMVC, 2017..
[31] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: CVPR, 2016..
[32] H. Zhao, M. Tian, S., Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: CVPR, 2017a..
[33] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: ICCV, 2017b..
[34] R. Zhao, W. Ouyang, X. Wang, Person re-identification by saliency learning, TPAMI 39 (2017) 356–370.
[35] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: ICCV, 2015..
[36] L. Zheng, Y., Yang, A. Hauptmann, Person re-identification: Past, present and future. arXiv, 2016..
[37] W.S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: CVPR, 2011..
[38] Z, Zheng, L. Zheng, Y., Yang, Pedestrian alignment network for large-scale person re-identification. arXiv, 2017a..
[39] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: ICCV, 2017b..
[40] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: CVPR, 2017..
[41] Q. Zhou, H. Fan, S. Zheng, H. Su, X. Li, S. Wu, H. Ling, Graph correspondence transfer for person re-identification, in: AAAI, 2018..
[42] S. Zhou, J. Wang, J. Wang, Y. Gong, N. Zheng, Point to set similarity based deep feature learning for person re-identification, in: CVPR, 2017..
[43] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017..
[44] Dapeng Tao, Yanan Guo, Mingli Song, Yaotang Li, Zhengtao Yu, Yuan Yan Tang, Person re-identification by dual-regularized kiss metric learning, IEEE Trans. Image Processing 25 (6) (2016) 2726–2738.
[45] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, Xuelong Li, Person re-identification by regularized smoothing kiss metric learning, IEEE Trans. Circuits Systems Video Technol. 23 (10) (2013) 1675–1685.
[46] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), pages 480–496, 2018..
[47] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification, in: 2018 ACM Multimedia Conference on Multimedia Conference, 2018, pp. 274–282.
[48] Joshua B Tenenbaum, William T Freeman, Separating style and content with bilinear models, Neural Comput. 12 (6) (2000) 1247–1283.
[49] Joao Carreira, Rui Caseiro, Jorge Batista, Cristian Sminchisescu. Semantic segmentation with second-order pooling, in: European Conference on Computer Vision, pages 430–443. Springer, 2012..
[50] Yang Gao, Oscar Beijbom, Ning Zhang, Trevor Darrell. Compact bilinear pooling, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pages 317–326, 2016..
[51] Shu Kong, Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017..
[52] Tsung-Yu Lin, Aruni RoyChowdhury, Subhransu Maji. Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE international conference on computer vision, pages 1449–1457, 2015..
[53] Florent Perronnin, Jorge Sánchez, Thomas Mensink, Improving the fisher kernel for large-scale image classification, in: European conference on computer vision, Springer, 2010, pp. 143–156.
[54] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pages 2169–2178. IEEE, 2006..
[55] Xiaobin Chang, Timothy M. Hospedales, Tao Xiang. Multi-level factorisation net for person re-identification. CoRR, abs/1803.09132, 2018..
[56] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, Adversarially occluded samples for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5098–5107.
[57] Wei Li, Xiatian Zhu, Shaogang Gong, Harmonious attention network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294.
[58] Zhigang Chang, Qin Zhou, Mingyang Yu, Shibao Zheng, Hua Yang, Tai-Pang Wu:Distribution Context Aware Loss for Person Re-identification, VCIP 1–4 (2019).

Zhigang Chang is a Ph.D. candidate in Department of Electronic Engineering of Shanghai Jiao Tong University, under the supervision of Prof. Shibao Zheng. Before that, I graduated from South China University of Technology on Electronic Information Engineering in 2016. He is now a visiting student in MM Lab at Nanyang Technological University under the supervision of Associate Prof. Chen Change Loy. His research interests are computer vision tasks like pedestrian detection and person re-identification.

Qin Zhou received her B.S. degree in information engineering from Xi'an Jiao Tong University, Xi'an, China, in 2013. She is currently a Ph.D. student at the Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include computer vision, machine learning, convex optimization and person re-identification. She is now a visiting student in Professor Haibin Ling's lab at Temple University.

Hang Su, is an assistant professor working in the Department of Computer Science and Technology at Tsinghua University, working with Prof. Bo Zhang and Prof. Jun Zhu. His research interests lie in the development of computer vision and machine learning algorithms for solving scientific and engineering problems. His current work involves both the foundations of interpretable machine learning and the applications of image/video analysis. Before joining Tsinghua, he received my Ph. D. degree from Shanghai Jiao Tong University and worked as a visiting scholar at Carnegie Mellon University.

Heng Fan is a PhD student in Department of Computer & Information Sciences at Temple University starting from 2016. His advisor is Prof. Haibin Ling. Prior to Temple University, he spent two and a half years as a master student (course completed) in Huazhong Agricultural University, where he worked closely with Prof. Jinhai Xiang. He received my B.S. degree from Huazhong Agricultural University in 2013. He works on visual object tracking, and am also interested in semantic segmentation and object detection.

Hua Yang, is an assistant professor working in Department of Electronic Engineering of Shanghai Jiao Tong University. Her research interests include video coding and network transmission, computer vision, intelligent video surveillance systems and applications.

Shibao Zheng, received his B.S. degree in communication engineering from Xidian University, Xi'an and M.S. degree in the signal and information processing from the 54th institute of CETC, Shijiazhuang, China, in 1983 and 1986, respectively. He is currently a professor of electronic engineering department and vice director of Elderly Health Information and Technology Institute, Shanghai Jiao Tong University (SJTU), Shanghai, China. And he is also a professor committee member of Shanghai Key Laboratory of Digital Media Processing and Transmission, and a Consultant Expert of ministry of public security in video surveillance field. His current research interests include urban video surveillance system, intelligent video analysis, and elderly health technology, etc.

Haibin Ling is now a SUNY Empire Innovation Professor in the Department of Computer Science of Stony Brook University, USA. Before that he received B.S. and M.S. from Peking University in 1997 and 2000, respectively, and Ph.D. from University of Maryland in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia; from 2006 to 2007, he worked as a postdoctoral scientist at UCLA; and from 2008 to 2019, he was a faculty member of the Department of Computer Sciences for Temple University. He received Best Student Paper Award of ACM UIST in 2003 and NSF CAREER Award in 2014. He serves as associate editors for IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Pattern Recognition (PR), and Computer Vision and Image Understanding (CVIU). He has served as Area Chairs for CVPR (2014, 2016, 2019, 2020) and ECCV (2020).