Multi-level cross-view consistent feature learning for person re-identification

Liu, Y., Zhang, Y., Bhanu, B., Coleman, S., & Kerr, D. (2021). Multi-level cross-view consistent feature learning for person re-identification. *Neurocomputing*, 435, 1-14. <https://doi.org/10.1016/j.neucom.2021.01.010>

[Link to publication record in Ulster University Research Portal](#)

Published in:
Neurocomputing

Publication Status:
Published (in print/issue): 07/05/2021

DOI:
[10.1016/j.neucom.2021.01.010](https://doi.org/10.1016/j.neucom.2021.01.010)

Document Version
Author Accepted version

General rights

Copyright for the publications made accessible via Ulster University's Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk.

Manuscript Number: NEUCOM-D-20-01462R1

Title: Multi-level Cross-view Consistent Feature Learning for Person Re-identification

Article Type: Full Length Article (LS)

Keywords: person re-identification; local deep features; ID-based cross-view multi-level dictionary learning; cross-view consistent word learning; multi-level metric learning

Corresponding Author: Professor Yunzhou Zhang,

Corresponding Author's Institution:

First Author: Yixiu Liu

Order of Authors: Yixiu Liu; Yunzhou Zhang; Bir Bhanu; Sonya Coleman; Dermot Kerr

Abstract: Person re-identification plays an important role in searching for a specific person in a camera network with non-overlapping cameras. The most critical problem for re-identification is feature representation. In this paper, a multi-level cross-view consistent feature learning framework is proposed for person re-identification. First, local deep, LOMO and SIFT features are extracted to form multi-level features. Specifically, local features from the lower and higher layers of a convolutional neural network (CNN) are extracted, these features complement each other as they extract apparent and semantic properties. Second, an ID-based cross-view multi-level dictionary learning (IDB-CMDL) is carried out to obtain sparse and discriminant feature representation. Third, a cross-view consistent word learning is performed to get the cross-view consistent BoVW histograms from sparse feature representation. Finally, a multi-level metric learning fuses multiple BoVW histograms, and learns the sample distance in the subspace for ranking. Experiments on the public CUHK03, Market1501, and DukeMTMC-ReID datasets show results that are superior to many state-of-the-art methods for person re-identification.

Abstract: Person re-identification plays an important role in searching for a specific person in a camera network with nonoverlapping cameras. The most critical problem for re-identification is feature representation. In this paper, a multi-level cross-view consistent feature learning framework is proposed for person re-identification. First, local deep, LOMO and SIFT features are extracted to form multi-level features. Specifically, local features from the lower and higher layers of a convolutional neural network (CNN) are extracted, these features complement each other as they extract apparent and semantic properties. Second, an ID-based cross-view multi-level dictionary learning (IDB-CMDL) is carried out to obtain sparse and discriminant feature representation. Third, a crossview consistent word learning is performed to get the cross-view consistent BoVW histograms from sparse feature representation. Finally, a multi-level metric learning fuses multiple BoVW histograms, and learns the sample distance in the subspace for ranking. Experiments on the public CUHK03, Market1501, and DukeMTMC-ReID datasets show results that are superior to many state-of-the-art methods for person re-identification.

Multi-level Cross-view Consistent Feature Learning for Person Re-identification

Yixiu Liu^a, Yunzhou Zhang^{a,*}, Bir Bhanu^b, Sonya Coleman^c, and Dermot Kerr^c

^aNortheastern University, College of Information Science and Engineering, Shenyang, China, 110819

^bUniversity of California, Riverside, Riverside, CA 92521

^cIntelligent Systems Research Centre, University of Ulster, Londonderry, UK

Abstract

Person re-identification plays an important role in searching for a specific person in a camera network with non-overlapping cameras. The most critical problem for re-identification is feature representation. In this paper, a multi-level cross-view consistent feature learning framework is proposed for person re-identification. First, local deep, LOMO and SIFT features are extracted to form multi-level features. Specifically, local features from the lower and higher layers of a convolutional neural network (CNN) are extracted, these features complement each other as they extract apparent and semantic properties. Second, an ID-based cross-view multi-level dictionary learning (IDB-CMDL) is carried out to obtain sparse and discriminant feature representation. Third, a cross-view consistent word learning is performed to get the cross-view consistent BoVW histograms from sparse feature representation. Finally, a multi-level metric learning fuses multiple BoVW histograms, and learns the sample distance in the subspace for ranking. Experiments on the public CUHK03, Market1501, and DukeMTMC-ReID datasets show results that are superior to many state-of-the-art methods for person re-identification.

Keywords: Person re-identification, Local deep features, ID-based cross-view multi-level dictionary learning, Cross-view consistent word learning, Multi-level metric learning.

1. Introduction

Recognizing an individual across non-overlapping cameras, referred to as person re-identification, has gained much attention among researchers [1, 2, 3, 4, 5]. It is becoming an increasingly important technique for multimedia and computer vision communities because of its wide range of potential applications, especially in video content analysis. Examples include security cameras and video surveillance [6, 7], video summarization [8], and so on. Although with great application prospects, person re-identification is still confronted with significant challenges in real-world scenarios due to large variations of same person but subtle variations of the different persons under changing illuminations, poses and viewpoints among different cameras. The persons who have the same identities are difficult to be distinguished without frontal face verification. While it is easy to misjudge different persons to be the same one because of the similar dress. In this paper, we solve this key issue by the proposed feature learning method that minimizes the difference of feature distribution of the same person in different camera views. The differences between the proposed method and common feature extraction methods are shown in Fig. 1.

Recently, deep learning has been widely used in almost all visual applications. Many studies [9, 10, 11] have shown that fine-tuning a deep model pre-trained on a large-scale dataset (*e.g.*, ImageNet [12]) is effective for other related domains and tasks. Many methods of feature representation for person re-identification based on Convolutional Neural Networks (CNNs) are emerging such as [13, 14] that are designed to obtain more robust deep features through altering the architecture of CNNs. What is more, many feature fusion strategies based on CNNs have been developed, such as [15, 16, 17], in which deep features are combined tightly with some low-level or mid-level attributes of pedestrian images. The fundamental difference between the proposed methods and those CNN related methods is that the latter only contains two parts: feature extraction and classification, while we have added a feature learning process externally. Besides, our work has mined the potential of both the high-level and low-level local features of CNN and we attempt find the optimal combination of apparent and semantic properties.

Although deep learning methods have achieved great success in improving the accuracy of person re-identification, it doesn't mean that hand-crafted features do not deserve any merit. More and more discriminative hand-crafted features have been developed to achieve accurate matching. Examples are: Local Maximal Occurrence Representation (LOMO) [18], Symmetry-Driven Accumulation of Local Features (SDALF) [19], local SIFT features [20], gabor Biologically inspired Covariance descriptors (gBiCov) [21], etc. In real scenarios, the available datasets for person re-identification are usually insufficient because of the difficulties in data acquisition and labeling for ground-truth, which lead to the weak generalization of the trained model. By taking

*Corresponding author

Email address: zhangyunzhou@mail.neu.edu.cn (Yunzhou Zhang)

Figure 1: The proposed method focuses on dealing with the misjudgement of the same pedestrians with dissimilar appearances caused by the change of camera views and the misjudgement of different pedestrians with the same view caused by similar appearance. The common feature extraction methods roughly classify the samples with high feature similarity, but ignore the two special cases mentioned above. As compared to these methods, our method can achieve smaller intraclass differences and larger interclass differences.

this fact into account, LOMO features and local SIFT features are used as the complement of deep features to improve the performance of person re-identification in this paper.

Our research aims to further process the constructed multi-level features to make them more applicable for pedestrian re-identification tasks. Dictionary learning is usually adopted to obtain sparse and discriminative feature representation, such as K-SVD [22], discriminative K-SVD [23], semi-supervised coupled dictionary learning (SSCDL) [24], and the cross-view multi-level dictionary learning (CMDL) [25]. We propose ID-based cross-view multi-level dictionary learning (IDB-CMDL). Unlike CMDL [25], which only maintains the view-consistency of the whole feature space, our work also ensures the view-consistency of the features of the same ID/label from different camera views.

In addition to considering the consistency of the features, we also consider the view-consistency of feature distribution. In order to simplify the statistics of feature distributions to minimize distribution differences, we propose cross-view consistent word learning based on Bag of visual words (BoVW) model. BoVW model was first applied to person re-identification in [26] as BoVW+color names (CN). The framework that we propose is very different from [26]. In this paper, multi-level local features are extracted to form the visual “words” of BoVW model, instead of color names only. An important finding is that the obtained BoVW histograms have Gaussian distribution, which inspired us to propose a multi-level metric learning. The multi-level BoVW histograms are normalized to equal mean Gaussian distributions for fusion, and then metric learning is performed.

In summary, this paper make the following contributions for person re-identification:

- A feature learning process outside of the traditional recognition model (feature extraction + classification) has been added that aims to pull closer to the same pedestrian in different camera views and push away different pedestrians with similar dresses.
- An ID-based cross-view multi-level dictionary learning is proposed that maintains the view-consistency of the features of the same ID/label from different camera views.
- A cross-view consistent word learning is proposed that minimizes the feature distribution differences and maintains the view-consistency of feature distribution.
- A multi-level metric learning is proposed to normalize and fuse BoVW histograms to perform metric learning.

2. Related work

The proposed method mainly covers four technical areas: hand-crafted methods, deep learning, dictionary learning and feature learning methods, which we describe in the following.

2.1. Hand-crafted methods for person re-identification

The early work in person re-identification focused on visual feature representation, and distance metric learning. In [27], Bazzani *et al.* designed a Symmetry-Driven Accumulation of Local Features (SDALF) descriptor which was invariant to pose, scale and illumination. Ma *et al.* [28] proposed a new descriptor building on the recent advances of Fisher Vectors, consisting of the pixel coordinates, intensity as well as the first and second order derivatives of each pixel. LOMO feature was proposed by Liao *et al.* in [18] to make a stable representation against viewpoint changes. It analyzes the horizontal occurrence of local features, and maximizes the occurrence. Zhao *et al.* [28] used dense SIFT descriptors as complementary to color histograms to handle

viewpoint and illumination changes. Zheng *et al.* [26] viewed person re-identification as a special task of image search, and proposed an unsupervised BoVW model to quantize and pool the color names descriptor. In distance metric learning methods for person re-identification, Engel *et al.* [29] considered person re-identification as a ranking problem by using RankSVM to learn similarity parameters. A metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) was proposed by Liao *et al.* [18]. A XQDA metric was learned in the derived discriminant low dimensional subspace. Shen *et al.* in [30] introduced an approach based on boosting to learn a correspondence structure which indicated the patch-wise matching probabilities between images from a target pair. In addition, KISSME [31] and EIML [32] also were effective metric learning methods.

2.2. Deep learning-based methods for person re-identification

Recently the state-of-the-art methods in person re-identification have been dominated by deep learning. The advantage of CNN is that it can not only optimize the feature representation, but also learn the discriminative metrics. Xiao *et al.* [14] introduced an algorithm based on Domain Guided Dropout to improve the feature learning procedure. This idea was based on the observation that some neurons learned representations are shared across several domains, while some others are effective only for a specific one when training a CNN with data from all the domains. In order to search the target person from a gallery of whole scene images, Xiao *et al.* proposed an Online Instance Matching (OIM) loss function to train the network effectively in [33]. Liu *et al.* [34] designed the end-to-end Comparative Attention Network (CAN) to selectively focus on parts of pairs adaptively comparing their appearance. A deep self-paced learning (DSPL) algorithm was proposed by Zhou *et al.* [35] to alleviate noisy background clutter and mutual occlusion. Zhu *et al.* in [36] proposed a Part-based Deep Hashing (PDH) by integrating deep learning and hashing into one framework to perform real-time search in a large gallery. In [37], an effective learning-to-rank algorithm was proposed by Chen *et al.* to minimize the cost corresponding to the ranking disorders. The ranking model was solved with a CNN that built the relation between input image pairs and their similarity scores.

2.3. Dictionary learning

Dictionary learning aims to learn a feature subspace which differentiate pedestrians from different IDs easier. In [24], two coupled dictionaries were learned from both labeled and unlabeled images of gallery and probe cameras in the training phase, then the features of pedestrian images from the probe camera were encoded by the sparse representation. After the above transformation, the features of the same person from different cameras became more similar than before. Zhu *et al.* [38] proposed a semi-supervised cross-view projection-based dictionary learning (SCPDL) method to achieve video-based person re-identification. With the transformation, the influence of variations within each video to the re-identification was reduced. In [39], a dictionary learning was embedded into metric learning, through which robust feature representation learning and discriminative metric learning were formulated into a unified framework. Li *et al.* [25] proposed a Cross-view Multi-level Dictionary Learning (CMDL) model at three different representation levels to solve the multi-view learning problem. The new strategy proposed was more efficient than the traditional dictionary learning.

2.4. Feature learning methods for person re-identification

Recently, many feature learning methods based on CNNs have been emerging, and the proposed method is exactly one of them. Chen *et al.* [40] proposed a Deep Pyramid Feature Learning (DPFL) model to jointly learn discriminative scale-specific features, so that the complementary advantages across scales could be mined. In [41], Yu *et al.* presented a divide and fusion strategy for person re-identification re-ranking. High-dimensional sub-features of different parts were extracted, from which context information was iteratively encoded into feature vector for further fusion and re-ranking. Similar to [40] and [41], Fu *et al.* [42] proposed a Horizontal Pyramid Matching (HPM) to capture partial features at different horizontal pyramid scales. Li *et al.* [43] presented a Multi-Scale Context-Aware Network (MSCAN) to capture powerful features over full body and body parts to release some difficulties in pose variations and background clutters. A braidNet was proposed by Wang *et al.* [44] to extract the comparison features of two input images by the designed WConv layers. In [45], Feng *et al.* extracted view-related features with a cross-view Euclidean constraint (CV-EC) and a cross-view center loss (CVCL) to narrow the gaps between the features from changing viewpoints. Similar to [45], the Visibility-aware Part Model (VPM) proposed by Sun *et al.* compared two images by focusing on their shared regions. We have compared the above methods with the proposed methods in [4.3] separately. Here, it is necessary to distinguish few similar re-id works, such as similarity consistency learning [46] and deep association learning (DAL) [47], from ours. Wu *et al.* proposed a camera-aware similarity consistency loss to learn consistent pairwise similarity distributions for intra-camera matching and cross-camera matching. While Chen *et al.* formulated a deep association learning scheme that constrained the association of each frame to the best-matched intra-camera representation and cross-camera representation. Unlike [46] and [47] which use the consistency of the similarity distributions or the consistency of the ranking associations to cluster for unsupervised person re-identification, the proposed feature learning method focus on the deep processing of existing features to formulate the view-consistent feature representations.

Based on the thorough analysis of the related work as described above, we have developed an approach with the following key characteristics. (1) We construct multi-level features by using the hand-crafted features (LOMO [18], SIFT [28]) and deep features. (2) We formulate the problem of maintaining the view-consistency

by adding several optimization objective terms to CMDL [25], we call the resulting method as ID based CMDL (IDB-CMDL). (3) Inspired by distributions of the obtained BoVW histograms [26], we propose a cross-view consistent word learning to maintain the view-consistency of feature distribution. (4) We propose a multi-level metric learning that normalizes and fuses the obtained BoVW histograms before metric learning. (5) A comprehensive set of experiments are carried out on various datasets and comparisons with state-of-the-art techniques are provided which show improved results by the proposed technique.

3. Technical Approach

Unlike most feature-based methods that focus on constructing view-consistent hand-crafted features or learning deep features, our method focuses on the subsequent processing of extracted features so that they can complement each other and be more discriminative. The proposed feature learning method consists of ID-based cross-view multi-level dictionary learning, cross-view consistent word learning, and multi-level metric learning, as shown in Fig. 2. These three parts will be detailed in section 3.1, 3.2 and 3.3 respectively.

Figure 2: The pipeline of the proposed feature learning method.

The complete person re-identification process (including local feature extraction omitted in Fig. 2) is shown in detail in Fig. 3. The input images are from different camera views. Local features (CNN, LOMO, SIFT) are extracted from input images. Then IDB-CMDL is performed to obtain multi-level sparse feature representation. After that, cross-view consistent word learning is performed to get the multi-level BoVW histograms by minimizing the difference in feature distribution. Followed by multi-level metric learning, we normalize and fuse the obtained BoVW histograms to perform metric learning. The weights of fusion are γ_1 and γ_2 .

Figure 3: The architecture of multi-level cross-view consistent feature learning proposed in this paper.

3.1. ID-based cross-view multi-level dictionary learning

The first step of feature learning method is to learn the most discriminative information from the extracted multi-level features, that is, ID-based cross-view multi-level dictionary learning. In this subsection, we will explain where IDB-CMDL comes from 3.1.1, how it works 3.1.2, and why it makes sense 3.1.3. It should be noted that feature extraction is based on existing works rather than our innovations, so we will introduce it as the prelude in the Experiments (4.1).

3.1.1. Preliminaries

Since IDB-CMDL is developed from existing algorithms, clarifying the inheritance relationship of related dictionary learning methods is necessary to explain the rationality of the proposed IDB-CMDL algorithm. Dictionary learning has been widely applied in many visual tasks for the reason that it aims to learn expressive feature representations. The initial dictionary learning is as follows:

$$X = DZ, \quad (1)$$

where $X \in \mathcal{R}^{d \times n}$ denotes a set of samples, $D \in \mathcal{R}^{d \times m}$ is the dictionary that is learned, and $Z \in \mathcal{R}^{m \times n}$ is the sparse coefficient. The representative approach, such as K-SVD [22], solves this problem by solving the following optimization problem to find the D that makes X as sparse as possible.

$$\min_{D, Z} \|X - DZ\|_1 \quad s.t. \quad \forall i, \|z_i\|_0 \leq T_0, \quad (2)$$

where $\|\cdot\|_0$ is the L0 norm, $\|\cdot\|_1$ is the L1 norm, and T_0 is a constant that controls the degree of sparsity. However, existing methods based on Eq. (1) still have two drawbacks, which make them can not be well applied

to some visual tasks. First, the expensive computational cost of the L1 norm in solving the sparse coefficient Z limits its application in large-scale visual tasks. Second, traditional dictionary learning methods mainly focus on single-view tasks, therefore they cannot directly handle the multi-view visual data. However, person re-identification is exactly a multi-view visual task, and the amount of data collected from multiple views is very large in some scenarios, such as Market1501 dataset. CDL [25] algorithm is proposed to tackle the above problems by reformulating the approximation in Eq. (1) as a linear encoding and reconstruction process in different views. Therefore, Eq. (1) is reformulated as $X \approx DPX$. PX denotes the linear encodings of sample set X .

Let $X_{(1)} \in R^{d \times n}$ and $X_{(2)} \in R^{d \times n}$ denote two training sets of two camera views. Then the process of dictionary learning can be expressed as

$$\begin{aligned} X_{(1)} &= D_{(1)}P_{(1)}X_{(1)} \\ X_{(2)} &= D_{(2)}P_{(2)}X_{(2)} \end{aligned} \quad (3)$$

Here, $D_{(1)}$ and $D_{(2)}$ represent the dictionaries of two different camera views, while $P_{(1)}$ and $P_{(2)}$ denote the corresponding linear projections in two camera views.

The function of CDL model can be presented as

$$\begin{aligned} \min_{D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)}} \quad & \|X_{(1)} - D_{(1)}P_{(1)}X_{(1)}\|_F^2 + \|X_{(2)} - D_{(2)}P_{(2)}X_{(2)}\|_F^2 + \lambda f(D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)}) \\ \text{s.t.} \quad & \|d_{1(:,i)}\| \leq 1, \|d_{2(:,i)}\| \leq 1, i = 1, \dots, m, \end{aligned} \quad (4)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $f(\cdot)$ denotes the regularization function, λ is a trade-off parameter, and $d_{1(:,i)}$ and $d_{2(:,i)}$ are the i -th columns in $D_{(1)}$ and $D_{(2)}$, respectively.

In order to apply the CDL model to person re-identification, the solution CDL is extended to CMDL [25] by customizing regularization functions at feature level l . Eq. (4) is rewrite as following,

$$\begin{aligned} \min_{D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}} \quad & \|X_{(1,l)} - D_{(1,l)}P_{(1,l)}X_{(1,l)}\|_F^2 + \|X_{(2,l)} - D_{(2,l)}P_{(2,l)}X_{(2,l)}\|_F^2 + \lambda_l f_l(D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}) \\ \text{s.t.} \quad & \|d_{1l(:,i)}\| \leq 1, \|d_{2l(:,i)}\| \leq 1, i = 1, \dots, m \end{aligned} \quad (5)$$

3.1.2. Model

In this paper, each image is characterized by multiple feature vectors, which means the local features are obtained with ID tags. It inspires us to extend CMDL to IDB-CMDL by minimizing the differences of features from the same person at different level l .

$$\begin{aligned} \min_{D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}} \quad & \|X_{(1,l)} - D_{(1,l)}P_{(1,l)}X_{(1,l)}\|_F^2 + \|X_{(2,l)} - D_{(2,l)}P_{(2,l)}X_{(2,l)}\|_F^2 + \|X_{(1,l,c)} - D_{(1,l)}P_{(1,l)}X_{(1,l,c)}\|_F^2 \\ & + \|X_{(2,l,c)} - D_{(2,l)}P_{(2,l)}X_{(2,l,c)}\|_F^2 + \lambda_l f_l(D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}) \\ \text{s.t.} \quad & \|d_{1l(:,i)}\| \leq 1, \|d_{2l(:,i)}\| \leq 1, i = 1, \dots, m, \end{aligned} \quad (6)$$

where the samples $X_{(1,l)}$ and $X_{(2,l)}$ are divided into $X_{(1,l,c)}$ and $X_{(2,l,c)}$, $c \in \{1, \dots, n\}$, respectively. n represents the number of identities under different two cameras in the same dataset. In order to minimize intra-class difference, we enhance the consistency of the learned sparse representations for same pedestrians across camera views. The regularization function $f_l(\cdot)$ in Eq. (6) is defined as $\|P_{(1,l)}X_{(1,l,c)} - P_{(2,l)}X_{(2,l,c)}\|_F^2$, then we have the following objective

$$\begin{aligned} \min_{D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}} \quad & \|X_{(1,l)} - D_{(1,l)}P_{(1,l)}X_{(1,l)}\|_F^2 + \|X_{(2,l)} - D_{(2,l)}P_{(2,l)}X_{(2,l)}\|_F^2 + \|X_{(1,l,c)} - D_{(1,l)}P_{(1,l)}X_{(1,l,c)}\|_F^2 \\ & + \|X_{(2,l,c)} - D_{(2,l)}P_{(2,l)}X_{(2,l,c)}\|_F^2 + \lambda_l (\|P_{(1,l)}X_{(1,l,c)} - P_{(2,l)}X_{(2,l,c)}\|_F^2) \\ \text{s.t.} \quad & \|d_{1l(:,i)}\| \leq 1, \|d_{2l(:,i)}\| \leq 1, i = 1, \dots, m \end{aligned} \quad (7)$$

Considering the computational efficiency, the relaxation variables $A_{(1,l)}$, $A_{(2,l)}$, $A_{(1,l,c)}$, $A_{(2,l,c)}$ are introduced, and the objective is rewritten as

$$\begin{aligned} \min_{D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}} \quad & \|X_{(1,l)} - D_{(1,l)}A_{(1,l)}\|_F^2 + \|X_{(2,l)} - D_{(2,l)}A_{(2,l)}\|_F^2 + \|X_{(1,l,c)} - D_{(1,l)}A_{(1,l,c)}\|_F^2 \\ & + \|X_{(2,l,c)} - D_{(2,l)}A_{(2,l,c)}\|_F^2 + \alpha (\|P_{(1,l)}X_{(1,l)} - A_{(1,l)}\|_F^2 + \|P_{(2,l)}X_{(2,l)} - A_{(2,l)}\|_F^2) \\ & + \beta (\|P_{(1,l)}X_{(1,l,c)} - A_{(1,l,c)}\|_F^2 + \|P_{(2,l)}X_{(2,l,c)} - A_{(2,l,c)}\|_F^2) + \lambda_l \|A_{(1,l,c)} - A_{(2,l,c)}\|_F^2 \\ \text{s.t.} \quad & \|d_{1l(:,i)}\| \leq 1, \|d_{2l(:,i)}\| \leq 1, i = 1, \dots, m \end{aligned} \quad (8)$$

The solution details of the model are shown in the **Appendix**. It should be noted that the above theoretical derivation only takes two-view case as an example. In fact, it is also applicable to multiple view scenarios. For instance, we learned six dictionaries $X_{(1,l)} \sim X_{(6,l)}$ and coefficient matrices $P_{(1,l)} \sim P_{(6,l)}$ from the Market1501 dataset, which contains 6 disjoint camera views. In particular, there were $C(6, 2) = 15$ cross-view terms like $\|P_{(1,l)}X_{(1,l,c)} - P_{(2,l)}X_{(2,l,c)}\|_F^2$ in Eq. (7) during optimization process.

3.1.3. The innovations compared with previous work

We analyzed the improvements of IDB-CMDL over CMDL in principle. For feature level l , CMDL focuses on minimizing reconstruction errors after sparse expression in each camera view, i.e., $\min d(X_{(1,l)}, D_{(1,l)} P_{(1,l)} X_{(1,l)})$ and $\min d(X_{(2,l)}, D_{(2,l)} P_{(2,l)} X_{(2,l)})$. CMDL takes advantages of the view-consistency information, and adaptively learn pairs of dictionaries to generate robust and compact representations for pedestrian images. Although it solves the cross-view problem to a certain extent, it only focuses on minimizing reconstruction errors from different views so that the learned coefficient matrix approximates the original feature matrix as much as possible. On this basis, the IDB-CMDL we proposed can not only maximize the inter-class difference, but also minimize the intra-class difference, making the learned coefficient matrix more discriminative for multi-view visual tasks.

Compared with the previous work CMDL, the most significant innovation of the proposed IDB-CMDL is that we maximize the inter-class difference by class-by-class reconstruction, and minimize the intra-class difference by enhancing cross-camera invariants. The cross-camera invariant refers to the inherent consistency information of the same pedestrian across camera views. (1). The person re-identification dataset is divided into many subsets. Specifically, features from two different camera views are divided into $X_{(1,l,1)} - X_{(1,l,n)}$ and $X_{(2,l,1)} - X_{(2,l,n)}$, respectively. Then we perform class-by-class reconstruction, that is, $\min d(X_{(1,l,c)}, D_{(1,l)} P_{(1,l)} X_{(1,l,c)})$ and $\min d(X_{(2,l,c)}, D_{(2,l)} P_{(2,l)} X_{(2,l,c)})$. In this way, the learned $D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}$ can clearly separate different classes, so as to maximize the inter-class difference. (2). Considering that the learned sparse representation of the same pedestrian in different views should have smaller difference compared to different pedestrians, we minimize intra-class difference by enhancing cross-camera invariants, that is $\min d(P_{(1,l)} X_{(1,l,c)}, P_{(2,l)} X_{(2,l,c)})$.

3.2. Cross-view consistent word learning

After IDB-CMDL, we find that there is still a huge computational cost to perform metric learning directly on the obtained sparse representations. Therefore, we propose a cross-view consistent word learning to improve computational efficiency by encoding sparse feature representations into BoVW histograms. As we know, BoVW model is one of the effective and efficient method to search images in massive data due to its insensitivity to visual angle. Fig. 4 shows how it works in our approach. Local features usually shift with the change of camera views, such as the green satchel in Fig. 4. It is likely to result in mismatch when directly classifying the concatenated local features. So we get the codebooks by K-means, and then get the BoVW histogram which only cares about the statistical distribution of local features in the whole image, not the spatial location.

Figure 4: BoVW solves the misalignment of the same person in different camera views.

Classical BoVW models utilize only SIFT descriptor to build visual words, and other important features are ignored. In this paper, three different local features are used as visual words after IDB-CMDL. In this section, we define that $Z_{(1,l,c)} = P_{(1,l)} X_{(1,l,c)}$, $Z_{(2,l,c)} = P_{(2,l)} X_{(2,l,c)}$. $z_{(1,l,c)}^i$ and $z_{(2,l,c)}^i$ are the i -th row of $Z_{(1,l,c)}$ and $Z_{(2,l,c)}$, respectively. K-means is used to divide $z_{(1,l,c)}^i$ into k clusters, so that each local feature of the same person is assigned to one of them by minimizing the non-negative cost function below.

$$\min \sum_{i=1}^{np} (\|z_{(1,l,c)}^i - c_{(1,l,c)}^j\|_2^2 \quad j = 1, \dots, k), \quad (9)$$

where $c_{(1,l,c)}^j$ is the clustering center corresponding to $Z_{(1,l,c)}$. np is the total number of the patches with the same label. The objective of k-means is to minimize the sum of squares of Euclidean distance between each local feature $z_{(1,l,c)}^i$ and its nearest cluster $c_{(1,l,c)}^j$. Similarly, we obtain the clusters corresponding to $Z_{(2,l,c)}$. After that, we obtain three codebooks by which each image can be represented linearly.

Different from the traditional BoVW methods, we try to maintain the consistency of word distribution across views while clustering, because the feature distribution (learned from Eq. (7)) across views should be basically the same. So a cross-view consistent word learning is proposed to make the word distribution of the same person across views as similar as possible. We introduce a cross-view multi-level consistent word learning loss:

$$L_{CMWL} = \sum_{v,l} d(\mathcal{P}(w_{v,l}) - \mathcal{P}(w_l)), \quad (10)$$

where $w_{v,l}$ is the v -th camera view l -th feature level word vector, $\mathcal{P}(w_l)$ is the l -th feature level word distribution of BoVW histogram, $\mathcal{P}(w_{v,l})$ is the v -th camera view l -th feature level word distribution of BoVW histogram, and $d(\cdot)$ is the distance between these two distribution.

We observe that $\mathcal{P}(w_{v,l})$ follows a normal distribution. In this work, we use the simplified 2-Wasserstein distance [48] to represent the cross-view multi-level consistent word learning loss:

$$L_{CMWL} = \sum_{v,l} \|\mu_{v,l} - \mu_l\|_2^2 + \|\sigma_{v,l} - \sigma_l\|_2^2, \quad (11)$$

where $\mu_{v,l}/\sigma_{v,l}$ is the mean/std vector of the v -th camera view l -th feature level word distribution, and μ_l/σ_l is the mean/std vector of the l -th feature level word distribution. The cross-view multi-level consistent word learning loss Eq. (11) is jointly optimized along with the objective described in Eq. (9) as an extra constraint to get the BoVW histograms with consistent distribution across views.

3.3. Multi-level metric learning

So far, we get the multi-level word vectors that reflect feature distribution. These word vectors with different attributes need to be fused, and finally classified by measuring similarity. Considering that the similarity between word vectors is different from the linear distance in geometric space, the fused word vectors need to be mapped to r dimensional subspaces $\Psi \in R^{d \times r}$, where the Euclidean distance can work better for classification or ranking, that is, metric learning. Compared with the previous metric learning methods, the proposed KL-XQDA in this paper is unusual in that it fits the following two observations: (1) the multi-level words follow Gaussian distribution, which can meet the precondition of XQDA (mean = 0, Gaussian distribution) after standardization and fusion. (2) the BoVW histogram reflects the distribution of features, which inspires us to use KL (Kullback-Leibler) divergence to compute the distribution difference of features in different views and embed it into XQDA.

First, the word vector $w_{v,l}$ is standardized as follows

$$w_{v,l} = \frac{w_{v,l} - \mu_{v,l}}{\sigma_{v,l}} \quad (12)$$

Then three levels of word vectors are fused into a word vector through weighted concatenation, which is expressed as

$$w = (w_{deep}, \gamma_1 w_{LOMO}, \gamma_2 w_{SIFT}), \quad (13)$$

where the fusion weight γ_1, γ_2 are the adaptive parameters, and the operation $(\cdot, \cdot)$ denotes vector concatenation.

Before computing KL divergence, we need to normalize w and map its elements between 0 and 1. In this way, each element of w represents the probability of the corresponding local feature. w is linearly normalized as

$$w = \frac{w - w_{\min}}{w_{\max} - w_{\min}} \quad (14)$$

Then we compute feature distribution difference $D(w_i, w_j)$ of two sample word vectors w_i and w_j from different views by KL divergence. Considering the asymmetry of KL divergence ($KL(w_i||w_j) \neq KL(w_j||w_i)$), $D(w_i, w_j)$ can be expressed as

$$\begin{aligned} D(w_i, w_j) &= KL(w_i, w_j) + KL(w_j, w_i) \\ &= \sum_{k=1}^L w_i^k \ln \frac{w_i^k}{w_j^k} + \sum_{k=1}^L w_j^k \ln \frac{w_j^k}{w_i^k} \\ &= (w_i \ln \frac{w_i}{w_j})^T (w_i \ln \frac{w_i}{w_j}) + (w_j \ln \frac{w_j}{w_i})^T (w_j \ln \frac{w_j}{w_i}), \end{aligned} \quad (15)$$

where w_i^k and w_j^k are the elements of the word vectors w_i and w_j , and L is the length of word vector.

Finally, we embed $D(w_i, w_j)$ into XQDA to learn a r dimensional subspace $\Psi \in R^{d \times r}$ where feature distribution difference is much reduced. The distance function between w_i and w_j can be expressed as

$$d(w_i, w_j) = (\sum_I^{-1} - \sum_E^{-1}) D(w_i, w_j), \quad (16)$$

where $\sum_I$ and $\sum_E$ denote the covariance matrices of intrapersonal variations Ω_I and extrapersonal variations Ω_E , respectively. Given the subspace $\Psi = (\varphi_1, \varphi_2, \dots, \varphi_r)$, this distance function Eq. (16) can be expressed as

$$d(w_i, w_j) = \varphi (\sum_I^{-1} - \sum_E^{-1}) \varphi^T D(w_i, w_j) \quad (17)$$

The samples of the two classes will still have the same mean but different variances in Ψ . In this case, the ratio of $\sigma_I(\varphi)$ and $\sigma_E(\varphi)$ can be used to increase the distance of the variance, so as to improve the discrimination of different samples. Notice that $\sigma_I(\varphi) = \varphi^T \sum_I \varphi$ and $\sigma_E(\varphi) = \varphi^T \sum_E \varphi$, therefore, φ can be optimized by maximizing the ratio function in Eq. (18):

$$r(\varphi) = \frac{\varphi^T \sum_E \varphi}{\varphi^T \sum_I \varphi} \quad (18)$$

The maximization of $r(\varphi)$ is equivalent to

$$\min_{\varphi} \varphi^T \sum_E \varphi, \quad s.t. \quad \varphi^T \sum_I \varphi = 1, \quad (19)$$

with φ , we can learn a discriminant subspace Ψ and a distance function $d(w_i, w_j)$, as defined in Eq. (17).

4. Experiments

4.1. Feature representation

4.1.1. Local deep features

Unlike many deep learning methods that use only the last fully connected layer as the feature, we use a combination of lower layer and higher layer to construct deep features. The idea is based on an observation that the lower feature maps in the network reflect the apparent attributes, while the higher feature maps represent the semantic properties, and these two sets of features can be complementary.

The VGG16 network has a total of 13 convolution layers (Conv1~Conv13). We roughly divide them into low layers (Conv1~Conv4) and high layers (Conv9~Conv13), and select one layer from each to extract features and complement each other by concatenation. Fig. 5 shows the process of extracting the deep local features. Each feature maps is gridded as non-overlapping patches, each of which is expanded into a one-dimensional feature vector. For each lower feature map of size 112×112 , the patch size is set to 16×16 . For each higher feature map of size 28×28 , the patch size is set to 4×4 . After L2 normalization, the patches of feature maps from the same spatial position of lower and higher feature maps, respectively, are concatenated into a $16 \times 16 + 4 \times 4 = 272$ dimensional local feature vector.

Figure 5: The schematic for feature extraction of local deep features.

We get the optimal layers $\text{Conv}n_1$ and $\text{Conv}n_2$ by minimizing the object function of the proposed feature learning method.

$$n_1, n_2 = \arg \min_{i,j} f(x_{deep}^i, x_{deep}^j, \gamma_1 x_{LOMO}, \gamma_2 x_{SIFT}), \quad (20)$$

where n_1 and n_2 are 3 and 10 in this paper, and $f(\cdot)$ represents the proposed multi-level cross-view consistent feature learning. The comparison experiments of n_1 and n_2 have been given in Fig. 7 of 4.4.

4.1.2. LOMO

In this paper, LOMO features [18] are extracted. However, different from the original, instead of utilizing only the patches that represent horizontal maximized occurrence in the same rows, all the patches are used to obtain the local features. In addition, downsampling is removed. In the end, an image of 160×60 is divided into $10 \times 3 = 30$ patches, each of which has $8 \times 8 \times 8$ color bins $+ 3^4 \times 2$ SILTP bins = 1186 dimensions.

4.1.3. SIFT

In this work, each image of size 160×60 is divided into 16×6 of overlapping local patches. Then the SIFT features are extracted from these patches. Each patch is divided into 4×4 cells, and the orientations of local gradients are quantized into 8 bins. In the end, for each patch, we obtain a $4 \times 4 \times 8 = 128$ dimensional SIFT feature vector, which is L2 normalized.

4.2. Datasets and implementation details

For evaluation, experiments were performed on three large-scale person re-id benchmarks, CUHK03 [49], Market1501 [26], and DukeMTMC-ReID [50]. Fig. 6 shows several samples of three datasets. We followed the standard person re-identification training/test split and test protocol, as shown in Table 1. We use the mean Average Precision (mAP) and cumulative matching characteristic (CMC) to evaluate the performance.

Table 1: Evaluation protocol for Person re-id on three datasets.

Datasets	#Images	#Identities	#Train	#Test
CUHK03	14,097	1,467	767	700
Market1501	32,668	1,501	751	750
DukeMTMC-ReID	36,411	1,402	702	702

We fine tune VGG16 network in the Tensorflow framework to obtain the local deep features. All the input images are resized to 224×224 . We set the batch size to 32 and epoch to 65. For IDB-CMDL, $D_{(1,l)}$, $D_{(2,l)}$, $P_{(1,l)}$

Figure 6: Examples from three datasets. Each column corresponds to the same person.

and $P_{(2,l)}$ are initialized to random matrixes ($l = 1, 2, 3$). The trade-off parameters $\lambda_l = 0.45, \lambda_2 = 0.35, \lambda_3 = 0.2$. The relaxation variables $\alpha_1 = 0.2, \alpha_1 = 0.3, \beta_1 = 0.15, \beta_2 = 0.2$. When we get the BoVW histograms from the local features by K-means algorithm (see Fig. 3), we classify the local features of the same identity into the same class. In other words, k is equal to the number of identities in the training set. Take the CUHK03 dataset as an example, $k = 767$ (see Table 1). For codebook fusion, the adaptive parameters $\gamma_1 = 0.3$ and $\gamma_2 = 0.2$. The feature learning method proposed in this paper is easy for multi-thread batch computing, which greatly improves the execution efficiency.

4.3. Comparisons with the state-of-the-art algorithms

4.3.1. Evaluation on CUHK03

We evaluated the proposed method on the CUHK03 dataset with both manually labelled and auto-detected cases. We adopted a more realistic setting in which the identities were divided into 767/700 for training/testing, instead of the previous 1367/100. In this way, the training set is smaller, with only 7,300 images (versus 12,936/16,522 in Market1501/DukeMTMC-ReID). Table 2 shows that in the labelled setting, our method is superior to the compared state-of-the-art algorithms, surpassing the 2nd best CRAN by 1.04% (69.24 - 68.2) in mAP and 1.67% (74.37 - 72.7) in Rank-1. For the detected setting, our method achieved almost the same result as the best CASN, far more than the 3rd best CRAN by 2.75% (67.65 - 64.9)/2.23% (72.13 - 69.9) in mAP/Rank-1. In addition, we compared our method with some feature learning methods that were listed separately in Table 2, like HPM, DPFL and DaF. And our method outperforms them by at least 10% in both mAP and Rank-1 (labelled setting). We also compared the methods associated with our descriptor, BoW+XQDA and LOMO+XQDA. The results show that our method outperforms them by at least 30% in both mAP and Rank-1, regardless of the setting. It shows that to some extent IDB-CMDL and feature fusion greatly improve the discrimination of the descriptor for person re-identification.

Table 2: Results on the CUHK03 dataset. Metric: mAP and Rank-1 (%). Our method using CUHK03 as the labelled and detected. 1st/2nd/3rd best in red/blue/green.

Method	References	Labelled		detected	
		mAP	Rank-1	mAP	Rank-1
CRAN [51]	TCSVT 2019	68.2	72.7	64.9	69.9
CASN [52]	CVPR 2019	64.4	71.5	68.0	73.7
Pose-Transfer [53]	CVPR 2018	42.0	45.1	38.7	41.6
SVDNet [54]	ICCV 2017	37.83	40.93	37.30	41.50
IDE-R+XQDA [55]	CVPR 2017	29.6	32.0	28.2	31.1
HPM [42]	AAAI 2019	57.5	63.9	-	-
DPFL [40]	ICCV Workshop 2017	40.5	43.0	37.0	40.7
DaF [41]	BMVC 2017	31.5	27.5	30.0	26.4
BoW+XQDA [26]	BMVC 2016	7.3	7.9	6.4	6.4
LOMO+XQDA [18]	CVPR 2015	13.6	14.8	11.5	12.8
Ours		69.24	74.37	67.65	72.13

4.3.2. Evaluation on Market1501

We evaluated the proposed method against 13 state-of-the-art methods on the Market1501 dataset. Market1501 is a large-scale dataset with enough images, making it particularly suitable for data-driven deep learning methods. Table 3 shows that the proposed method again outperforms all compared state-of-the-art methods with slight accuracy advantages both in mAP and Rank-1. More specifically, our method surpasses the 2nd best model CRAN by 0.66% (89.26 - 88.6)/0.78% (96.58 - 95.8) in mAP/Rank-1. The proposed method performed well in both Rank-5 and Rank-10, ranking second, only 0.18% and 0.12% lower than the best VPM. Besides, we

also compared our method with some feature learning methods that were listed separately in Table 3, such as VPM, JSTL_DGD+ICV-ECCL, DaF, and MSCAN. It can be seen that the results produced by our methods are not inferior to them. We can see that our method surpasses the 2nd best feature learning method VPM a lot, 9.26% (89.26 - 80.0)/3.58% (96.58 - 93.0) in mAP/Rank-1. And our method achieved almost the same result as the best VPM in Rank-5/Rank-10. This suggests the efficacy of our learning framework.

Table 3: Results on the Market1501 dataset. Metric: mAP, Rank-1, Rank-5 and Rank-10 (%). 1st/2nd/3rd best in red/blue/green.

Methods	References	mAP	Rank-1	Rank-5	Rank-10
CRAN [51]	TCSVT 2019	88.6	95.8	-	-
SPReID [56]	CVPR 2018	83.36	93.68	97.57	98.40
CASN [52]	CVPR 2019	82.8	94.4	-	-
HPM [42]	AAAI 2019	82.7	94.2	97.5	98.5
KPM [57]	CVPR 2018	75.3	90.1	96.7	97.9
Pose-Transfer [53]	CVPR 2018	68.92	87.65	-	-
PDC [58]	ICCV 2017	63.4	84.1	92.7	94.9
JLML [59]	IJCAI 2017	62.5	85.1	-	-
SVDNet [54]	ICCV 2017	62.1	82.3	92.3	95.2
VPM [60]	CVPR 2019	80.0	93.0	97.8	98.8
JSTL_DGD+ICV-ECCL [44]	TIP 2018	77.3	90.6	-	-
DaF [41]	BMVC 2017	72.4	82.3	-	-
MSCAN [43]	CVPR 2017	57.53	80.31	-	-
Ours		89.26	96.58	97.62	98.68

4.3.3. Evaluation on DukeMTMC-ReID

We evaluated the proposed method against 15 existing methods on recently released DukeMTMC-ReID dataset. Compared with Market1501, the camera in DukeMTMC-ReID has a wider field of view and a more complex scene, resulting in great changes in the image resolution and background, which makes the re-id task more challenging. Table 4 shows the clear performance superiority of the proposed method over all state-of-the-art methods with significant Rank-1 and Rank-5 advantages. Our model achieves a small margin (+0.93%/+1.18%) over the 2nd best CASN/HPM in Rank-1/Rank-5. Some feature learning methods like HPM, VPM, DPFL and BraidNet-CS+SRL are listed separately in Table 4. Experimental results show that the proposed method performs the best as compared with others. Since TJ-AIDL is in unsupervised learning, the accuracy rate is inherently lower. This demonstrates the efficacy of IDB-CMDL and the complementary role of hand-crafted features again.

Table 4: Results on the DukeMTMC-ReID dataset. Metric: mAP, Rank-1, Rank-5 and Rank-10 (%). 1st/2nd/3rd best in red/blue/green.

Methods	References	mAP	Rank-1	Rank-5	Rank-10
CRAN [51]	TCSVT 2019	74.7	87.6	-	-
CASN [52]	CVPR 2019	73.7	87.7	-	-
SPReID [56]	CVPR 2018	73.34	85.95	92.95	94.52
Two-stream+dMpRL-II [61]	TIP 2019	58.56	76.81	-	-
TJ-AIDL [62]	CVPR 2018	23.0	44.3	59.6	65.0
AOSReID [63]	CVPR 2018	62.10	79.17	-	-
SVDNet [54]	ICCV 2017	56.8	76.7	86.4	89.9
ACRN [64]	CVPR Workshop 2017	52.0	72.6	84.8	88.9
Pose-Transfer [53]	CVPR 2018	48.06	68.64	-	-
OIM Loss [33]	CVPR 2015	47.4	68.1	-	-
Basel.+LSRO [50]	ICCV 2017	47.1	67.7	-	-
HPM [42]	AAAI 2019	74.3	86.6	93.0	95.1
VPM [60]	CVPR 2019	72.6	83.6	91.7	94.2
DPFL [40]	ICCV Workshop 2017	60.6	79.2	-	-
BraidNet-CS+SRL [45]	CVPR 2018	59.49	76.44	-	-
Ours		74.34	88.63	94.18	94.97

4.4. Feature analysis

To illustrate the rationality of the combination of lower and higher convolution layers, considering too many combinations, we only compare different convolution layer combinations on CUHK03 labelled dataset. In addition, we also verify the necessity of the multi-level features constructed in this paper on Market1501 dataset.

In this paper, we demonstrate the complementary role of the deep features of the higher and lower layers learned from fine-tuning VGG16. There are 13 convolution layers in VGG16. We consider Conv1-Conv4 as

the lower layers and Conv9-Conv13 convolutional layer as the higher layers. We try different combinations of lower and higher layers to verify the rationality of our choice. As shown in Fig. 7, our combination (Conv3 + Conv10) achieved the best mAP result (69.2%, see Fig. 7(a)) and the best Rank-1 result (74.4%, see Fig. 7(b)) on the labelled CUHK03 dataset, 1.0% and 1.6% higher than the suboptimal (Conv3 + Conv11), respectively.

Figure 7: mAP and Rank-1 results for different combinations of lower and higher convolution layers on CUHK03 labelled dataset.

In addition, feature analysis experiments are conducted on Market1501 dataset to evaluate the complementarity among different features, and the results are shown in Table 5. To fully prove the effectiveness of each feature, feature analysis is also divided into three different levels, that is, $l = 1, 2, 3$. $l = 1$ represents the individual feature, $l = 2$ is a combination of any two features, and $l = 3$ is the full multi-level feature proposed in this paper. In particular, when analyzing individual feature ($l = 1$), the fusion parameters in Eq. (13) are ignored since the weight of individual feature should be 1, while when analyzing feature combinations ($l = 2, 3$), the fusion parameters remain the same as the original setting, $\gamma_1 = 0.3$, $\gamma_2 = 0.2$. It can be seen from Table 5 that as the most important component in the fusion feature, deep feature has got the best result in the individual feature analysis, and its effect is obviously greater than the other two (79.54% > 54.63% > 47.65% mAP), even better than their combination, LOMO & SIFT (+15.23% mAP/+16.62% Rank-1). What's more, we find that performance gradually improves with the level l . For example, from Deep to Deep & LOMO, then to Deep & LOMO & SIFT, mAP (same regularity on Rank-1, 5, 10) accuracy is gradually improved (from 79.54% to 86.14%, and then to 89.26%). Besides, we concluded that these three features contribute to the fusion in different degrees. Deep feature plays the most important role, LOMO and SIFT play an auxiliary role, and LOMO is slightly more useful than SIFT (by comparing Deep & LOMO with Deep & SIFT).

Table 5: Results on the Market1501 dataset. Metric: mAP, Rank-1, Rank-5 and Rank-10 (%).

Feature combinations	l	mAP	Rank-1	Rank-5	Rank-10
LOMO	1	54.63	56.78	63.21	66.84
SIFT	1	47.65	49.18	54.19	57.97
Deep	1	79.54	84.31	90.11	93.69
Deep & LOMO	2	86.14	93.23	94.47	95.82
Deep & SIFT	2	82.56	87.61	92.73	94.47
LOMO & SIFT	2	64.31	67.69	71.24	74.84
Deep & LOMO & SIFT	3	89.26	96.58	97.62	98.68

4.5. Ablation analysis

To further verify the effect of the three learning processes proposed in this paper, we decomposed them and conducted ablation experiments. Considering that removing the sparse process during analysis of IDB-CMDL results in very high feature dimensions, we only tested on a smaller CUHK03 dataset to reduce execution time.

We compared the four configurations (C1-C4) on the CUHK03 dataset with the labelled settings and detected settings respectively, as shown in Table 6 and Table 7. The abbreviation CVCWL represents cross-view consistent word learning, and MLML represents multi-level metric learning. $\times$ means to remove while $\checkmark$ is to retain. C4 is the full configuration, including the three processes of IDB-CMDL, CVCWL and MLML. The experimental settings of C1 - C3 are as follows: (1) for C1, we just remove the IDB-CMDL module because it is completely independent of the other two modules. (2) for C2, we only remove the consistency restriction of the word distribution (Eq. (11)) in optimization objective of CVCWL, which is exactly what we innovate compared with the traditional BoVW method. In other words, we replace CVCWL with the traditional BoVW method. (3) for C3, we only removed our innovation in MLML that focused on evaluating the feature distribution difference, and replaced the KL-XQDA with the existing XQDA method. Specifically, KL divergence $D(w_i, w_j)$ in Eq. (16) is replaced by Euclidean distance. It can be seen from the comparison of C1-C3 and C4 that each process has a positive effect on the improvement of recognition rate, because the mAP and Rank-1 of C4 are

higher than any of C1-C3. The influence ranking of three learning processes: IDB-CMDL>>MLML>CVCWL (refer to 54.33<<66.85<70.36 in Table 6, and 55.34<<65.65<68.94 in Table 7).

Table 6: Rank-1 and mAP results for different configurations on CUHK03 labelled datasets.

Config.	IDB-CMDL	CVCWL	MLML	Rank-1	mAP
C1	×	✓	✓	54.33	51.18
C2	✓	×	✓	70.36	66.14
C3	✓	✓	×	66.85	62.24
C4	✓	✓	✓	74.37	69.24

Table 7: Rank-1 and mAP results for different configurations on CUHK03 detected datasets.

Config.	IDB-CMDL	CVCWL	MLML	Rank-1	mAP
C1	×	✓	✓	55.34	52.23
C2	✓	×	✓	68.94	64.07
C3	✓	✓	×	65.65	61.72
C4	✓	✓	✓	72.13	67.65

4.6. Evaluation of fusion weights

We analyze the effect of the weights γ_1 and γ_2 of feature fusion on accuracy. Lots of comparative experiments have been done on Market1501 dataset to show that the accuracy (mAP, Rank-1, Rank-5, Rank-10) reach the highest when $\gamma_1 = 0.3$ and $\gamma_2 = 0.2$. We find the best combination of γ_1 and γ_2 in the range of 0 to 0.5 with step length of 0.1. The experimental results are shown in Fig. 8. As can be seen from Fig. 8, the accuracy is obviously lower than our optimal combination (0.3, 0.2) when $\gamma_1 = \gamma_2 = 0$ which means ignoring the auxiliary role of LOMO and SIFT, and considering only the discrimination of local deep features. These experiments fully demonstrate the complementary role of the hand-crafted feature and the rationality of weight parameters selection.

Figure 8: Accuracy results for different combinations of fusion weights γ_1 and γ_2 on Market1501 dataset.

5. Conclusions

In this paper, we proposed a multi-level cross-view consistent feature learning framework for person re-identification, which consists of three learning processes: IDB-CMDL, cross-view consistent word learning and multi-level metric learning. We overcome the challenge of viewpoint change in person re-identification from two aspects: feature consistency and feature distribution consistency. Compared to the state-of-the-art methods, our approach combined multiple techniques and achieved superior results. Specifically, we observed that the proposed method improves the accuracy of mAP and rank-1 by large margins for person re-identification.

Acknowledgments

The research is supported by National Natural Science Foundation of China (No. 61973066, 61471110), Central University Basic Scientific Research Business Expenses Special Funds (N172608005, N182608004).

Appendix A. Appendix

Here we give the detailed solution of IDB-CMDL and some detailed derivation of formulas in Eq. 3.3

Appendix A.1. Solution of IDB-CMDL

Although there are many variables in (7), we can alternatively optimize these variables as follows. Fix other variables and update $A_{(1,l)}$ and $A_{(2,l)}$. By ignoring the irrelevant variables with respect to $A_{(1,l)}$, the objective (8) is reduced to

$$\min_{A_{(1,l)}} \|X_{(1,l)} - D_{(1,l)}A_{(1,l)}\|_F^2 + \alpha \|P_{(1,l)}X_{(1,l)} - A_{(1,l)}\|_F^2 \quad (\text{A.1})$$

Setting the derivative with respect to $A_{(1,l)}$ to zero, we get the solution

$$A_{(1,l)} = (D_{(1,l)}^T D_{(1,l)} + \alpha I)^{-1} (D_{(1,l)}^T X_{(1,l)} + \alpha P_{(1,l)} X_{(1,l)}) \quad (\text{A.2})$$

where I is an identity matrix. Similarly, we get the solution to $A_{(2,l)}$ as

$$A_{(2,l)} = (D_{(2,l)}^T D_{(2,l)} + \alpha I)^{-1} (D_{(2,l)}^T X_{(2,l)} + \alpha P_{(2,l)} X_{(2,l)}) \quad (\text{A.3})$$

Fix other variables and update $A_{(1,l,e)}$ and $A_{(2,l,e)}$. The objective function regarding $A_{(1,l,e)}$ is reduced to

$$\min_{A_{(1,l,e)}} \|X_{(1,l,e)} - D_{(1,l)}A_{(1,l,e)}\|_F^2 + \beta \|P_{(1,l)}X_{(1,l,e)} - A_{(1,l,e)}\|_F^2 + \lambda_l \|A_{(1,l,e)} - A_{(2,l,e)}\|_F^2 \quad (\text{A.4})$$

Setting the derivative with respect to $A_{(1,l,e)}$ to zero, we get the solution

$$A_{(1,l,e)} = (D_{(1,l)}^T D_{(1,l)} + (\beta + \lambda_l)I)^{-1} (D_{(1,l)}^T X_{(1,l,e)} + \lambda_l A_{(2,l,e)} + \beta P_{(1,l)} X_{(1,l,e)}) \quad (\text{A.5})$$

Similarly, we get the solution to $A_{(2,l,e)}$ as

$$A_{(2,l,e)} = (D_{(2,l)}^T D_{(2,l)} + (\beta + \lambda_l)I)^{-1} (D_{(2,l)}^T X_{(2,l,e)} + \lambda_l A_{(1,l,e)} + \beta P_{(2,l)} X_{(2,l,e)}) \quad (\text{A.6})$$

Fix other variables and update $P_{(1,l)}$ and $P_{(2,l)}$. The objective function regarding $P_{(1,l)}$ is reduced to

$$\min_{P_{(1,l)}} \alpha \|P_{(1,l)}X_{(1,l)} - A_{(1,l)}\|_F^2 + \beta \|P_{(1,l)}X_{(1,l,e)} - A_{(1,l,e)}\|_F^2 \quad (\text{A.7})$$

Setting the derivative with respect to $P_{(1,l)}$ to zero, we get the solution

$$P_{(1,l)} = (\alpha A_{(1,l)} X_{(1,l)}^T + \beta A_{(1,l,e)} X_{(1,l,e)}^T) (\alpha X_{(1,l)} X_{(1,l)}^T + \beta X_{(1,l,e)} X_{(1,l,e)}^T)^{-1} \quad (\text{A.8})$$

Similarly, we get the solution to $P_{(2,l)}$ as

$$P_{(2,l)} = (\alpha A_{(2,l)} X_{(2,l)}^T + \beta A_{(2,l,e)} X_{(2,l,e)}^T) (\alpha X_{(2,l)} X_{(2,l)}^T + \beta X_{(2,l,e)} X_{(2,l,e)}^T)^{-1} \quad (\text{A.9})$$

Fix other variables and update $D_{(1,l)}$ and $D_{(2,l)}$. The objective function regarding $D_{(1,l)}$ is reduced to

$$\begin{aligned} \min_{D_{(1,l)}} \quad & \|X_{(1,l)} - D_{(1,l)}A_{(1,l)}\|_F^2 \\ & + \|X_{(1,l,e)} - D_{(1,l)}A_{(1,l,e)}\|_F^2 \\ \text{s.t.} \quad & \|d_{1l(:,i)}\| \leq 1, \|d_{2l(:,i)}\| \leq 1, i = 1, \dots, m \end{aligned} \quad (\text{A.10})$$

Problem (A.10) can be solved by the Alternating Direction Method Of Multipliers (ADMM) algorithm. We have similar solution to $D_{(2,l)}$. The above procedures are repeated until convergence. The complete procedures of IDB-CMDL is summarized in **Algorithm 1**.

Algorithm 1: ID-based cross-view multi-level dictionary learning (IDB-CMDL)

Input: Local feature sets in three different levels from two different camera views $X_{(1,l)}$ and $X_{(2,l)}$, $l = 1, 2, 3$.

Output: $D_{(1,l)}$, $D_{(2,l)}$, $P_{(1,l)}$, and $P_{(2,l)}$.

Initialize $D_{(1,l)}$, $D_{(2,l)}$, $P_{(1,l)}$, $P_{(2,l)}$, $A_{(1,l)}$, $A_{(2,l)}$, $A_{(1,l,e)}$, $A_{(2,l,e)}$, α , β , and λ_l ;

while not converge **do**

Fix other variables, update $A_{(1,l)}$ and $A_{(2,l)}$ according to (A.1), (A.2), (A.3);
 Fix other variables, update $A_{(1,l,e)}$ and $A_{(2,l,e)}$ according to (A.4), (A.5), (A.6);
 Fix other variables, update $P_{(1,l)}$ and $P_{(2,l)}$ according to (A.7), (A.8), (A.9);
 Fix other variables, update $D_{(1,l)}$ and $D_{(2,l)}$ according to (A.10);

end

return $D_{(1,l)}$, $D_{(2,l)}$, $P_{(1,l)}$, and $P_{(2,l)}$;

Appendix A.2. Derivation of some formulas in 3.3

The detailed derivation of distribution difference distance $D(w_i, w_j)$ in Eq. (15) is

$$\begin{aligned} D(w_i, w_j) &= KL(w_i, w_j) + KL(w_j, w_i) \\ &= \sum_{k=1}^L w_i^k \ln \frac{w_i^k}{w_j^k} + \sum_{k=1}^L w_j^k \ln \frac{w_j^k}{w_i^k} \\ &= \sum_{k=1}^L w_i^k (\ln w_i^k - \ln w_j^k) + w_j^k (\ln w_j^k - \ln w_i^k) \\ &= \sum_{k=1}^L w_i^k \ln w_i^k - w_i^k \ln w_j^k + w_j^k \ln w_j^k - w_j^k \ln w_i^k \\ &= \sum_{k=1}^L w_i^k \ln w_i^k - w_i^k \ln w_j^k + \sum_{k=1}^L w_j^k \ln w_j^k - w_j^k \ln w_i^k \\ &= (w_i \ln \frac{w_i}{w_j})^T (w_i \ln \frac{w_i}{w_j}) + (w_j \ln \frac{w_j}{w_i})^T (w_j \ln \frac{w_j}{w_i}) \end{aligned} \quad (\text{A.11})$$

The detailed derivation of metric distance $d(w_i, w_j)$ in Eq. (16) is

$$\begin{aligned} d(w_i, w_j) &= (w_i \ln \frac{w_i}{w_j})^T (\sum_I^{-1} - \sum_E^{-1}) (w_i \ln \frac{w_i}{w_j}) + (w_j \ln \frac{w_j}{w_i})^T (\sum_I^{-1} - \sum_E^{-1}) (w_j \ln \frac{w_j}{w_i}) \\ &= (\sum_I^{-1} - \sum_E^{-1}) [(w_i \ln \frac{w_i}{w_j})^T (w_i \ln \frac{w_i}{w_j}) + (w_j \ln \frac{w_j}{w_i})^T (w_j \ln \frac{w_j}{w_i})] \\ &= (\sum_I^{-1} - \sum_E^{-1}) D(w_i, w_j) \end{aligned} \quad (\text{A.12})$$

After mapping to subspace φ , the metric distance in Eq. (17) is derived as

$$\begin{aligned} d(w_i, w_j) &= (w_i \ln \frac{w_i}{w_j})^T \varphi (\sum_I^{-1} - \sum_E^{-1}) \varphi^T (w_i \ln \frac{w_i}{w_j}) + (w_j \ln \frac{w_j}{w_i})^T \varphi (\sum_I^{-1} - \sum_E^{-1}) \varphi^T (w_j \ln \frac{w_j}{w_i}) \\ &= \varphi (\sum_I^{-1} - \sum_E^{-1}) \varphi^T [(w_i \ln \frac{w_i}{w_j})^T (w_i \ln \frac{w_i}{w_j}) + (w_j \ln \frac{w_j}{w_i})^T (w_j \ln \frac{w_j}{w_i})] \\ &= \varphi (\sum_I^{-1} - \sum_E^{-1}) \varphi^T D(w_i, w_j) \end{aligned} \quad (\text{A.13})$$

The explanation of Eq. (18) and Eq. (19) can be found in 18.

References

- [1] M. Li, F. Shen, J. Wang, C. Guan, and J. Tang, "Person re-identification with activity prediction based on hierarchical spatial-temporal model," *Neurocomputing*, 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0925231217315837>
- [2] J. You, A. Wu, X. Li, and W. S. Zheng, "Top-push video-based person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1345–1353.
- [3] C. Zhang, W. Lin, and W. Yang, "Crossing generative adversarial networks for cross-view person re-identification," *Neurocomputing*, vol. 340, 2018.

- [4] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1565–1573.
- [5] M. I. Khedher, M. A. El-Yacoubi, and B. Dorizzi, "Fusion of appearance and motion-based sparse representations for multi-shot person re-identification," *Neurocomputing*, vol. 248, no. Supplement C, pp. 94 – 104, 2017, neural Networks : Learning Algorithms and Classification Systems. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0925231217304307>
- [6] B. Mirmahboub, M. L. Mekhalfi, and V. Murino, "Person re-identification by order-induced metric fusion," *Neurocomputing*, 2018.
- [7] L. Yixiu, Z. Yunzhou, C. Sonya, B. Bir, and L. Shuangwei, "A new patch selection method based on parsing and saliency detection for person re-identification," *Neurocomputing*, vol. 374, 2020.
- [8] J. Li, T. Yao, Q. Ling, and T. Mei, "Detecting shot boundary with sparse coding for video summarization," *Neurocomputing*, vol. 266, no. nov.29, pp. 66–78.
- [9] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015, pp. 36–45.
- [10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1717–1724.
- [11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," vol. 50, no. 1, pp. I–647, 2014.
- [12] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.
- [13] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3908–3916.
- [14] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1249–1258.
- [15] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–8.
- [16] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 907–915.
- [17] A. Feizi, "High-level feature extraction for classification and person re-identification," *IEEE Sensors Journal*, vol. 17, no. 21, pp. 7064–7073, Nov 2017.
- [18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2197–2206.
- [19] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2360–2367.
- [20] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 356–370, Feb 2017.
- [21] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image & Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.
- [22] M. Aharon, M. Elad, and A. Bruckstein, "-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [23] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition*, 2010, pp. 2691–2698.
- [24] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Computer Vision and Pattern Recognition*, 2014, pp. 3550–3557.

- [25] S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.
- [26] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1116–1124.
- [27] L. Bazzani, M. Cristani, and V. Murino, "Sdalf: Modeling human appearance with symmetry-driven accumulation of local features," vol. 63, no. 4, pp. 43–69, 2014.
- [28] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *International Conference on Computer Vision*, 2012, pp. 413–422.
- [29] C. Engel, P. Baumgartner, M. Holzmam, and J. F. Nutz, "Person re-identification by support vector ranking," in *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*, 2010, pp. 1–11.
- [30] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3200–3208.
- [31] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2288–2295.
- [32] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, Sept 2012, pp. 203–208.
- [33] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3376–3385.
- [34] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, July 2017.
- [35] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, and N. Zheng, "Deep self-paced learning for person re-identification," *Pattern Recognition*, 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S003132031730403X>
- [36] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4806–4817, Oct 2017.
- [37] S. Z. Chen, C. C. Guo, and J. H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, May 2016.
- [38] X. Zhu, X. Y. Jing, L. Yang, X. You, D. Chen, G. Gao, and Y. Wang, "Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.
- [39] Q. Zhou, S. Zheng, H. Ling, H. Su, and S. Wu, "Joint dictionary and metric learning for person re-identification," *Pattern Recognition*, vol. 72, no. Supplement C, pp. 196 – 206, 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320317302510>
- [40] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 2590–2600.
- [41] Y. Rui, Z. Zhou, B. Song, and B. Xiang, "Divide and fuse: A re-ranking approach for person re-identification," *BMVC*, 2017.
- [42] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," *2019 AAAI Conference on Artificial Intelligence*, 2019.
- [43] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 7398–7407.
- [44] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 1470–1478.
- [45] A. Wu, W. Zheng, and J. Lai, "Learning view-specific deep networks for person re-identification," *2019 IEEE International Conference on Computer Vision (ICCV)*, Oct 2019, pp. 6922–6931.

- [46] Z. Feng, J. Lai, and X. Xie, "Unsupervised Person Re-Identification by Camera-Aware Similarity Consistency Learning," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3472–3483, July 2018.
- [47] Y. Chen, X. Zhu, and S. Gong, "Deep Association Learning for Unsupervised Video Person Re-identification," *British Machine Vision Conference (BMVC)*, Oct 2018.
- [48] R. He, W. Xiang, S. Zhenan, and T. Tieniu, "Wasserstein cnn: Learning invariant features for nir-vis face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–1.
- [49] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 152–159.
- [50] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE International Conference on Computer Vision*, 2017.
- [51] C. Han, R. Zheng, C. Gao, and N. Sang, "Complementation-reinforced attention network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.
- [52] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 5728–5737.
- [53] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 4099–4108.
- [54] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *IEEE International Conference on Computer Vision*, 2017.
- [55] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3652–3661.
- [56] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 1062–1071.
- [57] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6886–6895.
- [58] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 3980–3989.
- [59] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," 2017.
- [60] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 393–402.
- [61] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1391–1403, March 2019.
- [62] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 2275–2284.
- [63] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 5098–5107.
- [64] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1435–1443.

Yixiu Liu received the B.E. degrees from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2016, and is currently working toward the Ph.D. degree at School of School of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests are in the area of computer vision, including person re-identification, pedestrian tracking.

Yunzhou Zhang received B.S. and M.S. degree in Mechanical and Electronic engineering from National University of Defense Technology, Changsha, China in 1997 and 2000, respectively. He received Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009. He is currently a professor with the Faculty of Robot Science and Engineering, Northeastern University, China. Now he leads the Cloud Robotics and Visual Perception Research Group. His research has been supported by funding from various sources such as National Natural Science Foundation of China, Ministry of science and technology of China, Ministry of Education of China and some famous high-tech companies. He has published many journal papers and conference papers in intelligent robots, computer vision and wireless sensor networks. His research interests include intelligent robot, computer vision, and sensor networks.

Bir Bhanu (M'82–F'95–LF'17) received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, and the M.B.A. degree from the University of California at Irvine, Irvine, CA. He was the Founding Professor of electrical engineering with the University of California at Riverside (UCR), Riverside, CA, and served as its first Chair from 1991 to 1994. He has been the Cooperative Professor of computer science and engineering (since 1991), bioengineering (since 2006), and mechanical engineering (since 2008). He served as the Interim Chair of the Department of Bioengineering from 2014 to 2016. He also served as the Director of the National Science Foundation Graduate Research and Training Program in video bioinformatics with UCR. He is currently the Bourns Presidential Chair in engineering, the Distinguished Professor of electrical and computer engineering, and the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory, UCR. He has published extensively and has 18 patents. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human-computer interactions, and biological, medical, military, and intelligence applications. He was a Senior Honeywell Fellow with Honeywell Inc. He is a Fellow of IEEE, AAAS, IAPR, SPIE, and AIMBE.

Sonya Coleman (M'11) received a BSc (Hons) in Mathematics, Statistics and Computing (first class) from the Ulster University, UK in 1999, and a PhD in Mathematics from the Ulster University, UK in 2003. She is a Professor and a leader in the Cognitive Robotics team of Intelligent Systems Research Centre. She is a Fellow of the Higher Education Academy. She has many publications in image processing, pattern recognition, computational intelligence and robotics. Her research has been supported by funding from various sources such as EPSRC, The Nuffield Foundation, The Leverhulme Trust and the European Commission. Additionally, she was co-investigator on the EU FP7 funded project RUBICON, the FP7 project VISUALISE and is currently co-investigator in the FP7 SLANDIAL project. She is also secretary of the Irish Pattern Recognition and Classification Society.

Dermot Kerr received a BSc(Hons) in Computing Science from the University of Ulster, UK in 2005, and a PhD in Computing and Engineering from the University of Ulster, UK in 2009. He is currently a research fellow in the School of Computing and Intelligent System at the University of Ulster, Magee. His current research interests are in mathematical image processing, feature detection, omnidirectional vision and robotics. Dr. Kerr is a member of the Irish Pattern Recognition and Classification Society.

<p>Yixiu Liu received the B.E. degrees from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2016, and is currently working toward the Ph.D. degree at School of School of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests are in the area of computer vision, including person re-identification, pedestrian tracking.</p>
<p>Yunzhou Zhang received B.S. degree and M.S. degree in Mechanical and Electronic engineering from National University of Defense Technology, Changsha, China in 1997 and 2000, respectively. He received Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009, where he is currently a professor with at the School of Information Science and Engineering, Northeastern University, China. His research interests include intelligent robot, computer vision, and sensor networks.</p>
<p>Bir Bhanu received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, and the M.B.A. degree from the University of California at Irvine, Irvine, CA. He was the Founding Professor of electrical engineering with the University of California at Riverside (UCR), Riverside, CA, and served as its first Chair from 1991 to 1994. He has been the Cooperative Professor of computer science and engineering (since 1991), bioengineering (since 2006), and mechanical engineering (since 2008). He served as the Interim Chair of the Department of Bioengineering from 2014 to 2016. He also served as the Director of the National Science Foundation Graduate Research and Training Program in video bioinformatics with UCR. He is currently the Bourns Presidential Chair in engineering, the Distinguished Professor of electrical and computer engineering, and the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory, UCR. He has published extensively and has 18 patents. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human-computer interactions, and biological, medical, military, and intelligence applications. In 1991, he was a Senior Honeywell Fellow with Honeywell Inc. He is a Fellow of AAAS, IAPR, SPIE, and AIMBE.</p>

Sonya Coleman received the B.Sc. (Hons) degree in mathematics, statistics, and computing and the Ph.D. degree in mathematics from the University of Ulster, Londonderry, U.K., in 1999 and 2003, respectively. She is currently a Lecturer in the School of Computing and Intelligent Systems, Magee College, University of Ulster. She has more than 50 publications primarily in the field of mathematical image processing, and much of the recent research undertaken by her has been supported by funding from EPSRC award EP/C006283/11, the Leverhulme Trust, and the Nuffield Foundation. Additionally, she is co-investigator on the EU FP7 funded project RUBICON. She is the author or coauthor of over 70 research papers on image processing, robotics, and computational neuroscience. Dr. Coleman was awarded the Distinguished Research Fellowship by the University of Ulster in recognition of her contribution to research in 2009.

Dermot Kerr received a BSc(Hons) in Computing Science from the University of Ulster, UK in 2005, and a PhD in Computing and Engineering from the University of Ulster, UK in 2009. He is currently a research fellow in the School of Computing and Intelligent Systems at the University of Ulster, Magee. His current research interests are in mathematical image processing, feature detection, omnidirectional vision and robotics. Dr. Kerr is a member of the Irish Pattern Recognition and Classification Society.

Yixiu Liu
Yunzhou Zhang
Bir Bhanu
Sonya Coleman

Dermot Kerr

***Source Files - Latex or Word**

[Click here to download Source Files - Latex or Word: Source files.zip](#)

Conflict of Interest Form

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Author’s name	Affiliation
Yixiu Liu	Northeastern University, Shenyang, China
Yunzhou Zhang	Northeastern University, Shenyang, China
Bir Bhanu	University of California, Riverside
Sonya Coleman	University of Ulster, Derry, UK
Dermot Kerr	University of Ulster, Derry, UK

Credit Author Statement

The following is the list of contribution from each author:

- Yixiu Liu:

Conceptualization, Methodology, Software, Investigation, Data Curation, Writing-Original Draft, Writing-Review&Editing, Visualization
- Yunzhou Zhang:

Formal analysis, Project administration, Writing- Original draft preparation, Funding acquisition
- Bir Bhanu:

Project administration, Project administration, Supervision.
- Sonya Coleman:

Resources, Supervision
- Dermot Kerr:

Software, Validation

Author’s name	Affiliation
Yixiu Liu	Northeastern University, Shenyang, China
Yunzhou Zhang	Northeastern University, Shenyang, China
Bir Bhanu	University of California, Riverside
Sonya Coleman	University of Ulster, Derry, UK
Dermot Kerr	University of Ulster, Derry, UK

Declaration of Interest Statement

We declared that they have no conflicts of interest to this work, and do not have any commercial or associative interest that represents a conflict of interest in connection with the work.

There is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, “Multi-level Cross-view Consistent Feature Learning for Person Re-identification”.

Statement by Yixiu Liu, Yunzhou Zhang, Bir Bhanu, Sonya Coleman, and Dermot Kerr.

Author’s name	Affiliation
Yixiu Liu	Northeastern University, Shenyang, China
Yunzhou Zhang	Northeastern University, Shenyang, China
Bir Bhanu	University of California, Riverside
Sonya Coleman	University of Ulster, Derry, UK
Dermot Kerr	University of Ulster, Derry, UK