Person Re-Identification by Camera Correlation Aware Feature Augmentation

Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, Jian-Huang Lai

Code is available at the project page:
http://isee.sysu.edu.cn/%7zhwshi/project/CRAFT.html

For reference of this work, please cite:

Ying-Cong Chen, Xiatian Zhu,Wei-Shi Zheng, and Jian-Huang Lai. Person Re-Identification by Camera Correlation Aware Feature Augmentation.IEEETransactions on Pattern Analysis and Machine Intelligence (DOI: 10.1109/TPAMI.2017.2666805)

Bib:
@article{craft,
title={Person Re-Identification by Camera Correlation Aware Feature Augmentation},
author={Chen, Ying-Cong, and Zhu, Xiatian, and Zheng, Wei-Shi and Lai, Jian-Huang},
journal={IEEETransactions on Pattern Analysis and Machine Intelligence (DOI: 10.1109/TPAMI.2017.2666805)}
}

Person Re-Identification by Camera Correlation Aware Feature Augmentation

Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, Jian-Huang Lai

Abstract—The challenge of person re-identification (re-id) is to match individual images of the same person captured by different non-overlapping camera views against significant and unknown cross-view feature distortion. While a large number of distance metric/subspace learning models have been developed for re-id, the cross-view transformations they learned are view-generic and thus potentially less effective in quantifying the feature distortion inherent to each camera view. Learning view-specific feature transformations for re-id (i.e., view-specific re-id), an under-studied approach, becomes an alternative resort for this problem. In this work, we formulate a novel view-specific person re-identification framework from the feature augmentation point of view, called **C**amera co**R**relation **A**ware **F**eature augmen**T**ation (CRAFT). Specifically, CRAFT performs cross-view adaptation by automatically measuring camera correlation from cross-view visual data distribution and adaptively conducting feature augmentation to transform the original features into a new adaptive space. Through our augmentation framework, view-generic learning algorithms can be readily generalized to learn and optimize view-specific sub-models whilst simultaneously modelling view-generic discrimination information. Therefore, our framework not only inherits the strength of view-generic model learning but also provides an effective way to take into account view specific characteristics. Our CRAFT framework can be extended to jointly learn view-specific feature transformations for person re-id across a large network with more than two cameras, a largely under-investigated but realistic re-id setting. Additionally, we present a domain-generic deep person appearance representation which is designed particularly to be towards view invariant for facilitating cross-view adaptation by CRAFT. We conducted extensively comparative experiments to validate the superiority and advantages of our proposed framework over state-of-the-art competitors on contemporary challenging person re-id datasets.

Index Terms—Person re-identification, adaptive feature augmentation, view-specific transformation.

1 Introduction

The extensive deployment of close-circuit television cameras in visual surveillance results in a vast quantity of visual data and necessitates inevitably automated data interpretation mechanisms. One of the most essential visual data processing tasks is to automatically re-identify individual person across non-overlapping camera views distributed at different physical locations, which is known as person re-identification (re-id). However, person re-id by visual matching is inherently challenging due to the existence of many visually similar people and dramatic appearance changes of the same person arising from the great cross-camera variation in viewing conditions such as illumination, viewpoint, occlusions and background clutter [1] (Figure 1).

In current person re-id literature, the best performers are discriminative learning based methods [2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21]. Their essential objective is to establish a reliable re-id matching model through learning identity discriminative information from the pairwise training data. Usually, this is achieved by either *view-generic* modelling (e.g., optimizing a common model for multiple camera views) [5,6,8,10,13,15] or *view-specific* modelling scheme (e.g., optimizing a separate model for each camera view) [16,17,18,19]. The former mainly focuses on the shared view-generic discriminative learning but does not explicitly take the individual view information (e.g., via camera view labels) into modelling. Given that person re-id inherently incurs dramatic appearance change across camera views due to the great difference in illumination, viewpoint or camera characteristics, the view-generic approach is inclined to be sub-optimal in quantifying the feature distortion caused by these variations in individual camera views. While the latter approach may enable to mitigate this problem by particularly considering view label information during modelling, most of these methods do not explicitly take into consideration the feature distribution alignment across camera views so that cross-view data adaptation cannot be directly quantified and optimized during model learning as view-generic counterparts do. Additionally, existing view-specific methods are often subject to limited scalability for person re-id across multiple (more than two) camera views in terms of implicit assumptions and formulation design.

In view of the analysis above, we formulate a novel view-specific person re-identification framework, named ***C**amera co**R**relation **A**ware **F**eature augmen**T**ation* (CRAFT), capable of performing cross-view feature adaptation by measuring cross-view correlation from visual data distribution and carrying out adaptive feature augmentation to transform the original features into a new augmented space. Specifically, we quantify the underlying camera

- *Ying-Cong Chen is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China, with the Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China, and is also with Department of Computer Science and Engineering, The Chinese University of Hong Kong. E-mail: yingcong.ian.chen@gmail.com*
- *Xiatian Zhu is with School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China; and also with School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom. E-mail: xiatian.zhu@qmul.ac.uk*
- *Wei-Shi Zheng is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China., and is also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yatsen University), Ministry of Education, China. E-mail: wszheng@ieee.org.*
- *Jian-Huang Lai are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China, and is also with Guangdong Province Key Laboratory of Information Security, P. R. China. E-mail: stsljh@mail.sysu.edu.cn.*

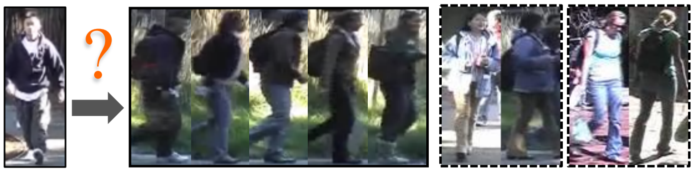


Fig. 1. Illustration of person re-id challenges [1]. **Left**: great visual similarity among different people. **Right**: large cross-view appearance variations of the same people, each person within a dotted box.

correlation in our framework by generalizing the conventional zero-padding, a non-parameterized feature augmentation mechanism, to a parameterized feature augmentation. As a result, any two cameras can be modelled adaptively but not independently, whereas the common information between camera views have already been quantified in the adaptive space. Through this augmentation framework, view-generic learning algorithms can be readily generalized to induce view-specific sub-models whilst involving simultaneously view-generic discriminative modelling. More concretely, we instantialize our CRAFT framework with Marginal Fisher Analysis (MFA) [22], leading to a re-id method instance called CRAFT-MFA. We further introduce camera view discrepancy regularization in order to append extra modelling capability for controlling the correlation degree between view-specific sub-models. This regularization can be viewed as a complementary means of incorporating camera correlation modelling on top of the proposed view-specific learning strategy. Moreover, our CRAFT framework can be flexibly deployed for re-id across multiple (> 2) camera views by jointly learning a unified model, which is largely under-studied in existing approaches.

Apart from cross-view discriminative learning, we also investigate domain-generic (i.e., independent of target data or domain) person appearance representation, with the aim to make person features towards view invariant for facilitating the cross-view adaptation process using CRAFT. In particular, we explore the potential of deep learning techniques for person appearance description, inspired by the great success of deep neural networks in other related applications like object recognition and detection [23,24,25]. Differing significantly from existing deep learning based re-id approaches [26,27,28,29] that typically learn directly from the small person re-id training data, we instead utilize large auxiliary less-related image data. This strategy allows to not only avoid the insufficient training data problem and address the limited scalability challenge in deep re-id models, but also yield domain-generic person features with more tolerance to view variations.

The main **contributions** of this work include: **(I)** We propose a camera correlation aware feature augmentation person re-id framework called CRAFT. Our framework is able to generalize existing view-generic person re-identification models to perform view-specific learning. A kernelization formulation is also presented. **(II)** We extend our CRAFT framework to jointly learn view-specific feature transformations for person re-id across a large network involving more than two cameras. Although this is a realistic scenario, how to build an effective unified re-id model for an entire camera network is still under-explored in existing studies. **(III)** We present a deep convolutional network based appearance feature extraction method in order to extract domain-generic and more view invariant person features. To our knowledge, this is the first attempt that explores deep learning with large auxiliary non-person image data for constructing discriminative re-id features. For evaluating our method, we conducted extensive comparisons between CRAFT and a variety of state-of-the-art models on VIPeR [30], CUHK01 [31], CUHK03 [26], QMUL GRID [32], and Market-1501 [33] person re-id benchmarks.

2 Related Work

Distance metric learning in person re-id. Supervised learning based methods [6,8,9,11,12,13,14,15,34,35,36,37,38,39,40,41,42,43] dominate current person re-id research by achieving state-of-the-art performance, whilst a much fewer unsupervised re-id methods [44,45,46,47,48,49] have been proposed with much inferior results yielded. This is because large cross-camera variations in viewing conditions may cause dramatic person appearance changes and arise great difficulties for accurately matching identities. Discriminative learning from re-id training data is typically considered as a necessary and effective strategy for reliable person re-id. Notable re-id learning models include PRDC [8], LADF [9], KISSME [13], PCCA [6], LFDA [10], XQDA [12], PSD [15], Metric Ensemble [14], DNS [39], SCSP [40], and so forth.

All the above re-id methods are mainly designed for learning the common view-generic discriminative knowledge, but ignoring greatly the individual view-specific feature variation under each camera view. This limitation can be relaxed by the recent view-specific modelling strategy capable of learning an individual matching function for each camera view. Typical methods of such kind include the CCA (canonical correlation analysis) based approaches ROCCA [16], refRdID [17], KCCA-based re-id [18], and CVDCA [19]. However, they are less effective in extracting the shared discriminative information between different views, because these CCA-based methods [16,17,18] do not directly/simultaneously quantify the commonness and discrepancy between views during learning transformation, so that they cannot identify accurately what information can be shared between views. While CVDCA [19] attempts to quantify the inter-view discrepancy, it is theoretically restricted due to the stringent Gaussian distribution assumption on person image data, which may yield sub-optimal modelling at the presence of typically complex/significant cross-view appearance changes. View-specific modelling for person re-id remains under studied to a great extent.

In this work, we present a different view-specific person re-id framework, characterized by a unique capability of generalizing view-generic distance metric learning methods to perform view-specific person re-id modelling, whilst still preserving their inherent learning strength. Moreover, our method can flexibly benefit many existing distance metric/subspace-based person re-id models for substantially improving re-id performance.

Feature representation in person re-id. Feature representation is another important issue for re-id. Ideal person image features should be sufficiently invariant against viewing condition changes and generalized across different cameras/domains. To this end, person re-id images are often represented by hand-crafted appearance pattern based features, designed and computed according to human domain knowledge [4,12,44,45,46,50,51,52,53,54,55]. These features are usually constituted by multiple different types of descriptors (e.g., color, texture, gradient, shape or edge) and greatly domain-generic (e.g., no need to learn from target labelled training data). Nowadays, deep feature learning for person re-id has attracted increasing attention [26,27,29,56,57,58,59,60,61,62]. These alternatives allow benefiting from the powerful modelling capacity of neural networks, and are thus suitable for joint learning even given very heterogeneous training data [29]. Often, they require a very large collection of labelled training data and can easily suffer from the model overfitting risk in realistic applications when data are not sufficiently provided [14,39]. Also, these existing deep re-id features are typically domain-specific.

In contrast, our method exploits deep learning techniques for automatically mining more diverse and view invariant appearance patterns (*versus* restricted hand-crafted ones) from auxiliary less-relevant image data (*versus* using the sparse person re-id training data), finally leading to more reliable person representation. Furthermore, our deep feature is largely domain-generic with no need for labelled target training data. Therefore, our method possesses simultaneously the benefits of the two conventional feature extraction paradigms above. The advantages of our proposed features over existing popular alternatives are demonstrated in our evaluations (Section 5.2 and Table 5).

Domain adaptation. In a broader context, our cross-view adaptation for re-id is related to but different vitally from domain adaptation (DA) [63,64,65,66,67,68,69,70]. Specifically, the aim of existing DA methods is to diminish the distribution discrepancy between the source and target domains, which is similar conceptually to our method. However, DA models assume typically that the training and test classes (or persons) are overlapped, whilst our method copes with disjoint training and test person (class) sets with the objective to learn a discriminative model that is capable of generalizing to previously unseen classes (people). Therefore, conventional DA methods are less suitable for person re-id.

Feature augmentation. Our model formulation is also related to the zero padding technique [71,72] and feature data augmentation methods in [67,69]. However, ours is different substantially from these existing methods. Specifically, [67] is designed particularly for a heterogeneous modelling problem (i.e., different feature representations are used in distinct domains), whilst person re-id is typically homogeneous and therefore not suitable. More critically, beyond all these conventional methods, our augmentation formulation uniquely considers the relation between transformations of different camera views (even if more than two) and embeds intrinsic camera correlation into the adaptive augmented feature space for facilitating cross-view feature adaptation and finally person identity association modelling.

3 Towards View Invariant Structured Person Representation

We want to construct more view change tolerant person re-id features. To this end, we explore the potential of deep convolutional neural network, encouraged by its great generalization capability in related applications [25,73]. Typically, one has only sparse (usually hundreds or thousands) labelled re-id training samples due to expensive data annotation. This leads to great challenges for deriving effective domain-generic features using deep models [26,27,74]. We resolve the sparse training data problem by learning the deep model with a large auxiliary image dataset, rather than attempting to learn person discriminative features from the small re-id training data. Our intuition is that, a generic person description relies largely on the quality of atomic appearance patterns. Naturally, our method can be largely domain-generic in that appearance patterns learned from large scale data are likely to be more general. To this purpose, we first exploit the AlexNet[1] [23] convolutional neural network (CNN) to learn from the ILSVRC 2012 training dataset for obtaining diverse patterns. Then, we design reliable descriptors to characterize the appearance of a given person image. We mainly leverage the lower convolutional (conv) layers, unlike [24] that utilizes the higher layers (e.g., the 5^{th} conv layer, $6^{\text{th}}/7^{\text{th}}$ fully connected (fc) layers). The reasons are: (1) Higher layers can be more likely to be task-specific (e.g., sensitive to general object categories rather than person identity in our case), and may have worse transferability than lower ones [24,73] (see our evaluations in Table 5). In contrast, lower conv layers correspond to low-level visual features such as color, edge and elementary texture patterns [73], and naturally have better generality across different tasks. (2) Features from higher layers are possibly contaminated by dramatic variations in human pose or background clutter due to their large receptive fields and thus not sufficiently localized for person re-id.

We present two person descriptors based on feature maps of the first two conv layers[2], as detailed below.

Structured person descriptor. The feature data from convnet layers are highly structured, e.g., they are organized in form of multiple 2-D feature maps. Formally, we denote by $\boldsymbol{F}_c \in \mathbb{R}^{h_c \times w_c \times m_c}$ the feature maps of the c-th ($c \in \{1,2\}$) layer, with h_c/w_c the height/width of feature maps, and m_c the number of conv filters. In our case, $h_1 = w_1 = 55, m_1 = 96$; and $h_2 = w_2 = 27, m_2 = 256$. For the c-th layer, $\boldsymbol{F}_c(i,j,\kappa)$ represents the activation of the κ-th filter on the image patch centred at the pixel (i,j). Given the great variations in person pose, we further divide each feature map into horizontal strips with a fixed height h_s for encoding spatial structure information and enhancing pose invariance, similar to ELF [51] and WHOS [46]. We empirically set $h_s = 5$ in our evaluation for preserving sufficient spatial structure. We then extract the intensity histogram (with bin size 16) from each strip for every feature map. The concatenation of all these strip based histograms forms the Histogram of Intensity Pattern (**HIP**) descriptor, with feature dimension $37376 = 16(\text{bins}) \times 11(\text{strips}) \times 96(m_1) + 16(\text{bins}) \times 5(\text{strips}) \times 256(m_2)$. As such, HIP is inherently structured, containing multiple components with different degrees of pattern abstractness.

View-invariant person descriptor. The proposed HIP descriptor encodes completely the activation information of all feature maps, regardless their relative saliency and noise degree. This ensures pattern completeness, but being potentially sensitive to cross-view covariates such as human pose changes and background discrepancy. To mitigate this issue, we propose to selectively use these feature maps for introducing further view-invariance capability. This is realized by incorporating the activation ordinal information [77,78], yielding another descriptor called Histogram of Ordinal Pattern (**HOP**). We similarly encode the spatial structure information by the same horizontal decomposition as HIP.

Specifically, we rank all activations $\{\boldsymbol{F}_c(i,j,\kappa)\}_{\kappa=1}^{m_c}$ in descendant order and get the top-κ feature map indices, denoted as

$$\boldsymbol{p}_c(i,j) = [v_1, v_2, \cdots, v_\kappa], \tag{1}$$

where v_i is the index of the i-th feature map in the ranked activation list. We fix $\kappa = 20$ in our experiments. By repeating the same ranking process for each image patch, we obtain a new representation $\boldsymbol{P}_c \in \mathbb{R}^{h_c \times w_c \times \kappa}$ with elements $\boldsymbol{p}_c(i,j)$. Since $\boldsymbol{P}_c$ can be considered as another set of feature maps, we utilize a similar pooling way as HIP to construct its histogram-like representation, but with bin size m_c for the c-th layer. Therefore, the feature

[1] Other networks such as the VggNet [75] and GoogLeNet [76] architectures can be considered without any limitation.

[2] More conv layers can be utilized similarly but at the cost of increasing the feature dimension size.

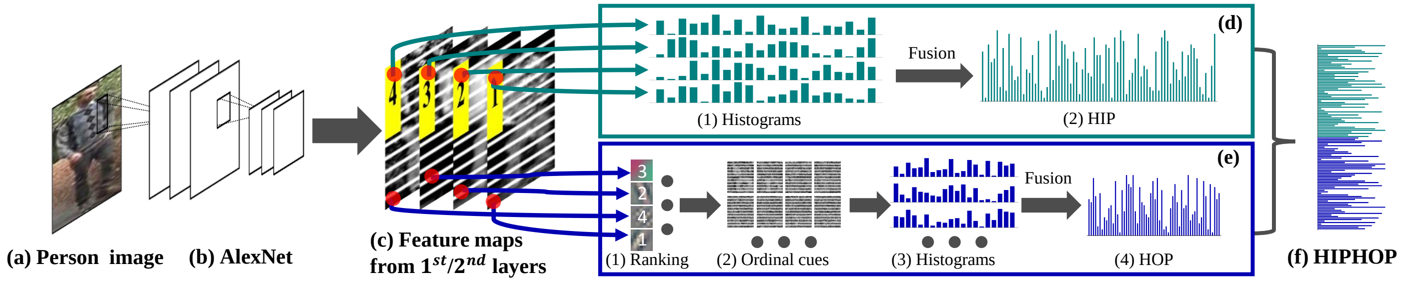


Fig. 2. Illustration of extracting the proposed HIPHOP feature. (a) A resized input person image with 227×227 pixel size. (b) Forward propagate the resized image through the whole AlexNet architecture. (c) Obtain the feature maps of the 1st and 2nd convolutional layers. (d) Compute the HIP descriptor by pooling activation intensity into histograms across different horizontal strips and feature maps. (e) Extract the HOP descriptor by ranking the localized activations and then pooling the top-κ feature map indices over all horizontal strips and feature maps. (f) Construct the final HIPHOP feature by fusion.

dimension of HOP is $46720 = 96(\text{bins}, m_1) \times 11(\text{strips}) \times 20(\kappa) + 256(\text{bins}, m_2) \times 5(\text{strips}) \times 20(\kappa)$. Together with HIP, we call our final fused person feature as **HIPHOP**, with the total dimension $84096 = 46720 + 37376$.

Feature extraction overview. We depict the main steps of constructing our HIPHOP feature. First, we resize a given image into the size of 227×227, as required by AlexNet (Figure 2(a)). Second, we forward propagate the resized image through the AlexNet (Figure 2(b)). Third, we obtain the feature maps from the 1st and 2nd conv layers (Figure 2(c)). Fourth, we compute the HIP (Figure 2(d)) and HOP (Figure 2(e)) descriptors. Finally, we composite the HIPHOP feature for the given person image by vector concatenation (Figure 2(f)). For approximately suppressing background noise, we impose an Epanechnikov kernel [46] as weight on each activation map before computing histogram.

4 Camera Correlation Aware Feature Augmentation for Re-id

We formulate a novel view-specific person re-id framework, namely **C**amera co**R**relation **A**ware **F**eature augmen**T**ation (CRAFT), to adapt the original image features into another view adaptive space, where many view-generic methods can be readily deployed for achieving view-specific discrimination modelling. In the following, we formulate re-id as a feature augmentation problem, and then present our CRAFT framework. We first discuss the re-id under two non-overlapping camera views and later generalize our model under multiple (more than two) camera views.

4.1 Re-Id Model Learning Under Feature Augmentation

Given image data from two non-overlapping camera views, namely camera a and camera b, we reformulate person re-id in a feature augmentation framework. Feature augmentation has been exploited in the domain adaptation problem. For example, Daumé III [69] proposed the feature mapping functions $\rho^s(\boldsymbol{x}) = [\boldsymbol{x}^\top, \boldsymbol{x}^\top, (\mathbf{0}_d)^\top]^\top$ (for the source domain) and $\rho^t(\boldsymbol{x}) = [\boldsymbol{x}^\top, (\mathbf{0}_d)^\top, \boldsymbol{x}^\top]^\top$ (for the target domain) for homogeneous domain adaptation, with $\boldsymbol{x} \in \mathbb{R}^d$ denoting the sample feature, $\mathbf{0}_d$ the d column vector of all zeros, d the feature dimension, and the superscript $^\top$ the transpose of a vector or a matrix. This can be viewed as incorporating the original feature into an augmented space for enhancing the similarities between data from the same domain and thus increasing the impact of same-domain (or same-camera) data. For person re-id, this should be unnecessary given its cross-camera matching nature. Without original features in augmentation, they are resorted to zero padding, a technique widely exploited in signal transmission [71,72].

Zero padding. Formally, the zero padding augmentation can be formulated as:

$$\tilde{\boldsymbol{X}}^a_{\text{zp}} = \begin{bmatrix} \boldsymbol{I}_{d\times d} \\ \boldsymbol{O}_{d\times d} \end{bmatrix} \boldsymbol{X}^a, \qquad \tilde{\boldsymbol{X}}^b_{\text{zp}} = \begin{bmatrix} \boldsymbol{O}_{d\times d} \\ \boldsymbol{I}_{d\times d} \end{bmatrix} \boldsymbol{X}^b \ , \tag{2}$$

where $\boldsymbol{X}^a = [\boldsymbol{x}^a_1, \dots, \boldsymbol{x}^a_{n_a}] \in \mathbb{R}^{d\times n_a}$ and $\boldsymbol{X}^b = [\boldsymbol{x}^b_1, \dots, \boldsymbol{x}^b_{n_b}] \in \mathbb{R}^{d\times n_b}$ represent the column-wise image feature matrix from camera a and b; $\tilde{\boldsymbol{X}}^a_{\text{zp}} = [\tilde{\boldsymbol{x}}^a_{\text{zp},1}, \dots, \tilde{\boldsymbol{x}}^a_{\text{zp},n_a}] \in \mathbb{R}^{2d\times n_a}$ and $\tilde{\boldsymbol{X}}^b_{\text{zp}} = [\tilde{\boldsymbol{x}}^b_{\text{zp},1}, \dots, \tilde{\boldsymbol{x}}^b_{\text{zp},n_b}] \in \mathbb{R}^{2d\times n_b}$ refer to augmented feature matrices; n_a and n_b are the training sample numbers of camera a and camera b, respectively; $\boldsymbol{I}_{d\times d}$ and $\boldsymbol{O}_{d\times d}$ denote the $d \times d$ identity matrix and zero matrix, respectively.

Re-id reformulation. The augmented features $\tilde{\boldsymbol{X}}^\phi_{\text{zp}}$ ($\phi \in \{a, b\}$) can be incorporated into different existing view-generic distance metric or subspace learning algorithms. Without loss of generality, we take the subspace learning as example in the follows. Specifically, the aim of discriminative learning is to estimate the optimal projections $\hat{\boldsymbol{W}} \in \mathbb{R}^{2d\times m}$ (with m the subspace dimension) such that after projection $\boldsymbol{z}^\phi = \hat{\boldsymbol{W}}^\top \tilde{\boldsymbol{x}}^\phi_{\text{zp}}$, where $\phi \in \{a, b\}$ and $\tilde{\boldsymbol{x}}^\phi_{zp}$ is an augmented feature defined in Eqn. (2), one can effectively discriminate between different identities by Euclidean distance. Generally, the objective function can be written as

$$\hat{\boldsymbol{W}} = \min_{\boldsymbol{W}} f_{\text{obj}}(\boldsymbol{W}^\top \tilde{\boldsymbol{X}}_{\text{zp}}), \tag{3}$$

where $\tilde{\boldsymbol{X}}_{\text{zp}} = [\tilde{\boldsymbol{X}}^a_{\text{zp}}, \tilde{\boldsymbol{X}}^b_{\text{zp}}] = [\tilde{\boldsymbol{x}}_{\text{zp},1}, \dots, \tilde{\boldsymbol{x}}_{\text{zp},n}]$, $n = n_a + n_b$. $\tilde{\boldsymbol{X}}_{\text{zp}}$ is the combined feature data matrix from camera a and camera b.

Clearly, $\hat{\boldsymbol{W}}$ can be decomposed into two parts as:

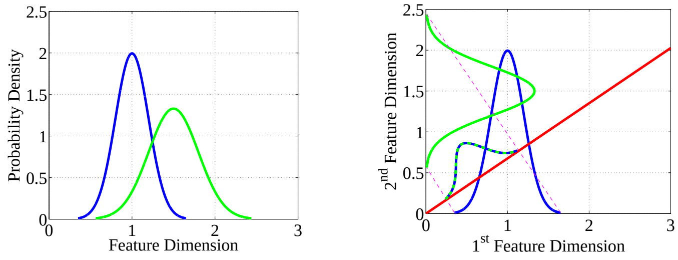
$$\hat{\boldsymbol{W}} = [(\hat{\boldsymbol{W}}^a)^\top, (\hat{\boldsymbol{W}}^b)^\top]^\top, \tag{4}$$

with $\hat{\boldsymbol{W}}^a \in \mathbb{R}^{d\times m}$ and $\hat{\boldsymbol{W}}^b \in \mathbb{R}^{d\times m}$ corresponding to the respective projections (or sub-models) for camera a and b. This is due to the zero padding based feature augmentation (Eqn. (2)):

$$\begin{aligned} \hat{\boldsymbol{W}}^\top \tilde{\boldsymbol{X}}^a_{\text{zp}} &= (\hat{\boldsymbol{W}}^a)^\top \boldsymbol{X}^a, \\ \hat{\boldsymbol{W}}^\top \tilde{\boldsymbol{X}}^b_{\text{zp}} &= (\hat{\boldsymbol{W}}^b)^\top \boldsymbol{X}^b. \end{aligned} \tag{5}$$

Clearly, zero padding allows *view-generic* methods to simultaneously learn two *view-specific* sub-models, i.e., $\hat{\boldsymbol{W}}^a$ for view a and $\hat{\boldsymbol{W}}^b$ for view b, and realize *view-specific* modelling [16,17,18], likely better aligning cross-view image data distribution [19,69].

For better understanding, we take an example in 1-D feature space. Often, re-id feature data distributions from different camera views are misaligned due to the inter-camera viewing condition discrepancy (Figure 3(a)). With labelled training data, the transformation learned in Eqn. (3) aims to search for an optimized projection in the augmented feature space (the red line in Figure



(a) Original 1-D feature space

(b) Augmented 2-D feature space

Fig. 3. An illustration of zero padding based feature augmentation. (a) The data distribution in the original feature space from camera view a (the blue curve) and b (the green curve). (b) The augmented feature space by zero padding. The dashed blue and green curves represent the projected features with respect to the projection basis indicated by the solid red line. The two dashed lines imply feature projection operation. Note that the probability density axis is not plotted in (b) for demonstration simplicity.

3(b)) such that cross-view data distributions are aligned and thus good for matching images of the same person across camera views. Clearly, the zero-padding treats each camera view independently by optimizing two separated view-specific sub-models and therefore allows to better quantify the feature distortion of either camera view. Nonetheless, as compared to a single view-generic model, this doubled modelling space may unfavourably loosen inter-camera inherent correlation (e.g., the "same" person with "different" appearance in the images captured by two cameras with distinct viewing conditions). This may in turn make the model optimization less effective in capturing appearance variation across views and extracting shared view-generic discriminative cues.

To overcome the above limitation, we design particularly the camera correlation aware feature augmentation, which allows for adaptively incorporating the common information between camera views into the augmented space whilst retaining the capability of well modelling the feature distortion of individual camera views.

4.2 Camera Correlation Aware Feature Augmentation

The proposed feature augmentation method is performed in two steps: (I) We quantify automatically the commonness degree between different camera views. (II) We exploit the estimated camera commonness information for adaptive feature augmentation.

(I) Quantifying camera commonness by correlation. We propose exploiting the correlation in image data distributions for camera commonness quantification. Considering that many different images may be generated by any camera, we represent a camera by a set of images captured by itself, e.g., *a set of feature vectors*. We exploit the available training images captured by both cameras for obtaining more reliable commonness measure.

Specifically, given image features $\boldsymbol{X}^a$ and $\boldsymbol{X}^b$ for camera a and b, respectively, we adopt the principle angles [79] to measure the correlation between the two views. In particular, first, we obtain the linear subspace representations by the principle component analysis, $\boldsymbol{G}^a \in \mathbb{R}^{n_a \times r}$ for $\boldsymbol{X}^a$ and $\boldsymbol{G}^b \in \mathbb{R}^{n_b \times r}$ for $\boldsymbol{X}^b$, with r the dominant component number. In our experiments, we empirically set $r = 100$. Either $\boldsymbol{G}^a$ or $\boldsymbol{G}^b$ can then be seen as a data point on the Grassmann manifold – a set of fixed-dimensional linear subspaces of Euclidean space. Second, we measure the similarity between the two manifold points with their principle angles ($0 \leq \theta_1 \leq \cdots \leq \theta_k \leq \cdots \leq \theta_r \leq \frac{\pi}{2}$) defined as:

$$\begin{aligned} \cos(\theta_k) = & \max_{\boldsymbol{q}_j \in \mathrm{span}(\boldsymbol{G}^a)} \max_{\boldsymbol{v}_k \in \mathrm{span}(\boldsymbol{G}^b)} \boldsymbol{q}_k^\top \boldsymbol{v}_k, \\ \text{s.t.} \quad & \boldsymbol{q}_k^\top \boldsymbol{q}_k = 1, \quad \boldsymbol{v}_k^\top \boldsymbol{v}_k = 1, \\ & \boldsymbol{q}_k^\top \boldsymbol{q}_i = 0, \quad \boldsymbol{v}_k^\top \boldsymbol{v}_i = 0, \quad i \in [0, k-1], \end{aligned} \tag{6}$$

where $\mathrm{span}(\cdot)$ denotes the subspace spanned by the column vectors of a matrix. The intuition is that, principle angles have good geometry interpretation (e.g., related to the manifold geodesic distance [80,81]) and their cosines $\cos(\theta_k)$ are known as *canonical correlations*. Finally, we estimate the camera correlation (or commonness degree) ω as:

$$\omega = \frac{1}{r} \sum_{k=1}^{r} \cos(\theta_k), \tag{7}$$

with $\cos(\theta_k)$ computed by Singular Value Decomposition:

$$(\boldsymbol{G}^a)^\top \boldsymbol{G}^b = \boldsymbol{Q} \cos(\Theta) \boldsymbol{V}^\top, \tag{8}$$

where $\cos(\Theta) = \mathrm{diag}(\cos(\theta_1), \cos(\theta_2), \cdots \cos(\theta_r))$, $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_r]$, and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r]$.

(II) Adaptive feature augmentation. Once obtaining the camera correlation ω, we want to incorporate it into feature augmentation. To achieve this, we generalize the zero padding (Eqn. (2)) to:

$$\tilde{\boldsymbol{X}}^a_{\text{craft}} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{M} \end{bmatrix} \boldsymbol{X}^a, \qquad \tilde{\boldsymbol{X}}^b_{\text{craft}} = \begin{bmatrix} \boldsymbol{M} \\ \boldsymbol{R} \end{bmatrix} \boldsymbol{X}^b \ , \tag{9}$$

where $\boldsymbol{R}$ and $\boldsymbol{M}$ refer to the $d \times d$ augmentation matrices. So, zero padding is a special case of the proposed feature augmentation (Eqn. (9)) where $\boldsymbol{R} = \boldsymbol{I}_{d \times d}$ and $\boldsymbol{M} = \boldsymbol{O}_{d \times d}$.

With some view-generic discriminative learning algorithm, we can learn an optimal model $\boldsymbol{W} = [(\boldsymbol{W}^a)^\top, (\boldsymbol{W}^b)^\top]^\top$ in our augmented space. Then, feature mapping functions can be written:

$$\begin{aligned} f_a(\tilde{\boldsymbol{X}}^a_{\text{craft}}) = \boldsymbol{W}^\top \tilde{\boldsymbol{X}}^a_{\text{craft}} = (\boldsymbol{R}^\top \boldsymbol{W}^a + \boldsymbol{M}^\top \boldsymbol{W}^b)^\top \boldsymbol{X}^a, \\ f_b(\tilde{\boldsymbol{X}}^b_{\text{craft}}) = \boldsymbol{W}^\top \tilde{\boldsymbol{X}}^b_{\text{craft}} = (\boldsymbol{M}^\top \boldsymbol{W}^a + \boldsymbol{R}^\top \boldsymbol{W}^b)^\top \boldsymbol{X}^b. \end{aligned} \tag{10}$$

Compared to zero padding (Eqn. (5)), it is clear that the feature transformation for each camera view is not treated independently in our adaptive feature augmentation (Eqn. (10)). Instead, the transformations of all camera views are intrinsically correlated and meanwhile being view-specific.

However, it is non-trivial to estimate automatically the augmentation matrices $\boldsymbol{R}$ and $\boldsymbol{M}$ with the estimated camera correlation information accommodated for enabling more accurate cross-view discriminative analysis (Sec. 4.3). This is because a large number of ($2d^2$) parameters are required to be learned given the typically high feature dimensionality d (e.g., tens of thousands) but only a small number of (e.g., hundreds) training data available. Instead of directly learning from the training data, we propose to properly design $\boldsymbol{R}$ and $\boldsymbol{M}$ for overcoming this problem as:

$$\boldsymbol{R} = \frac{2 - \omega}{\varpi} \boldsymbol{I}_{d \times d}, \qquad \boldsymbol{M} = \frac{\omega}{\varpi} \boldsymbol{I}_{d \times d}, \tag{11}$$

where ω is the camera correlation defined in Eqn. (7) and $\varpi = \sqrt{(2-\omega)^2 + \omega^2}$ is the normalization term. In this way, camera correlation is directly embedded into the feature augmentation process. Specifically, when $\omega = 0$, which means the two camera views are totally uncorrelated with no common property, we have $\boldsymbol{M} = \boldsymbol{O}_{d \times d}$ and $\boldsymbol{R} = \boldsymbol{I}_{d \times d}$, and our feature augmentation degrades to zero padding. When $\omega = 1$, which means the largest camera correlation, we have $\boldsymbol{R} = \boldsymbol{M}$ thus potentially similar view-specific sub-models, i.e., strongly correlated each other. In other words, $\boldsymbol{M}$ represents the shared degree across camera views whilst $\boldsymbol{R}$ stands for view specificity strength, with their balance controlled by the inherent camera correlation.

By using Eqn. (11), the view-specific feature mapping functions in Eqn. (10) can be expressed as:

$$\begin{aligned} f_a(\tilde{\boldsymbol{X}}^a_{\text{craft}}) &= \underbrace{\frac{2-\omega}{\varpi}(\boldsymbol{W}^a)^\top}_{\text{specificity}} \boldsymbol{X}^a + \underbrace{\frac{\omega}{\varpi}(\boldsymbol{W}^b)^\top}_{\text{adaptiveness}} \boldsymbol{X}^a, \\ f_b(\tilde{\boldsymbol{X}}^b_{\text{craft}}) &= \underbrace{\frac{\omega}{\varpi}(\boldsymbol{W}^a)^\top}_{\text{adaptiveness}} \boldsymbol{X}^b + \underbrace{\frac{2-\omega}{\varpi}(\boldsymbol{W}^b)^\top}_{\text{specificity}} \boldsymbol{X}^b. \end{aligned} \tag{12}$$

Obviously, the mapped discriminative features for each camera view depend on its respective sub-model (weighted by $\frac{2-\omega}{\varpi}$, corresponding to view-specific modelling) as well as the other sub-model (weighted by $\frac{\omega}{\varpi}$, corresponding to view-generic modelling). As such, our proposed transformations realize the joint learning of both view-generic and view-specific discriminative information. We call this *cross-view adaptive* modelling

Model formulation analysis - We examine the proposed formulation by analyzing the theoretical connection among model parameters $\{\boldsymbol{R}, \boldsymbol{M}, \boldsymbol{W}\}$. Note that in our whole modelling, the augmentation matrices $\boldsymbol{R}$ and $\boldsymbol{M}$ (Eqn. (11)) are appropriately designed with the aim for embedding the underlying camera correlation into a new feature space, whereas $\boldsymbol{W}$ (Eqn. (3)) is automatically learned from the training data. Next, we demonstrate that learning $\boldsymbol{W}$ alone is sufficient to obtain the optimal solution.

Formally, we denote the optimal augmentation matrices:

$$\boldsymbol{R}^{\text{opt}} = \boldsymbol{R} + \nabla\boldsymbol{R}, \quad \boldsymbol{M}^{\text{opt}} = \boldsymbol{M} + \nabla\boldsymbol{M}, \tag{13}$$

with $\nabla\boldsymbol{R}$ the difference (e.g., the part learned from the training data by some ideal algorithm) between our designed $\boldsymbol{R}$ and the assumed optimal one $\boldsymbol{R}^{\text{opt}}$ (similarly for $\nabla\boldsymbol{M}$). The multiplication operation between $\boldsymbol{M}$ (or $\boldsymbol{R}$) and $\boldsymbol{W}$ in Eqn. (12) suggests that

$$\begin{cases} (\boldsymbol{R}+\nabla\boldsymbol{R})^\top\boldsymbol{W}^a + (\boldsymbol{M}+\nabla\boldsymbol{M})^\top\boldsymbol{W}^b \\ \quad = \boldsymbol{R}^\top(\boldsymbol{W}^a + \nabla\boldsymbol{W}^a) + \boldsymbol{M}^\top(\boldsymbol{W}^b + \nabla\boldsymbol{W}^b) \\ (\boldsymbol{M}+\nabla\boldsymbol{M})^\top\boldsymbol{W}^a + (\boldsymbol{R}+\nabla\boldsymbol{R})^\top\boldsymbol{W}^b \\ \quad = \boldsymbol{M}^\top(\boldsymbol{W}^a + \nabla\boldsymbol{W}^a) + \boldsymbol{R}^\top(\boldsymbol{W}^b + \nabla\boldsymbol{W}^b) \end{cases}, \tag{14}$$

where $\nabla\boldsymbol{W}^a$ and $\nabla\boldsymbol{W}^b$ are:

$$\begin{pmatrix} \nabla\boldsymbol{W}^a \\ \nabla\boldsymbol{W}^b \end{pmatrix} = \begin{pmatrix} \boldsymbol{R}^\top & \boldsymbol{M}^\top \\ \boldsymbol{M}^\top & \boldsymbol{R}^\top \end{pmatrix}^{-1} \begin{pmatrix} \nabla\boldsymbol{R} & \nabla\boldsymbol{M} \\ \nabla\boldsymbol{M} & \nabla\boldsymbol{R} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{W}^a \\ \boldsymbol{W}^b \end{pmatrix}. \tag{15}$$

Eqn. (15) indicates that the learnable parts $\nabla\boldsymbol{R}$ and $\nabla\boldsymbol{M}$ can be equivalently obtained in the process of optimizing $\boldsymbol{W}$. This suggests no necessity of directly learning $\boldsymbol{R}$ and $\boldsymbol{M}$ from the training data as long as $\boldsymbol{W}$ is inferred through some optimization procedure. Hence, deriving $\boldsymbol{R}$ and $\boldsymbol{M}$ as in Eqn. (11) should not degrade the effectiveness of our whole model, but instead making our entire formulation design more tractable and more elegant with different components fulfilling a specific function.

4.3 Camera View Discrepancy Regularization

As aforementioned in Eqn. (10), our transformations of all camera views are not independent to each other. Although these transformations are view-specific, they are mutually correlated in practice because they quantify the same group of people for association between camera views. View-specific modelling (Eqns. (4) and (12)) allows naturally for regularizing the mutual relation between different sub-models, potentially incorporating complementary correlation modelling between camera views in addition to Eqn. (10). To achieve this, we enforce the constraint on sub-models by introducing a Camera View Discrepancy (CVD) regularization as:

$$\gamma_{\text{cvd}} = ||\boldsymbol{W}^a - \boldsymbol{W}^b||^2, \tag{16}$$

Moreover, this CVD constraint can be combined well with the common ridge regularization as:

$$\begin{aligned} \gamma &= ||\boldsymbol{W}^a - \boldsymbol{W}^b||^2 + \eta_{\text{ridge}}\text{tr}(\boldsymbol{W}^\top\boldsymbol{W}) \\ &= \text{tr}(\boldsymbol{W}^\top \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix} \boldsymbol{W}) + \eta_{\text{ridge}}\text{tr}(\boldsymbol{W}^\top\boldsymbol{W}) \\ &= (1+\eta_{\text{ridge}})\text{tr}(\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W}), \end{aligned} \tag{17}$$

where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I} & -\beta\boldsymbol{I} \\ -\beta\boldsymbol{I} & \boldsymbol{I} \end{bmatrix}, \quad \beta = \frac{1}{1+\eta_{\text{ridge}}}.$$

$\text{tr}(\cdot)$ denotes the trace operation, and $\eta_{\text{ridge}} > 0$ is a tradeoff parameter for balancing the two terms. The regularization γ can be readily incorporated into existing learning methods [10,12,13,22,82] for possibly obtaining better model generalization. Specifically, we define an enriched objective function on top of Eqn. (3) as:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} f_{\text{obj}}(\boldsymbol{W}^\top\tilde{\boldsymbol{X}}_{\text{craft}}) + \lambda\text{tr}(\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W}), \tag{18}$$

where λ controls the influence of γ. Next, we derive the process for solving Eqn. (18).

View discrepancy regularized transformation. Since $\beta = \frac{1}{1+\eta_{\text{ridge}}} < 1$, the matrix $\boldsymbol{C}$ is of positive-definite. Therefore, $\boldsymbol{C}$ can be factorized into the form of

$$\boldsymbol{C} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top, \tag{19}$$

with $\boldsymbol{\Lambda}$ a diagonal matrix and $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}$. So, $\boldsymbol{P}^\top\boldsymbol{C}\boldsymbol{P} = \boldsymbol{\Lambda}$. By defining

$$\boldsymbol{W} = \boldsymbol{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{H}, \tag{20}$$

we have

$$\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W} = \boldsymbol{H}^\top\boldsymbol{H}. \tag{21}$$

Thus, Eqn. (18) can be transformed equivalently to:

$$\hat{\boldsymbol{H}} = \arg\min_{\boldsymbol{H}} f_{\text{obj}}(\boldsymbol{H}^\top\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}^\top\tilde{\boldsymbol{X}}_{\text{craft}}) + \lambda\text{tr}(\boldsymbol{H}^\top\boldsymbol{H}). \tag{22}$$

We define the transformed data matrix from all views

$$\ddot{\boldsymbol{X}}_{\text{craft}} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}^\top\tilde{\boldsymbol{X}}_{\text{craft}} = [\ddot{\boldsymbol{x}}_1, \cdots, \ddot{\boldsymbol{x}}_n], \tag{23}$$

which we call *view discrepancy regularized transformation*. So, Eqn. (22) can be simplified as:

$$\hat{\boldsymbol{H}} \quad = \arg\min_{\boldsymbol{H}} f_{\text{obj}}(\boldsymbol{H}^\top\ddot{\boldsymbol{X}}_{\text{craft}}) + \lambda\text{tr}(\boldsymbol{H}^\top\boldsymbol{H}). \tag{24}$$

Optimization. Typically, the same optimization algorithm as the adopted view-generic discriminative learning method can be exploited to solve the optimization problem. For providing a complete picture, we will present a case study with a specific discriminative learning method incorporated into the proposed CRAFT framework.

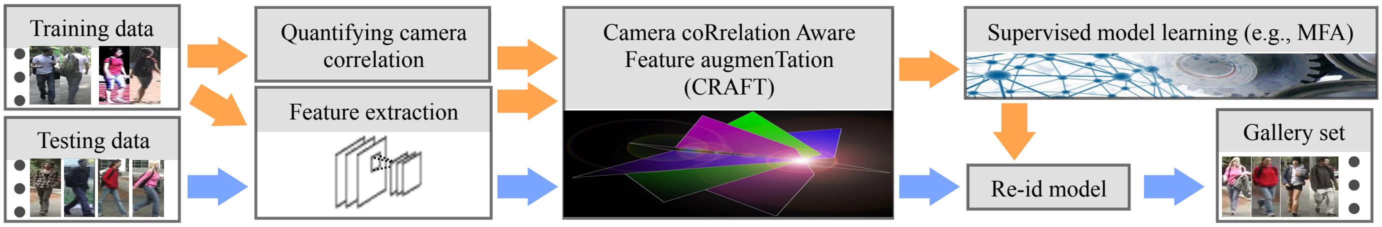

Training data
Testing data
Quantifying camera correlation
Feature extraction
Camera coRrelation Aware Feature augmenTation (CRAFT)
Supervised model learning (e.g., MFA)
Re-id model
Gallery set

Fig. 4. Pipeline of the proposed person re-id approach. Training state is indicated with orange arrows, and testing stage with blue arrows.

```
Algorithm 1: Learning CRAFT-MFA
  Input: Training data $\boldsymbol{X}^a$ and $\boldsymbol{X}^b$ with identity labels;
  Output: Augmentation matrices $\boldsymbol{R}$ and $\boldsymbol{M}$, projection matrix $\hat{\boldsymbol{W}}$;
1  (I) Camera correlation aware feature augmentation (Section 4.2)
2  - Estimate the camera correlation $\omega$ (Eqn. (7));
3  - Compute augmentation matrix $\boldsymbol{R}$ and $\boldsymbol{M}$ (Eqn. (11));
4  - Transform original features $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}}_{\text{craft}}$ (Eqn. (9));
5  (II) Camera view discrepancy regularization (Section 4.3)
6  - Obtain the camera view discrepancy regularization (Eqn. (16));
7  - Get the fused regularization; (Eqn. (17));
8  - Decompose $\boldsymbol{C}$ into $\boldsymbol{P}$ and $\boldsymbol{\Lambda}$ (Eqn. (19));
9  - Transform $\tilde{\boldsymbol{X}}_{\text{craft}}$ into $\ddot{\boldsymbol{X}}_{\text{craft}}$ (Eqn. (23));
10 (III) Optimization
11 - Obtain $\hat{\boldsymbol{H}}$ with the MFA algorithm (Eqn. (25));
12 - Calculate $\tilde{\boldsymbol{W}}$ with $\boldsymbol{P}$, $\boldsymbol{\Lambda}$, and $\hat{\boldsymbol{H}}$ (Eqn. (20)).
```

4.4 CRAFT Instantialization

In our CRAFT framework, we instantilize a concrete person re-id method using the Marginal Fisher Analysis (MFA) [22], due to its several important advantages over the canonical Linear Discriminant Analysis (LDA) model [82]: (1) no strict assumption on data distribution and thus more general for discriminative learning, (2) a much larger number of available projection directions, and (3) a better capability of characterizing inter-class separability. We call this method instance as "CRAFT-MFA". Algorithm 1 presents an overview of learning the CRAFT-MFA model.

Specifically, we consider each person identity as an individual class. That is, all images of the same person form the whole same-class samples, regardless of being captured by either camera. Formally, given the training data $\boldsymbol{X} = [\boldsymbol{X}^a, \boldsymbol{X}^b] = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$ with $n = n_a + n_b$, we first transform them to $\ddot{\boldsymbol{X}}_{\text{craft}} = [\ddot{\boldsymbol{x}}_1, \cdots, \ddot{\boldsymbol{x}}_n]$ (Lines 1-9 in Algorithm 1). $\hat{\boldsymbol{H}}$ can then be obtained by solving the MFA optimization problem (Line 11 in Alg. 1):

$$\begin{aligned} &\min_{\boldsymbol{H}} \sum_{i\neq j} \boldsymbol{A}^c_{ij} ||\boldsymbol{H}^\top(\ddot{\boldsymbol{x}}_i - \ddot{\boldsymbol{x}}_j)||_2^2 + \lambda \text{tr}(\boldsymbol{H}^\top \boldsymbol{H}) \\ &\text{s.t.} \sum_{i\neq j} \boldsymbol{A}^p_{ij} ||\boldsymbol{H}^\top(\ddot{\boldsymbol{x}}_i - \ddot{\boldsymbol{x}}_j)||_2^2 = 1, \end{aligned} \tag{25}$$

where the element $\boldsymbol{A}^c_{ij}$ of the intrinsic graph $\boldsymbol{A}^c$ is:

$$\boldsymbol{A}^c_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } i \in N^+_{k_1}(j) \text{ or } j \in N^+_{k_1}(i) \\ 0 & \text{otherwise} \end{array} \right. ,$$

with $N^+_{k_1}(i)$ denoting the set of k_1 nearest neighbor indices of sample $\boldsymbol{x}_i$ in the same class. And the elements $\boldsymbol{A}^p_{ij}$ of the penalty graph $\boldsymbol{A}^p$ are defined as:

$$\boldsymbol{A}^p_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } (i,j) \in P_{k_2}(y_i) \text{ or } (i,j) \in P_{k_2}(y_j) \\ 0 & \text{otherwise} \end{array} \right. ,$$

where y_i and y_j refer to the class/identity label of the i^{th} and j^{th} sample, respectively, $P_{k_2}(y_i)$ indicates the set of data pairs that are k_2 nearest pairs among $\{(i,j)|y_i \neq y_j\}$. Finally, we compute the optimal $\hat{\boldsymbol{W}}$ with Eqn. (20) (Line 12 in Algorithm 1). Note that this process generalizes to other view-generic discriminative learning algorithms [10,12,13,82] (see evaluations in Table 4).

4.5 Kernel Extension

The objective function Eqn. (24) assumes linear projection. However, given complex variations in viewing condition across cameras, the optimal subspace for person re-id may not be obtainable by linear models. Thus, we further kernelize our feature augmentation (Eqn. (9)) by projecting the original feature data into a reproducing kernel Hilbert space $\mathcal{H}$ with an implicit function $\phi(\cdot)$. The inner-product of two data points in $\mathcal{H}$ can be computed by a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. In our evaluation, we utilized the nonlinear Bhattacharyya kernel function due to (1) its invariance against any non-singular linear augmentation and translation and (2) its additional consideration of data distribution variance and thus more reliable [83]. We denote $\boldsymbol{k}(\boldsymbol{x})$ as the kernel similarity vector of a sample $\boldsymbol{x}$, obtained by:

$$\boldsymbol{k}(\boldsymbol{x}) \quad = [k(\boldsymbol{x}_1, \boldsymbol{x}), k(\boldsymbol{x}_2, \boldsymbol{x}), \cdots, k(\boldsymbol{x}_n, \boldsymbol{x})]^\top, \tag{26}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ are all samples from all views, and $n = n_a + n_b$ is the number of training samples. Then the mapping function can be expressed as:

$$f_{\text{ker}}(\boldsymbol{x}) = \boldsymbol{U}^\top \phi(\boldsymbol{X})^\top \phi(\boldsymbol{x}) = \boldsymbol{U}^\top \boldsymbol{k}(\boldsymbol{x}), \tag{27}$$

where $\boldsymbol{U} \in \mathbb{R}^n$ represents the parameter matrix to be learned. The superscript for camera id is omitted for simplicity. Conceptually, Eqn. (27) is similar to the linear case Eqn. (12) if we consider $\boldsymbol{k}(\boldsymbol{x})$ as a feature representation of $\boldsymbol{x}$. Hence, by following Eqn. (9), the kernelized version of our feature augmentation can be represented as:

$$\tilde{\boldsymbol{k}}^a_{\text{craft}}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{M} \end{bmatrix} \boldsymbol{k}^a(\boldsymbol{x}), \quad \tilde{\boldsymbol{k}}^b_{\text{craft}}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{M} \\ \boldsymbol{R} \end{bmatrix} \boldsymbol{k}^b(\boldsymbol{x}), \tag{28}$$

where $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ are the same as in Eqn. (11) but with a different dimension, $\boldsymbol{k}^a(\boldsymbol{x})$ and $\boldsymbol{k}^b(\boldsymbol{x})$ denote the sample kernel similarity vectors for camera a and b, respectively. Analogously, both view discrepancy regularization (Eqn. (17)) and model optimization (Eqn. (24)) can be performed similarly as in the linear case.

4.6 Extension to More than Two Camera Views

There often exist multiple (more than two) cameras deployed across a typical surveillance network. Suppose there are $J(> 2)$ non-overlapping camera views. Therefore, person re-id across the whole network is realistically critical, but joint quantification on person association across multiple camera views is largely under-studied in the current literature [84,85]. Compared to camera pair based re-id above, this is a more challenging situation due to: (1) intrinsic difference between distinct camera pairs in viewing conditions which makes the general cross-camera feature mapping more complex and difficult to learn; (2) quantifying simultaneously the correlations of multiple camera pairs is non-trivial in both formulation and computation. In this work, we propose to address this challenge by jointly learning adaptive view-specific re-id models for all cameras in a unified fashion.

Specifically, we generalize our camera correlation aware feature augmentation (Eqn. (9)) into multiple (J) camera cases as:

$$\tilde{\boldsymbol{x}}^{\phi}_{\text{craft}} = [\underbrace{\boldsymbol{M}_{i,1}, \boldsymbol{M}_{i,2}, \cdots, \boldsymbol{M}_{i,i-1}}_{\#:\, i-1}, \boldsymbol{R}_i, \underbrace{\boldsymbol{M}_{i,i+1}, \boldsymbol{M}_{i,i+2}, \cdots, \boldsymbol{M}_{i,J}}_{\#:\, J-i}]^{\top}\boldsymbol{x}^{\phi}, \tag{29}$$

with

$$\boldsymbol{M}_{i,j} = \frac{\omega_{i,j}}{\varpi_i}\boldsymbol{I}_{d\times d}, \tag{30}$$

where $\omega_{i,j}$ denotes the correlation between camera i and j, estimated by Eqn. (7). Similar to Eqn. (11), we design

$$\boldsymbol{R}_i = \frac{2-\frac{1}{J-1}\sum_{j\neq i}\omega_{i,j}}{\varpi_i} \quad , \tag{31}$$

with

$$\varpi_i = \sqrt{(2-\frac{1}{J-1}\sum_{j\neq i}\omega_{i,j})^2 + \sum_{j\neq i}\omega_{i,j}^2} \quad . \tag{32}$$

Similarly, we extend the CVD regularization Eqn. (16) to:

$$\gamma_{\text{cvd}} = \sum_{i,j\in\{1,2,\dots,J\} \text{ and } i\neq j} ||\boldsymbol{W}^i - \boldsymbol{W}^j||^2. \tag{33}$$

Thus, the matrix $\boldsymbol{C}$ of γ (Eqn. (17)) is expanded as:

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{I} & -\beta'\boldsymbol{I} & -\beta'\boldsymbol{I} & \cdots & -\beta'\boldsymbol{I} \\ -\beta'\boldsymbol{I} & \boldsymbol{I} & -\beta'\boldsymbol{I} & \cdots & -\beta'\boldsymbol{I} \\ \vdots & \vdots & \vdots & \vdots & \\ -\beta'\boldsymbol{I} & -\beta'\boldsymbol{I} & -\beta'\boldsymbol{I} & \cdots & \boldsymbol{I} \end{pmatrix}, \tag{34}$$

where $\beta' = \frac{\beta}{J-1}$. The following view discrepancy regularized transformation and model optimization can be carried out in the same way as in Sec. 4.3. Clearly, re-id between two cameras is a special case of the extended model when $J = 2$. Therefore, our extended CRAFT method allows to consider and model the intrinsic correlation between camera views in the entire network.

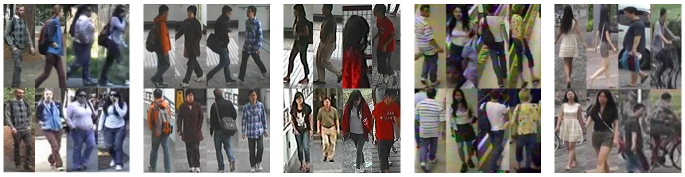
4.7 Person Re-identification by CRAFT

Once a discriminative model ($\boldsymbol{W}$ or $\boldsymbol{U}$) is learned from the training data using the proposed method, we can deploy it for person re-id. Particularly, first, we perform feature augmentation to transform all original features $\{\boldsymbol{x}\}$ to the CRAFT space $\{\tilde{\boldsymbol{x}}_{\text{craft}}\}$ (Eqn. (29)). Second, we match a given probe person $\tilde{\boldsymbol{x}}^p_{\text{craft}}$ from one camera against a set of gallery people $\{\tilde{\boldsymbol{x}}^g_{\text{craft},i}\}$ from another camera by computing the Euclidean distance $\{\text{dist}(\tilde{\boldsymbol{x}}^p_{\text{craft}}, \tilde{\boldsymbol{x}}^g_{\text{craft},i})\}$ in the prediction space (induced by Eqn. (12) or (27)). Finally, the gallery people are sorted in ascendant order of their assigned distances to generate the ranking list. Ideally, the true match(es) can be found among a few top ranks. The pipeline of our proposed re-id approach is depicted in Figure 4.

5 EXPERIMENTS

5.1 Datasets and Evaluation Settings

Datasets. We extensively evaluated the proposed approach on five person re-id benchmarks: VIPeR [30], CUHK01 [31], CUHK03 [26], QMUL GRID [32], and Market-1501 [33]. All datasets are very challenging due to unknown large cross-camera divergence in viewing conditions, e.g., illumination, viewpoint, occlusion and background clutter (Figure 5). The *VIPeR* dataset consists of 1264 images from 632 persons observed from two surveillance cameras with various viewpoints and background settings. As these images are of low spatial resolution, it is very difficult to extract reliable appearance features (Figure 5(a)). The *CUHK01* dataset consists of 971 people observed from two outdoor cameras. Each person has two samples per camera view. Compared with VIPeR, this dataset has higher spatial resolution and thus more appearance information are preserved in images (Figure 5(b)). The *CUHK03* dataset consists of 13164 images from 1360 people collected from six non-overlapping cameras. In evaluation, we used the automatically detected person images which represent a more realistic yet challenging deployment scenario, e.g., due to more severe misalignment caused by imperfect detection (Figure 5(c)). The *QMUL GRID* dataset consists of 250 people image pairs from eight different camera views in a busy underground station. Unlike all the three datasets above, there are 775 extra identities or imposters. All images of this dataset are featured with lower spatial resolution and more drastic illumination variation (Figure 5(d)). The *Market-1501* dataset contains person images collected in front of a campus supermarket at Tsinghua University. A total of six cameras were used, with five high-resolution and one low-resolution. This dataset consists of 32,668 annotated bounding boxes of 1501 people (Figure 5(e)).



(a) VIPeR (b) CUHK01 (c) CUHK03 (d) GRID (e) Market

Fig. 5. Example images from different person re-id datasets. For every dataset, two images in a column correspond to the same person.

Evaluation protocol. For all datasets, we followed the standard evaluation settings for performing a fair comparison with existing methods as below: **(I)** On the VIPeR, CUHK01 and QMUL GRID benchmarks, we split randomly the whole people into two halves: one for training and one for testing. The cumulative matching characteristic (CMC) curve was utilized to measure the performance of the compared methods. As CUHK01 is a multi-shot (e.g., multiple images per person per camera view) dataset, we computed the final matching distance between two people by averaging corresponding cross-view image pairs. We repeated the experiments for 10 trials and reported the average results. **(II)** On CUHK03, we followed the standard protocol [26] – repeating 20 times of random 1260/100 people splits for model training/test and comparing the averaged matching results. **(III)** On Market-1501, we utilized the standard training (750) and testing (751) people split provided by the authors of [33]. Apart from CMC, we also used other two performance metrics: (1) Rank-1 accuracy, and (2) mean Average Precision (mAP), i.e. first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes.

5.2 Evaluating Our Proposed Re-Id Approach

We evaluated and analyzed the proposed person re-id approach in these aspects: (1) Effect of our CRAFT framework (Eqns. (9) and (11)); (2) Comparison between CRAFT and domain adaptation; (3) Effect of our CVD regularization γ_{cvd} (Eqn. (16)); (4) Generality of CRAFT instantialization; (5) Effect of our HIPHOP feature; (6) Complementary of HIPHOP on existing popular features.

TABLE 1
Comparing top matching ranks (%) on VIPeR and CUHK01. The 1^{st} and 2^{nd} best results are indicated in red and blue color respectively.

Dataset	VIPeR [30]				CUHK01 [31]				CUHK03 [26]				Market-1501 [33]				QMUL GRID [32]			
Rank (%)	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
OriFeat	43.3	72.7	84.1	93.4	64.3	85.1	90.6	94.6	63.4	88.0	93.0	96.1	65.4	84.0	89.3	93.1	21.0	42.9	53.0	62.7
+ Kernelization	47.0	75.4	86.8	94.4	69.5	89.3	93.5	96.5	78.6	94.9	96.8	98.4	66.0	84.4	89.3	93.2	19.0	42.2	51.9	61.4
ZeroPad [71]	37.5	69.9	82.8	92.1	66.4	85.8	90.5	94.7	76.0	91.9	94.8	95.3	38.2	62.5	71.7	80.0	7.2	26.0	40.3	55.8
+ Kernelization	40.0	72.8	85.0	93.4	71.8	89.6	93.8	96.5	80.0	92.7	94.4	95.3	49.5	72.4	80.0	85.8	6.1	21.8	36.3	51.4
BaseFeatAug [69]	45.5	76.1	87.4	95.1	59.0	81.1	87.0	92.4	78.3	94.6	97.3	98.9	65.3	83.6	88.8	92.6	20.1	47.6	58.8	70.0
+ Kernelization	47.3	77.8	89.0	95.2	63.0	83.5	89.0	93.6	83.4	97.0	98.1	99.1	65.3	83.6	88.8	92.6	20.1	47.6	58.8	70.0
CRAFT	47.8	77.1	87.8	95.1	70.0	87.4	92.0	95.5	78.5	94.7	97.5	98.9	67.9	85.1	90.0	93.4	25.4	50.2	61.8	74.2
+ Kernelization	50.3	80.0	89.6	95.5	74.5	91.2	94.8	97.1	84.3	97.1	98.3	99.1	68.7	87.1	90.8	94.0	22.4	49.9	61.8	71.7

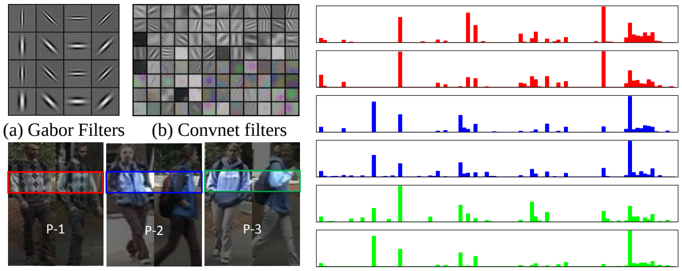
Effect of our CRAFT framework. We evaluated the proposed CRAFT method by comparing with (a) *baseline feature augmentation* (BaseFeatAug) [69], (b) *zero padding* (ZeroPad) [71], and (c) *original features* (OriFeat). Their kernelized versions using the Bhattacharyya kernel function were also evaluated. For fair comparison, our CRAFT-MFA method is utilized for re-id model learning in all compared methods.

Table 1 shows the re-id results. It is evident that our CRAFT approach outperformed consistently each baseline method on all the four re-id datasets, either using kernel or not. For example, CRAFT surpasses OriFeat, ZeroPad and BaseFeatAug by $4.5\%/5.7\%/15.1\%/2.5\%/4.4\%$, $10.3\%/3.6\%/2.5\%/29.7\%/18.2\%$, $2.3\%/11.0\%/0.2\%/2.6\%$ $/5.3\%$ at rank-1 on VIPeR/CUHK01/CUHK03/Market-1501/QMUL GRID, respectively. Similar improvements were observed in the kernelized case. It is also found that, without feature augmentation, OriFeat produced reasonably good person matching accuracies, whilst both ZeroPad and BaseFeatAug may deteriorate the re-id performance. This plausible reasons are: (1) The small training data size may lead to some degree of model overfitting given two or three times as many parameters as OriFeat; and (2) Ignoring camera correlation can result in sub-optimal discrimination models. The re-id performance can be further boosted using the kernel trick in most cases except QMUL GRID. This exception may be due to the per-camera image imbalance problem on this dataset, e.g., 25 images from the 6th camera *versus* 513 images from the 5th camera. In the following, we used the kernel version of all methods unless otherwise stated.

Comparison between CRAFT and domain adaptation. We compared our CRAFT with two representative domain adaptation models: TCA [63] and TFLDA [64]. It is evident from Table 3 that the proposed CRAFT always surpasses TCA and TFLDA with a clear margin in the re-id performance on all datasets. This is because: (1) TCA is not discriminant and thus yielded much poor accuracies; and (2) both TCA and TFLDA assume that the target domain shares the same class labels as the source domain, which however is not valid in re-id applications. This suggests the advantage and superiority of our cross-view adaptive modelling over conventional domain adaptation applied to person re-id.

Effect of our CVD regularization. For evaluating the exact impact of the proposed regularization γ_{cvd} (Eqns. (16) and (33)) on model generalization, we compared the re-id performance of our full model with a stripped down variant without γ_{cvd} during model optimization, called "CRAFT(no γ_{cvd})". The results in Table 2 demonstrate the usefulness of incorporating γ_{cvd} for re-id. This justifies the effectiveness of our CVD regularization in controlling the correlation degree between view-specific sub-models.

Generality of CRAFT instantialization. We evaluated the generality of CRAFT instantiation by integrating different supervised learning models. Five popular state-of-the-art methods were considered: (1) MFA [22], (2) LDA [82], (3) KISSME [13], (4) LFDA [10], (5) XQDA [12]. Our kernel feature was adopted.



(a) Gabor Filters (b) Convnet filters
(c) Person image pairs (d) Corresponding strip level HOP features

Fig. 6. Illustration of our HIP and HOP person descriptors. (a) Gabor filters used in ELF18 [19]. (b) Convolutional filters from the 1^{st} AlexNet layer. (c) Horizontal stripes of 6 images from 3 different people (P-1, P-2 and P-3). (d) HOP histograms extracted from the corresponding image strips (i.e., indicated with a rectangular of the same color as histogram bars) in (c). Partial HOP descriptors are shown for clear visualization.

The results are given in Table 4. It is evident that our CRAFT framework is general for incorporating different existing distance metric learning algorithms. Specifically, we find that CRAFT achieved better re-id results than other feature augmentation competitors with any of these learning algorithms. The observation also suggests the consistent superiority and large flexibility of the proposed approach over alternatives in learning discriminative re-id models. We utilize CRAFT-MFA for the following comparisons with existing popular person re-id methods.

Effect of our HIPHOP feature. We compared the proposed HIPHOP feature with several extensively adopted re-id representations: (1) ELF18 [19], (2) ColorLBP [54], (3) WHOS [46], and (4) LOMO [12]. The Bhattacharyya kernel was utilized for all compared visual features.

The re-id results are reported in Table 5. It is evident that our HIPHOP feature is overall much more effective than the compared alternatives for person re-id. For example, using HIPHOP improved the rank-1 rate from 42.3% by LOMO to 50.3% on VIPeR, from 66.0% by WHOS to 74.5% on CUHK01, from 77.9% by LOMO to 84.3% on CUHK03, and from 52.2% by WHOS to 68.7% on Market-1501. This suggests the great advantage and effectiveness of more view invariant appearance representation learned from diverse albeit generic auxiliary data. A few reasons are: (1) By exploiting more complete and diverse filters (Figure 6(b)) than ELF18 (Figure 6(a)), more view change tolerant person appearance details shall be well encoded for discriminating between visually alike people; (2) By selecting salient patterns, our HOP descriptor possesses some features more tolerant to view variety, which is critical to person re-id due to the potential cross-view pose and viewing condition variation. This is clearly illustrated in Figure 6(c-d): (i) Cross-view images of the same person have similar HOP patterns; (ii) Visually similar people (P-2, P-3) share more commonness than visually distinct people (P-1, P-2) in HOP histogram. Besides, we evaluated the discrimination power of features from different conv layers in Table 5. More specifically, it is shown in Table 5 that the $1^{st}/2^{nd}$ conv layer based features (i.e., HIP/HOP, being low-level)

TABLE 2
Evaluating the effect of our CVD regularization.

Dataset	VIPeR [30]				CUHK01 [31]				CUHK03 [26]				Market-1501 [33]				QMUL GRID [32]			
Rank (%)	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
CRAFT(no γ_{cvd})	46.3	77.9	88.1	95.4	73.8	90.6	94.2	96.9	83.9	97.0	98.2	**99.1**	66.6	85.9	90.7	93.7	15.8	45.0	57.7	60.0
CRAFT	**50.3**	**80.0**	**89.6**	**95.5**	**74.5**	**91.2**	**94.8**	**97.1**	**84.3**	**97.1**	**98.3**	**99.1**	**68.7**	**87.1**	**90.8**	**94.0**	**22.4**	**49.9**	**61.8**	**71.7**

TABLE 3
Comparison between CRAFT and domain adaptation.

Dataset	VIPeR [30]				CUHK01 [31]				CUHK03 [26]				Market-1501 [33]				QMUL GRID [32]			
Rank (%)	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
TCA [63]	11.1	23.4	31.0	38.5	7.0	16.4	22.2	30.1	5.5	16.2	26.4	42.8	8.9	18.7	24.1	30.1	9.8	22.2	29.8	38.3
TFLDA [64]	46.4	75.8	86.7	93.9	69.6	88.7	92.8	96.2	76.7	94.4	96.5	98.0	62.5	81.3	87.0	91.6	19.5	42.5	51.6	61.8
CRAFT	**50.3**	**80.0**	**89.6**	**95.5**	**74.5**	**91.2**	**94.8**	**97.1**	**84.3**	**97.1**	**98.3**	**99.1**	**68.7**	**87.1**	**90.8**	**94.0**	**22.4**	**49.9**	**61.8**	**71.7**

TABLE 4
Evaluating the generality of CRAFT instantialization. In each row, the $1^{\text{st}}/2^{\text{nd}}$ best results (%) for each rank are indicated in red / blue color.

	Method	**CRAFT**				OriFeat				ZeroPad [71]				BaseFeatAug [69]			
	Rank (%)	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
VIPeR	MFA [22]	50.3	80.0	89.6	95.5	47.0	75.4	86.8	94.4	40.0	72.8	85.0	93.4	47.3	77.8	89.0	95.2
	LDA [82]	49.3	79.1	89.2	95.3	46.4	75.0	86.5	94.1	40.0	73.8	84.7	93.3	47.4	78.0	88.9	95.3
	KISSME [13]	50.1	79.1	88.9	95.0	46.3	75.1	86.5	94.0	38.7	71.1	83.8	92.3	48.4	77.6	88.7	94.9
	LFDA [10]	49.3	79.1	89.2	95.3	46.4	75.0	86.5	94.1	40.0	73.8	84.7	93.3	47.4	77.7	88.8	95.2
	XQDA [12]	50.3	79.1	88.9	95.0	46.3	75.1	86.5	94.0	38.7	71.1	83.8	92.3	47.1	76.9	88.9	94.7
CUHK01	MFA [22]	74.5	91.2	94.8	97.1	69.5	89.3	93.5	96.5	71.8	89.6	93.8	96.5	63.0	83.5	89.0	93.6
	LDA [82]	73.8	90.3	93.9	96.6	69.4	88.3	92.8	96.1	71.3	88.8	92.9	96.1	63.0	83.5	88.9	93.4
	KISSME [13]	73.0	89.6	93.6	96.3	69.2	87.6	92.5	95.8	67.3	85.9	90.8	94.7	63.6	83.0	88.0	92.9
	LFDA [10]	73.8	90.3	93.9	96.6	69.4	88.3	92.8	96.1	71.3	88.8	92.9	96.1	62.6	83.1	88.5	93.3
	XQDA [12]	73.0	89.5	93.6	96.3	69.1	87.6	92.6	95.8	69.9	87.3	92.1	95.5	61.3	82.2	87.8	93.1
CUHK03	MFA [22]	84.3	97.0	98.3	99.1	78.6	94.9	96.8	98.3	80.0	92.7	94.4	95.3	83.4	97.0	98.1	99.1
	LDA [82]	80.2	96.6	98.2	99.0	76.6	94.6	96.6	98.0	79.5	94.7	96.4	98.3	80.0	96.2	97.3	99.0
	KISSME [13]	76.2	93.7	96.9	98.5	64.7	88.7	93.4	96.2	73.6	92.5	95.9	98.0	72.4	91.6	95.3	97.8
	LFDA [10]	80.9	96.4	98.3	99.0	76.3	93.5	96.3	97.8	79.7	95.6	97.4	98.0	80.4	96.3	98.2	98.9
	XQDA [12]	79.8	96.0	98.0	99.0	78.7	94.3	97.4	98.7	78.3	95.8	97.4	98.2	79.2	95.5	97.3	98.1
GRID	MFA [22]	22.4	49.9	61.8	71.7	19.0	42.2	51.9	61.4	6.1	21.8	36.3	51.4	17.0	45.8	58.2	67.9
	LDA [82]	22.1	50.1	61.6	71.0	19.0	42.2	51.7	61.7	6.6	23.4	37.3	50.7	17.4	46.6	57.6	68.3
	KISSME [13]	22.4	50.4	61.4	71.5	19.5	41.8	51.6	61.0	5.5	21.4	35.8	50.8	17.0	46.5	57.4	67.5
	LFDA [10]	22.2	50.1	61.5	71.0	19.0	42.0	51.6	61.7	6.6	23.4	37.3	50.7	17.4	46.6	57.6	68.3
	XQDA [12]	22.3	50.9	61.6	71.4	19.8	42.5	51.7	61.8	6.1	22.7	36.5	51.4	17.0	46.5	57.4	67.8
Market	MFA [22]	68.7	87.1	90.8	94.0	66.0	84.4	89.3	93.2	49.5	72.4	80.0	85.8	65.5	84.5	90.5	93.4
	LDA [82]	62.9	82.2	88.6	92.4	62.6	81.9	87.6	92.2	47.9	72.9	81.9	87.8	60.8	81.8	87.3	91.8
	KISSME [13]	61.4	82.4	88.5	92.6	58.7	78.7	85.3	90.1	51.2	76.3	84.3	89.5	51.2	75.7	83.6	89.7
	LFDA [10]	68.0	85.5	90.9	94.5	61.7	79.5	85.7	90.6	47.7	72.4	80.9	87.9	65.4	83.6	89.5	93.3
	XQDA [12]	61.3	81.9	87.6	92.1	55.5	76.6	84.1	89.3	51.0	77.1	84.3	89.4	44.2	71.1	81.4	88.1

TABLE 5
Evaluating the effectiveness of our HIPHOP feature. Best results (%) by single and fused features are indicated in blue and red color respectively. Our CRAFT-MFA is utilized for each type of feature in this experiment.

Dataset	VIPeR [30]				CUHK01 [31]				CUHK03 [26]				Market-1501 [33]				QMUL GRID [32]			
Rank (%)	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
ELF18 [19]	38.6	69.2	82.1	91.4	51.8	75.9	83.1	88.9	65.8	89.9	93.8	97.0	46.4	70.0	78.3	84.9	19.1	41.8	55.1	67.0
ColorLBP [54]	16.1	36.8	49.4	64.9	32.3	53.1	62.9	73.0	29.8	57.3	70.6	82.9	13.6	28.2	36.2	45.9	5.8	15.0	22.8	32.2
WHOS [46]	37.1	67.7	79.7	89.1	66.0	85.5	91.0	95.1	72.0	92.9	96.4	98.4	52.2	76.0	83.3	89.6	20.7	47.4	58.8	69.8
LOMO [12]	42.3	74.7	86.5	94.2	65.4	85.3	90.5	94.1	77.9	95.5	97.9	98.9	47.2	72.9	81.7	88.2	21.4	43.0	53.1	65.5
HIPHOP	50.3	80.0	89.6	95.5	74.5	91.2	94.8	97.1	84.3	97.1	98.3	99.1	68.7	87.1	90.8	94.0	22.4	49.9	61.8	71.7
HOP only	47.6	75.6	86.6	93.8	68.9	87.7	92.7	96.3	76.4	89.5	93.6	97.3	57.9	78.7	85.6	90.5	18.0	48.2	59.9	70.2
HIP only	47.8	76.4	86.4	94.3	70.0	88.0	92.2	95.5	79.5	92.3	96.2	97.4	67.5	85.0	89.7	92.9	21.6	46.2	59.4	71.0
fc6 only	15.0	36.6	50.9	66.0	19.6	42.1	54.0	66.3	21.2	43.3	56.1	67.2	13.5	31.3	42.3	54.1	9.6	21.1	29.3	40.2
fc7 only	10.0	27.3	39.3	55.1	11.7	27.3	37.0	48.6	12.7	27.9	36.2	48.4	9.2	24.2	34.1	45.3	5.1	14.5	20.6	32.4
HIPHOP+ELF18	52.5	81.9	91.3	96.5	75.6	91.6	95.1	97.1	86.5	97.4	98.6	99.4	69.9	86.9	90.9	94.4	23.1	50.7	60.9	72.4
HIPHOP+ColorLBP	51.0	80.3	90.2	95.7	74.3	90.8	94.4	96.9	84.0	97.0	98.6	99.5	69.3	86.9	91.4	94.1	22.9	48.4	59.2	70.0
HIPHOP+WHOS	52.5	81.0	90.5	96.2	75.9	92.1	95.2	97.3	85.0	97.4	98.7	99.5	70.3	88.1	91.7	94.6	24.5	53.3	64.4	75.0
HIPHOP+LOMO	54.2	82.4	91.5	96.9	78.8	92.6	95.3	97.8	87.5	97.4	98.7	99.5	72.3	88.2	91.9	95.0	26.0	50.6	62.5	73.3

are shown more effective than those from the $6^{\text{th}}/7^{\text{th}}$ conv layer (i.e., fc6/fc7, being abstract). This confirms early findings [24,73] that higher layers are more task-specific, thus less generic to distinct tasks.

Recall that the ordinal feature HOP (Eqn. (1)) is computed based on top-κ activations of the feature maps. We further examined the effect of κ on the feature quality on the VIPeR dataset. For obtaining a detailed analysis, we tested the HOP feature extracted from the $1^{\text{st}}/2^{\text{nd}}$ conv layer and both layers, separately. Figure 7 shows the impact of setting different κ values ranging from 5 to 50 in terms of rank-1 recognition rate. The observations suggest clearly that κ is rather insensitive with a wide satisfiable range. We set $\kappa = 20$ in all other evaluations.

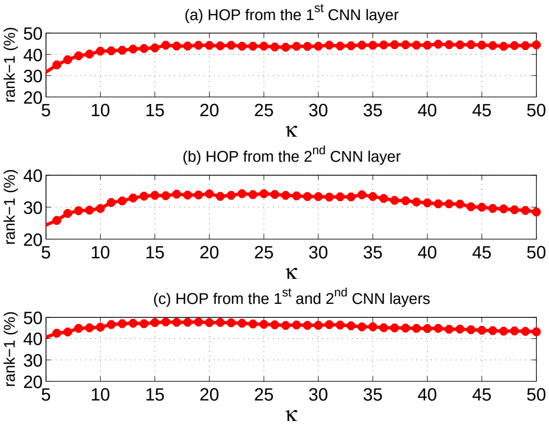


Fig. 7. Evaluating the effect of κ in the HOP descriptor.

Complementary of HIPHOP on existing re-id features. Considering the different design nature of our person representation as compared to previous re-id features, we evaluated the complementary effect of our HIPHOP with ELF18, ColorLBP, WHOS and LOMO (see the last four rows in Table 5). It is found that: (1) After combining our HIPHOP, all these existing features can produce much better re-id performance. This validates the

TABLE 6
Comparing state-of-the-art methods on *VIPeR* [30].

Rank (%)	1	5	10	20
RDC [8]	15.7	38.4	53.9	70.1
KISSME [13]	22.0	-	68.0	-
LFDA [10]	24.2	52.0	67.1	82.0
RPLM [54]	27.0	-	69.0	83.0
MtMCML [85]	28.8	59.3	75.8	88.5
LADF [9]	29.3	61.0	76.0	88.1
SalMatch [86]	30.2	52.3	65.5	79.2
MFA [11]	32.2	66.0	79.7	90.6
kLFDA [11]	32.3	65.8	79.7	90.9
Ref-reid [17]	33.3	-	78.4	88.5
SCNCD [3]	33.7	62.7	74.8	85.0
Siamese-Net [56]	34.4	62.2	75.9	87.2
CIND-Net [27]	34.8	63.6	75.6	84.5
CorStruct [87]	34.8	68.7	82.3	91.8
PolyMap [88]	36.8	70.4	83.7	91.7
KCCA [18]	37.0	-	85.0	93.0
DGD [29]	38.6	-	-	-
XQDA [12]	40.0	68.1	80.5	91.1
MLAPG [15]	40.7	69.9	82.3	92.4
RDC-Net [74]	40.5	60.8	70.4	84.4
DNS [39]	42.3	71.5	82.9	92.1
KEPLER [89]	42.4	-	82.4	90.7
LSSCDL [90]	42.7	-	84.3	91.9
CVDCA [19]	43.3	72.7	83.5	92.2
Metric Ensemble [14]	45.9	77.5	88.9	95.8
TCP [60]	47.8	74.7	84.8	89.2
CRAFT-MFA	**50.3**	**80.0**	**89.6**	**95.5**

favourable complementary role of our HIPHOP feature for existing ones. (2) Interestingly, these four fusions produce rather similar and competitive re-id performance, as opposite to the large differences in the results by each individual existing feature (see the first four rows in Table 5). This justifies the general complementary importance of the proposed feature for different existing re-id features. Next, we utilize "HIPHOP+LOMO" as the default multi-type feature fusion in the proposed CRAFT-MFA method due to its slight superiority over other combinations. This is termed as **CRAFT-MFA(+LOMO)** in the remaining evaluations.

5.3 Comparing State-of-the-Art Re-Id Methods

We compared extensively our method CRAFT-MFA with state-of-the-art person re-id approaches. In this evaluation, we considered two scenarios: (1) Person re-id between two cameras; (2) Person re-id across multiple (>2) cameras. Finally, we compared the person re-id performance by different methods using multiple feature types. We utilized the best results reported in the corresponding papers for fair comparison across all compared models.

(I) Person re-id between two cameras. This evaluation was carried out on VIPeR [30] and CUHK01 [31] datasets, each with a pair of camera views. We compared our CRAFT-MFA with both metric learning and recent deep learning based methods.

Comparisons on VIPeR - We compared with 26 state-of-the-art re-id methods on VIPeR. The performance comparison is reported in Table 6. It is evident that our CRAFT-MFA method surpassed all competitors over top ranks clearly. For instance, the rank-1 rate is improved notably from 47.8% (by the 2nd best method TCP [60]) to 50.3%. This shows the advantages and effectiveness of our approach in a broad context.

Comparisons on CUHK01 - We further conducted the comparative evaluation on the CUHK01 dataset with the results shown in Table 7. It is evident that our CRAFT-MFA method generated the highest person re-id accuracies among all the compared methods. Specifically, the best alternative (DGD) was outperformed notably by our method with a margin of $7.9\%(= 74.5\% - 66.6\%)$ rank-1 rate. Relative to VIPeR, CRAFT-MFA obtained more performance increase on CUHK01, e.g., improving rank-1 rate by $2.5\%(= 50.3\% - 47.8\%)$ on VIPeR *versus* 7.9% on CUHK01.

TABLE 7
Comparing state-of-the-art methods on *CUHK01* [31].

Rank (%)	1	5	10	20
LMNN [5]	13.4	31.3	42.3	54.1
ITML [7]	16.0	35.2	45.6	59.8
eSDC [45]	19.7	32.7	40.3	50.6
GM [31]	20.0	-	56.0	69.3
SalMatch [86]	28.5	45.9	55.7	68.0
Ref-reid [17]	31.1	-	68.6	79.2
MLF [2]	34.3	55.1	65.0	74.9
CIND-Net [27]	47.5	71.6	80.3	87.5
CVDCA [19]	47.8	74.2	83.4	89.9
Metric Ensemble [14]	53.4	76.4	84.4	90.5
TCP [60]	53.7	84.3	91.0	96.3
XQDA [12]	63.2	83.9	90.0	94.9
MLAPG [15]	64.2	85.5	90.8	94.9
DNS [39]	65.0	85.0	89.9	94.4
DGD [29]	66.6	-	-	-
CRAFT-MFA	**74.5**	**91.2**	**94.8**	**97.1**

TABLE 8
Comparing state-of-the-art methods on *CUHK03* [26].

Rank (%)	1	5	10	20	mAP (%)
SDALF [44]	4.9	21.0	31.7	-	-
ITML [7]	5.1	17.7	2.8.3	-	-
LMNN [5]	6.3	18.7	29.0	-	-
eSDC [45]	7.7	22.0	33.0	-	-
KISSME [13]	11.7	33.3	48.0	-	-
FPNN [26]	19.9	50.0	64.0	78.5	-
BoW [33]	23.0	42.4	52.4	64.2	22.7
CIND-Net [27]	45.0	76.0	83.5	93.2	-
XQDA [12]	46.3	78.9	88.6	94.3	-
LSSCDL [90]	51.2	80.8	89.6	-	-
MLAPG [15]	51.2	83.6	92.1	96.9	-
SI-CI [59]	52.2	84.9	92.4	96.7	-
DNS [39]	53.7	83.1	93.0	94.8	-
S-LSTM [62]	57.3	80.1	88.3	-	46.3
Gated-SCNN [61]	68.1	88.1	94.6	-	58.8
DGD [29]	75.3	-	-	-	-
CRAFT-MFA	**84.3**	**97.1**	**98.3**	**99.1**	**72.41**

This is as expected, because person images from VIPeR are much more challenging for re-id due to poorer imaging quality, more complex illumination patterns, and more severe background clutter (see Figure 5(a-b)). This also validates the generality and capability of the proposed method in coping with various degrees of person re-id challenges when learning view-specific and view-generic discriminative re-id information.

(II) Person re-id across more than two cameras. Real-world person re-id applications often involve a surveillance network with many cameras. It is therefore critical to evaluate the performance of associating people across a whole camera network, although the joint learning and quantification is largely under-studied in the current literature. In this multi-camera setting, we exploit the generalized CRAFT model (Eqn. (29)) to learn an adaptive sub-model for each camera view in a principled fashion. This evaluation was performed on three multi-camera re-id datasets: CUHK03 [26] (with 6 cameras in a university campus), QMUL GRID [32] (with 8 cameras in an underground station), and Market-1501 [33] (with 6 cameras near a university supermarket).

Comparisons on CUHK03 - We evaluated our approach by comparing the state-of-the-arts on CUHK03 [26]. This evaluation was conducted using detected images. It is shown in Table 8 that our method significantly outperformed all competitors, e.g., the top-2 Gated-SCNN/DGD by $16.2\%/9.0\%$ at rank-1, respectively.

Comparisons on QMUL GRID - The re-id results of different methods on QMUL GRID are presented in Table 9. It is found that the proposed CRAFT-MFA method produced the most accurate results among all competitors, similar to the observation on CUHK03 above. In particular, our CRAFT-MFA method outperformed clearly the 2nd best model LSSCDL, e.g., with similar top-1 matching rate but boosting rank-10 matching from 51.3% to 61.8%. This justifies the superiority of our CRAFT model and person appearance feature

TABLE 9
Comparing state-of-the-art methods on *QMUL GRID* [32].

Rank (%)	1	5	10	20
PRDC [8]	9.7	22.0	33.0	44.3
LCRML [91]	10.7	25.8	35.0	46.5
MRank-PRDC [92]	11.1	26.1	35.8	46.6
MRank-RSVM [92]	12.2	27.8	36.3	46.6
MtMCML [85]	14.1	34.6	45.8	59.8
PolyMap [88]	16.3	35.8	46.0	57.6
MLAPG [15]	16.6	33.1	41.2	53.0
KEPLER [89]	18.4	39.1	50.2	61.4
XQDA [12]	19.0	42.2	52.6	62.2
LSSCDL [90]	**22.4**	-	51.3	61.2
CRAFT-MFA	**22.4**	**49.9**	**61.8**	**71.7**

TABLE 10
Comparing state-of-the-art methods on *Market-1501* [33]. ($^+$): the results reported in [39] were utilized.

Query/person	Single Query		Multiple Query	
Metric	rank-1 (%)	mAP (%)	rank-1 (%)	mAP (%)
BoW [33]	34.4	14.1	42.6	19.5
KISSME$^+$ [13]	40.5	19.0	-	-
MFA$^+$ [22]	45.7	18.2	-	-
kLFDA$^+$ [66]	51.4	24.4	52.7	27.4
XQDA$^+$ [12]	43.8	22.2	54.1	28.4
DNS [39]	55.4	29.9	68.0	41.9
S-LSTM [62]	-	-	61.6	35.3
Gated-SCNN [61]	65.9	39.6	76.0	48.5
CRAFT-MFA	**68.7**	**42.3**	**77.0**	**50.3**

in a more challenging realistic scenario.

Comparisons on Market-1501 - We compared the performance on Market-1501 with these methods: Bag-of-Words (BoW) based baselines [33], a Discriminative Null Space (DNS) learning based model [39], and four metric learning methods KISSME [13], MFA [22], kLFDA [66], XQDA [12], S-LSTM [62], Gated-SCNN [61]. We evaluated both the single-query and multi-query (using multiple probe/query images per person during the deployment stage) settings. It is evident from Table 10 that our CRAFT-MFA method outperformed all competitors under both single-query and multi-query settings. By addressing the small sample size problem, the DNS model achieves more discriminative models than other metric learning algorithms. However, all these methods focus on learning view-generic discriminative information whilst overlooking largely useful view-specific knowledge. Our CRAFT-MFA method effectively overcome this limitation by encoding camera correlation into an extended feature space for jointly learning both view-generic and view-specific discriminative information. Additionally, our method benefits from more view change tolerant appearance patterns deeply learned from general auxiliary data source for obtaining more effective person description, and surpassed recent deep methods Gated-SCNN and S-LSTM. All these evidences validate consistently the effectiveness and capability of the proposed person visual features and cross-view re-id model learning approach in multiple application scenarios.

(III) Person re-id with multiple feature representations. Typically, person re-id can benefit from using multiple different types of appearance features owing to their complementary effect (see Table 5). Here, we compared our CRAFT-MFA(+LOMO) method with competitive re-id models using multiple features. This comparison is given in Table 11. It is found that our CRAFT-MFA(+LOMO) method notably outperformed all compared methods utilizing two or more types of appearance features, particularly on CUHK01, CUHK03, and Market-1501. Along with the extensive comparison with single feature based methods above, these observations further validate the superiority and effectiveness of our proposed method under varying feature representation cases.

TABLE 11
Comparing state-of-the-art methods using multiple types of appearance feature representations.

Dataset	VIPeR [30]			
Rank (%)	1	5	10	20
Late Fusion [93]	30.2	51.6	62.4	73.8
MLF+LADF [2]	43.4	73.0	84.9	93.7
Metric Ensemble [14]	45.9	77.5	88.9	95.8
CVDCA (fusion) [19]	47.8	76.3	86.3	94.0
FFN-Net (fusion) [57]	51.1	81.0	91.4	**96.9**
DNS (fusion) [39]	51.2	82.1	90.5	95.9
SCSP [40]	53.5	**82.6**	**91.5**	96.7
GOG (fusion) [55]	49.7	-	88.7	94.5
CRAFT-MFA	50.3	80.0	89.6	95.5
CRAFT-MFA(+LOMO)	**54.2**	82.4	**91.5**	**96.9**
Dataset	CUHK01 [31]			
Rank (%)	1	5	10	20
Metric Ensemble [14]	53.4	76.4	84.4	90.5
FFN-Net (fusion) [57]	55.5	78.4	83.7	92.6
GOG (fusion) [55]	67.3	86.9	91.8	95.9
DNS (fusion) [39]	69.1	86.9	91.8	95.4
CRAFT-MFA	74.5	91.2	94.8	97.1
CRAFT-MFA(+LOMO)	**78.8**	**92.6**	**95.3**	**97.8**
Dataset	CUHK03 [26]			
Rank (%)	1	5	10	20
DNS (fusion) [39]	54.7	84.8	94.8	95.2
GOG (fusion) [55]	65.5	88.4	93.7	-
CRAFT-MFA	84.3	97.1	98.3	99.1
CRAFT-MFA(+LOMO)	**87.5**	**97.4**	**98.7**	**99.5**
Dataset	QMUL GRID [32]			
Rank (%)	1	5	10	20
SCSP [40]	24.2	44.6	54.1	65.2
GOG (fusion) [55]	24.7	47.0	58.4	69.0
CRAFT-MFA	22.4	49.9	61.8	71.7
CRAFT-MFA(+LOMO)	**26.0**	**50.6**	**62.5**	**73.3**
Dataset	Market-1501 [33]			
Query/person	Single Query		Multiple Query	
Metric	rank-1 (%)	mAP (%)	rank-1 (%)	mAP (%)
BoW(+HS) [33]	-	-	47.3	21.9
DNS (fusion) [39]	61.0	35.7	71.6	46.0
SCSP [40]	51.9	26.4	-	-
CRAFT-MFA	68.7	42.3	77.0	50.3
CRAFT-MFA(+LOMO)	**71.8**	**45.5**	**79.7**	**54.3**

5.4 Discussion

Here, we discuss the performance of HIPHOP in case that deep model fine-tuning on the available labelled target data is performed in prior to feature extraction. Our experimental results suggest that this network adaptation can only bring marginal re-id accuracy gain on our HIPHOP feature, e.g., $< 1\%$ rank-1 increase for all datasets except Market-1501 (1.4%), although the improvement on using fc6 and fc7 feature maps (which are much inferior to HIPHOP either fine-tuned or not, see the supplementary file for more details) is clearer. This validates empirically our argument that lower layer conv filters can be largely task/domain generic and expressive, thus confirming the similar earlier finding [25] particularly in person re-id context. This outcome is reasonable considering that the amount of training person images could be still insufficient (e.g., 632 on VIPeR, 1940 on CUHK01, 12197 on CUHK03, 250 on QMUL GRID, 12936 on Market-1501) for producing clear benefit, especially for lower conv layers which tend to be less target task specific [25]. Critically, model fine-tuning not only introduces the cost of extra network building complexity but also *unfavourably* renders our feature extraction method domain-specific – relying on a sufficiently large set of labelled training data in the target domain. Consequently, we remain our *domain-generic* (i.e., independent of target labelled training data and deployable universally) re-id feature extraction method as our main person image representation generation way.

6 Conclusion

We have presented a new framework called CRAFT for person re-identification. CRAFT is formed based on the camera correlation aware feature augmentation. It is capable of jointly learning both view-generic and view-specific discriminative information for person re-id in a principled manner. Specifically, by creating automatically a camera correlation aware feature space, view-generic learning algorithms are allowed to induce view-specific sub-models which simultaneously take into account the shared view-generic discriminative information so that more reliable re-id models can be produced. The correlation between per-camera sub-models can be further constrained by our camera view discrepancy regularization. Beyond the common person re-id between two cameras, we further extend our CRAFT framework to cope with re-id jointly across a whole network of more than two cameras. In addition, we develop a general feature extraction method allowing to construct person appearance representations with desired view-invariance property by uniquely exploiting less relevant auxiliary object images other than the target re-id training data. That is, our feature extraction method is universally scalable and deployable regardless of the accessibility of labelled target training data. Extensive comparisons against a wide range of state-of-the-art methods have validated the superiority and advantages of our proposed approach under both camera pair and camera network based re-id scenarios on five challenging person re-id benchmarks.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002),NSFC (61522115, 61573387, 61661130157), and the RS-Newton Advanced Fellowship (NA150459), and Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265. This work was finished when Mr. Chen was a master student at Sun Yat-sen University. The corresponding author and principal investigator for this paper is Wei-Shi Zheng.

References

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, 2014, vol. 1.

[2] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identfiation," in *CVPR*, 2014.

[3] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014.

[4] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE TPAMI*, vol. 35, no. 7, pp. 1622–1634, 2013.

[5] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *NIPS*, 2006.

[6] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007.

[8] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE TPAMI*, vol. 35, no. 3, pp. 653–668, 2013.

[9] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013.

[10] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.

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

[12] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[13] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[14] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015.

[15] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," *ICCV*, 2015.

[16] L. An, S. Yang, and B. Bhanu, "Person re-identification by robust canonical correlation analysis," *IEEE SPL*, vol. 22, no. 8, pp. 1103–1107, 2015.

[17] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person re-identification with reference descriptor," *IEEE TCSVT*, vol. 26, no. 4, pp. 776–787, 2016.

[18] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *ICDSC*, 2014.

[19] Y.-C. Chen, W.-S. Zheng, P. C. Yuen, and J. Lai, "An asymmetric distance model for cross-view feature mapping in person re-identification," in *IEEE TCSVT*, 2015.

[20] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE TPAMI*, vol. 38, pp. 591–606, 2016.

[21] X. Wang, W.-S. Zheng, X. Li, and J. Zhang, "Cross-scenario transfer person re-identification," *IEEE TCSVT*, 2015.

[22] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE TPAMI*, 2007.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[24] 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," in *ICML*, 2014.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

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

[27] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," *CVPR*, 2015.

[28] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *CVPR*, 2015.

[29] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016.

[30] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *VS-PETS Workshop*, 2007.

[31] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.

[32] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *CVPR*, 2009.

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

[34] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking." in *BMVC*, 2010.

[35] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.

[36] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *ECCV*. Springer, 2014, pp. 688–703.

[37] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *ICCV*, 2015, pp. 4678–4686.

[38] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *ICCV*, 2015, pp. 3765–3773.

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

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

[41] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *CVPR*, 2016.

[42] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *IEEE TPAMI*, vol. 38, no. 12, pp. 2501–2514, 2016.

[43] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *ECCV*. Springer, 2016, pp. 405–422.

[44] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.

[45] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[46] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE TPAMI*, 2014.

[47] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised dictionary learning," in *BMVC*, 2015.

[48] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *ICIP*, 2016, pp. 769–773.

[49] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong, "Person re-identification by unsupervised video matching," *Pattern Recognit.*, vol. 65, pp. 197–210, 2017.

[50] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *ICCV*, 2007.

[51] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.

[52] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *BMVC*, 2011.

[53] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *AVSS*, 2010.

[54] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012.

[55] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016.

[56] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for practical person re-identification," *arXiv e-prints*, 2014.

[57] 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 *WACV*, 2016.

[58] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE TIP*, vol. 25, no. 5, pp. 2353–2367, 2016.

[59] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *CVPR*, 2016.

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

[61] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016.

[62] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016.

[63] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE TNN*, 2011.

[64] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE TKDE*, 2010.

[65] B. Geng, D. Tao, and C. Xu, "Daml: Domain adaptation metric learning," *IEEE TIP*, 2011.

[66] M. Xiao and Y. Guo, "Feature space independent semi-supervised domain adaptation via kernel matching," *IEEE TPAMI*, 2014.

[67] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE TPAMI*, 2014.

[68] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE TPAMI*, vol. 36, pp. 2288–2302, 2014.

[69] H. Daumé III, "Frustratingly easy domain adaptation," in *Annual Meeting of the Association for Computational Linguistics*, June 2007.

[70] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016.

[71] B. Muquet, Z. Wang, G. B. Giannakis, M. De Courville, and P. Duhamel, "Cyclic prefixing or zero padding for wireless multicarrier transmissions?" *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 2136–2148, 2002.

[72] D. Wang, N. Canagarajah, D. Redmill, and D. Bull, "Multiple description video coding based on zero padding," in *International Symposium on Circuits and Systems*, vol. 2, 2004, pp. II–205.

[73] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[74] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv e-prints*, 2014.

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

[77] Z. Sun, T. Tan, Y. Wang, and S. Z. Li, "Ordinal palmprint represention for personal identification," in *CVPR*, 2005.

[78] P. Sinha, "Perceiving and recognizing three-dimensional forms," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[79] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[80] Y.-C. Wong, "Differential geometry of grassmann manifolds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 57, no. 3, p. 589, 1967.

[81] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *ICML*, 2008.

[82] A. M. Martínez and A. C. Kak, "Pca versus lda," *IEEE TPAMI*, vol. 23, no. 2, pp. 228–233, 2001.

[83] E. Choi and C. Lee, "Feature extraction based on the bhattacharyya distance," *Pattern Recognit.*, vol. 36, no. 8, 2003.

[84] A. Chakraborty, A. Das, and A. Roy-Chowdhury, "Network consistent data association," *IEEE TPAMI*, vol. PP, no. 99, pp. 1–1, October 2015.

[85] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE TIP*, vol. 23, no. 8, pp. 3656–3670, 2014.

[86] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *ICCV*, 2013.

[87] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *ICCV*, 2015.

[88] 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 *CVPR*, 2015.

[89] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE TIP*, vol. 24, no. 12, pp. 5645–5658, 2015.

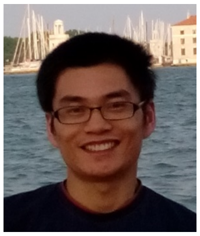
[90] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *CVPR*, 2016.

[91] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting global similarities," in *ICPR*, 2014.

[92] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *ICIP*, 2013.

[93] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *CVPR*, 2015.

Ying-Cong Chen received his BEng and Master degree from Sun Yat-sen University in 2013 and 2016 respectively. Now he is a PhD student in the Chinese University of Hong Kong. His research interest includes computer vision and machine learning.



Xiatian Zhu received his B.Eng. and M.Eng. from University of Electronic Science and Technology of China, and his Ph.D. (2015) from Queen Mary University of London. He won The Sullivan Doctoral Thesis Prize (2016), an annual award representing the best doctoral thesis submitted to a UK University in the field of computer or natural vision. His research interests include computer vision, pattern recognition and machine learning.

Wei-Shi Zheng is now a Professor at Sun Yat-sen University. He has now published more than 90 papers, including more than 60 publications in main journals (TPAMI,TIP,PR) and top conferences (ICCV, CVPR,IJCAI). His research interests include person/object association and activity understanding in visual surveillance. He has joined Microsoft Research Asia Young Faculty Visiting Programme. He is a recipient of Excellent Young Scientists Fund of the NSFC, and a recipient of Royal Society-Newton Advanced Fellowship. Homepage: http://isee.sysu.edu.cn/%7ezhwshi/

Jian-Huang Lai is Professor of School of Data and Computer Science in Sun Yat-sen university. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications. He has published over 100 scientific papers in the international journals and conferences on image processing and pattern recognition e.g., IEEE TPAMI, IEEE TNN, IEEE TIP, IEEE TSMC (Part B), Pattern Recognition, ICCV, CVPR and ICDM.