Learning View-Specific Deep Networks for Person Re-Identification

Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie

***Abstract*—In recent years, a growing body of research has focused on the problem of person re-identification (re-id). The re-id techniques attempt to match the images of pedestrians from disjoint non-overlapping camera views. A major challenge of re-id is the serious intra-class variations caused by changing viewpoints. To overcome this challenge, we propose a deep neural network-based framework which utilizes the view information in the feature extraction stage. The proposed framework learns a view-specific network for each camera view with a cross-view Euclidean constraint (CV-EC) and a cross-view center loss (CV-CL). We utilize CV-EC to decrease the margin of the features between diverse views and extend the center loss metric to a view-specific version to better adapt the re-id problem. Moreover, we propose an iterative algorithm to optimize the parameters of the view-specific networks from coarse to fine. The experiments demonstrate that our approach significantly improves the performance of the existing deep networks and outperforms the state-of-the-art methods on the VIPeR, CUHK01, CUHK03, SYSU-mReId, and Market-1501 benchmarks.**

***Index Terms*—Person re-identification, view-specific deep networks, cross-view Euclidean constraint, cross-view center loss.**

I. Introduction

WITH the increasing ubiquity of closed-circuit television cameras, the problem of person re-identification (re-id) has attracted considerable research attention. The purpose of re-id techniques is to match the images of pedestrians from disjoint camera views. The application of re-id techniques in intelligent video surveillance systems will be beneficial to enhance security in public areas. Conducting re-id over disjoint non-overlapping camera views is challenging because of the dramatic visual changes caused by variations in illumination, image quality, and especially viewpoint.

In the existing literature, most re-id methods consist of a feature extraction process and a view-invariant discriminative transformation or metric [1–7]. Figure 1 displays a traditional re-id framework for addressing the cross-view problem. Conventional re-id approaches first extract view-generic features for different views. Then, a view-invariant model is learned to narrow the gaps between the images of the intra-class pedestrians and increase the distances between the inter-class samples. Three issues prevent traditional methods from solving the re-id problem. First, traditional features characterize pedestrians without any view-specific knowledge. View-generic features are inadequate for solving a re-id problem when dramatic changes in appearance occur between disjoint camera views. Second, feature extraction and view-invariant model learning are independent. During training, the view-invariant model propagates no information back to the view-generic features. View information, which is commonly exploited by discriminative model learning algorithms, is rarely utilized to improve the feature extraction process. Learning a model with excellent generalizability simply by metric learning is difficult because the view-generic features of the same person vary across different views. Third, traditional methods learn a view-invariant model shared by all views. Utilizing a shared model for all views ignores the discrepancies among different views. Compared with the shared model, view-specific models can cover stronger view-related information. With such information, view-specific models can achieve better performance than view-generic ones. However, the amount of computational resources grows dramatically with the increasing number of camera views. Currently, few studies have focused on the learning of view-specific models.

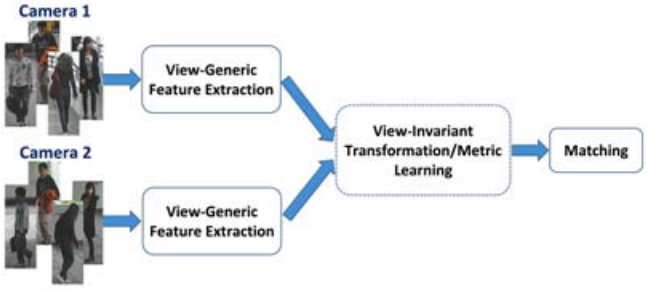

Fig. 1. Traditional re-id framework for addressing the cross-view problem. Traditional re-id framework conducts the same implementation to extract view-generic features for different camera views. The feature extraction process is usually followed by a view-invariant discriminative transformation or metric, which is shared across the changing viewpoints in most cases. The dashed window indicates that the step can be skipped.

Although view information may be useful for improving the performance of the existing methods, learning view-specific features or transformations for re-id is still under study. The current literature focuses on learning view-invariant but view-generic discriminative information. The existing approaches mainly integrate view information into metric learning methods by projecting the view-generic features onto a view-invariant common space [8–11]. To further exploit view-

Corresponding author: Jianhuang Lai (e-mail: stsljh@mail.sysu.edu.cn).

Zhanxiang Feng is with the School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou 510006, China, and with the Xin Hua College of Sun Yat-Sen University, Guangzhou 510006, China (e-mail: fengzhx@mail2.sysu.edu.cn)

Jianhuang Lai and Xiaohua Xie are with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, China, and with the Guangdong Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006, China (e-mail: stsljh@mail.sysu.edu.cn and xiexiaoh6@mail.sysu.edu.cn).

related information, some recent works [12–16] attempt to learn a view-specific dictionary or transformation for each view and project the view-generic features onto the corresponding view-specific spaces. Although view-specific models can minimize the discrepancies among changing viewpoints, the process of view-generic feature extraction ignores the view information hidden in the low-level features. The using of view-specific features may improve the performance of the existing frameworks. Nevertheless, few studies have explored the topic of view-specific feature extraction for re-id issue.

Data-driven features are more effective than empirically designed features in exploiting view-specific information. Deep learning is the most effective technique to learn discriminative features from the training data. Deep networks have been successfully applied in various computer vision tasks, including image classification [17, 18], face recognition [19, 20], and action recognition [21, 22]. In the last few years, deep learning-based approaches have been proved to be effective for re-id [23–34]. Deep models may exploit the view information and extract discriminative view-related features by using a back propagation algorithm. However, the existing approaches share the parameters of deep networks across all views, particularly for the layers that extract low-level features. Therefore, view-specific information may be ignored during feature extraction stage.

According to the above discussion, we propose to extract view-specific features by using deep networks. To the best of our knowledge, our approach is the first to extract view-specific features by using deep networks for re-id. To narrow the gaps between features from changing viewpoints, we integrate a cross-view Euclidean constraint (CV-EC) into the proposed framework. The goal of CV-EC is to decrease the distances between the deep features of the same person from disjoint camera views. Together with CV-EC, we integrate another loss called the cross-view center loss (CV-CL) to improve the discriminative ability of the view-specific deep networks. Center loss has been proved to be effective for face recognition [35] by narrowing the margin between the samples and their corresponding class centers. In this paper, we extend the center loss to a view-specific version that better fits the re-id problem. Moreover, we propose an iterative optimization algorithm (ICV-ECCL) to learn CV-EC and CV-CL alternatively and optimize the parameters of the view-specific networks from coarse to fine. Finally, we extend CV-EC and CV-CL to a multi-view version to deal with the application using more than two cameras.

The study makes the following contributions.

- We propose a novel framework to learn view-specific networks for feature extraction in re-id and demonstrate that the utilization of view-specific information is crucial for extracting robust re-id representations.
- We propose the CV-EC and CV-CL constraints to integrate view information into view-specific deep networks and minimize the distances between the features across varying views. The proposed framework is adaptive to any baseline network and benchmark.
- We propose an iterative algorithm (ICV-ECCL) to optimize the parameters of view-specific networks from coarse to fine.
- We evaluate the proposed framework by extensive comparisons between ICV-ECCL and a variety of methods on the VIPeR, CUHK01, CUHK03, SYSU-mREID, and Market-1501 benchmarks. The experimental results validate that the proposed framework significantly improves the performance of the existing deep networks and outperforms the state-of-the-art methods.

The rest of this paper is organized as follows: In Section II, we review the related works. We then introduce the overall framework and the optimization algorithm in Section III. Section IV presents the experimental results of the comparisons and the self-evaluation. Section V concludes the paper.

II. Related Work

A. View-Generic Features for Re-Id

View-generic features are widely used for re-id. The existing works extract view-generic features to describe the visual characteristics of the objective pedestrians captured in disjoint cameras. The color [36–41], shape [42–44], texture [45–47], and spatio-temporal features [48, 49] are the most commonly used characteristics to describe the target pedestrian regions. Descriptors such as LBP [46], HOG [43, 50], Gabor [9, 51], ELF [52], and SIFT [36, 53] have been successfully applied to match persons from disjoint views. Researchers have also extracted global features [37, 39], local features [54, 55], and patch-based features [56, 57] to integrate structural constraints into view-generic features. Moreover, some studies combine multiple features to form better descriptors [58, 59]. In addition to hand-crafted features, some unsupervised learning-based view-generic representations have been used for re-id, including sparse coding [11, 12], bag-of-words [41, 60], and fisher vector [61]. Moreover, saliency [36, 40] and pedestrian attributes [62–65] have been used to match people across different cameras.

B. View-Invariant Transformations and Metrics for Re-Id

The learning of view-invariant transformations and metrics for re-id is of great interest to researchers. View-invariant transformation and metric learning algorithms can exploit label information and narrow the gaps between different views. In the existing literature, supervised learning-based approaches demonstrate superior performance to that of unsupervised learning-based approaches. KISSME [1], LMNN [44], SCSP [2], RankSVM [54], PCCA [11], LFDA [4], and ITML [66] have been proposed to learn discriminative view-invariant representations.

In addition to label information, view information has been used to learn robust re-id representations. Traditionally, researchers project the features from both camera views onto a common space by a shared view-invariant model [8–11]. In recent years, some works have attempted to learn view-specific transformations or models [12–16]. Further, view-specific models have been proved to be superior to the shared models regarding the generalizability across the appearance variations caused by viewpoint changes. Chen et al. [13]

proposed a view-specific re-id framework by the feature augmentation of different views. The framework executed a view-specific learning algorithm to measure the camera correlations and transform the features from disjoint views to a new adaptive space by adaptive augmentation. Chen et al. [14] also introduced an asymmetric distance model for cross-view feature mapping to extract discriminative features against the changing viewpoints. The asymmetric distance model learns camera-specific projections to transform low-level features from both views to a common space. Li et al. [15] proposed a cross-view projective dictionary learning-based method, which learned a specific dictionary for each view, to cope with the re-id task.

C. Deep Learning Models for Re-Id

Deep learning is playing an increasingly significant role for the re-id task. The existing deep network-based methods focus on designing Siamese networks that combine the input pairs of pedestrians from multiple camera views [23–28]. For example, Yi et al. [23] and Li et al. [24] used a Siamese neural network to optimize the parameters of deep networks with pairs of inputs. Ahmed et al. [25] further extended the Siamese neural network by embedding a layer to compute the cross-input neighborhood differences. Varior et al. [27] jointly combined a Siamese neural network with LSTM to recurrently extract spatial-structured features. Varior et al. also proposed a Gated S-CNN [28] to match pedestrians from disjoint views horizontally.

Another common strategy for learning view-invariant deep features is to learn a discriminative cross-view metric through the ensemble layers [29–33]. Chen et al. [29] proposed a deep ranking model to improve the ranking of the correct match of probe images with a learning-to-rank algorithm. Su et al. [30] proposed a semi-supervised deep attribute learning-based method for re-id. They first trained a deep network with labeled attributes in an extra dataset and then fine-tuned the network with the re-id dataset by using the attribute triplet loss. Liu et al. [32] proposed a novel soft attention-based model called the end-to-end comparable attention network (CAN). The CAN model stimulates human perception in judging whether the parts are from the same person or not.

D. Conclusion on Current Approaches

View information needs to be further exploited by the existing re-id approaches. On one hand, the existing traditional re-id frameworks utilize view information simply by using view-invariant transformations or metrics. The main drawback of the view-specific model-based methods lies in the process of feature extraction. For most cases, the existing methods conduct the same implementation to extract view-generic features from different views, which weakens the strength of the view-related information during feature extraction and deteriorates the performance of the overall framework. To overcome the drawback of view-specific models and improve their generalizability, we need to extract view-specific discriminative features.

On the other hand, few works pay attention to learn view-specific deep features for re-id in an end-to-end manner. Most deep learning-based methods share the parameters of a single network for disjoint views. Such approaches learn the view-generic information while neglecting the view-specific information. The learning of an individual deep network for each camera view can cover more view-specific information and improve the generalizability towards viewpoint changes. On the basis of the above observations, we focus on training view-specific deep networks to extract view-related features, which is crucial for improving the generalizability of the re-id models.

III. Proposed Framework

A. Overview

By sharing the parameters of the lower layers, the existing works extract the view-generic low-level features for changing viewpoints. View information is thus ignored during low-level feature extraction. To solve this problem, we propose to learn view-specific deep networks to extract discriminative view-related features for re-id. Figure 2 illustrates the proposed framework. Our framework is different from conventional deep models in that it learns different deep networks for different views. Further, we integrate CV-EC into the framework to align the deep features from disjoint views. We also introduce CV-CL to narrow the gaps between the features across different views. Furthermore, CV-EC and CV-CL are utilized iteratively (ICV-ECCL) to update the parameters of view-specific networks. Finally, we extend CV-EC and CV-CL to a multi-view version.

B. Cross-View Euclidean Constraint

As shown in Figure 2, a view-specific network is learned for each view in the proposed framework. We seek to learn the view-related features during training. For view-specific deep networks, the cross-view intra-class distance may be very large without the consideration of any cross-view constraint. We need to minimize the cross-view intra-class distances between pairs of view-specific features. To address this problem, we integrate a constraint into the view-specific networks to guarantee that the features of the same people under disparate views are as close as possible.

In this section, we introduce the main principle of CV-EC. To simplify the discussion, we only consider the situation of two cameras. In fact, the proposed method can be extended to multiple cameras. Given the deep features from disjoint views as $\{\boldsymbol{x}_{ip}^1, \boldsymbol{x}_{iq}^2\}, 1 \leq i \leq M, 1 \leq p \leq K_i^1, 1 \leq q \leq K_i^2$, where i denotes the identity of the pedestrians, M represents the number of identities, (ip, iq) refer to the pth and qth feature of the ith identity from view 1 and view 2, respectively, and K_i^1 and K_i^2 denote the numbers of the ith identity from different views, CV-EC can be formulated as follows:

$$\mathcal{L}_{cv-ec} = \frac{1}{2M} \sum_{i=1}^{M} \frac{1}{K_i^1 K_i^2} \sum_{p=1}^{K_i^1} \sum_{q=1}^{K_i^2} \parallel \boldsymbol{x}_{ip}^1 - \boldsymbol{x}_{iq}^2 \parallel_2^2 . \quad (1)$$

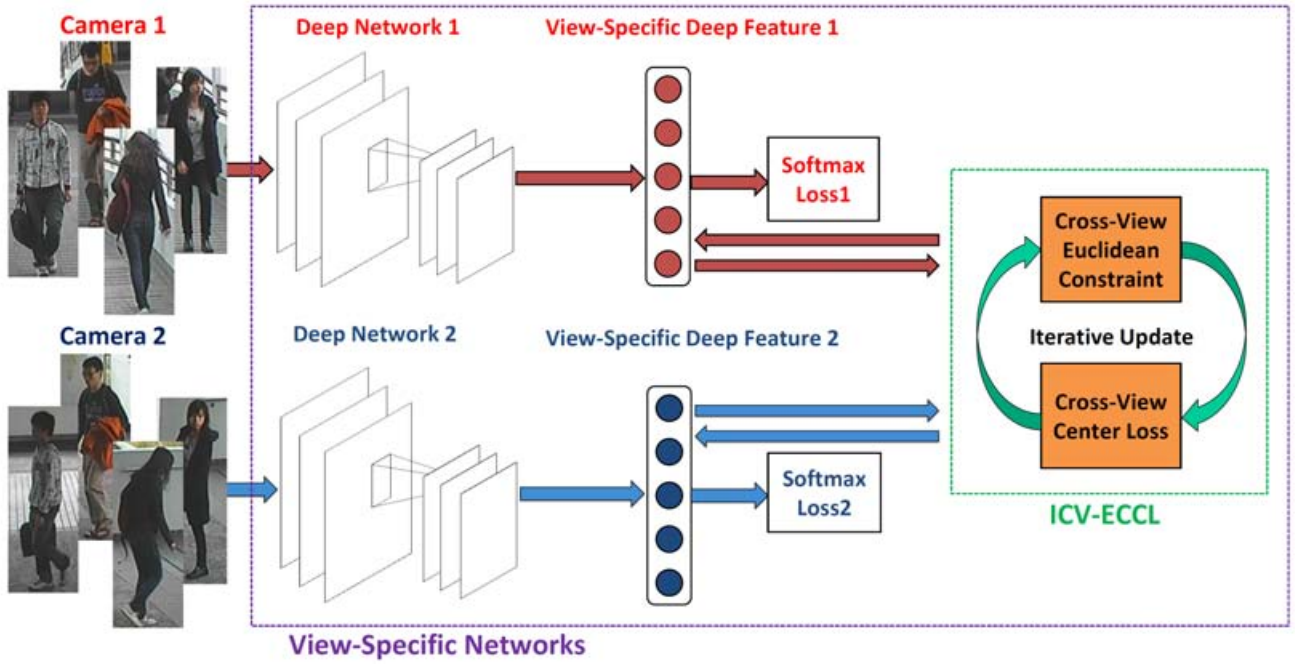

Fig. 2. Proposed re-id framework. View-specific deep networks are trained to extract view-related features. The cross-view Euclidean constraint and the cross-view center loss are integrated into the view-specific networks to narrow the gaps between the features from changing viewpoints, and they are employed in an iterative manner.

CV-EC aims to minimize the cross-view intra-class distances between the embedding feature pairs from different view-specific networks. We choose to implement the CV-EC metric between the last fully-connected layers. Thus, we can obtain view-specific information when extracting low-level features. The proposed framework jointly minimizes CV-EC and the view-specific softmax losses to extract discriminative features. The formula can be written as follows:

$$\mathcal{L}_1 = \sum_{v=1}^{2} \mathcal{L}_s^v + \lambda_1 \mathcal{L}_{cv-ec}, \tag{2}$$

where $\mathcal{L}_s^v$ denotes the vth view-specific softmax loss and λ_1 represents the regularization factor between the softmax loss and CV-EC. The softmax loss [35] for the vth view can be formulated as follows:

$$\mathcal{L}_s^v = -\sum_{n=1}^{N_v} log \frac{\mathrm{e}^{(\boldsymbol{W}_{y_n^v}^v)^T \boldsymbol{x}_n^v + b_{y_n^v}^v}}{\sum_{m=1}^{M} \mathrm{e}^{(\boldsymbol{W}_m^v)^T \boldsymbol{x}_n^v + b_m^v}}, \tag{3}$$

where $\boldsymbol{x}_n^v$ denotes the nth deep feature from view v, belonging to the y_n^vth class; $\boldsymbol{W}_m^v$ represents the mth column of the weights $\boldsymbol{W^v}$ in the last fully-connected layer from view v; b^v refers to the view-specific bias term; and N_v indicates the size of the training samples in a mini-batch.

C. Cross-View Center Loss

Center loss intends to learn discriminative features by penalizing the distances between the deep features and their corresponding centers. We can apply center loss to the re-id problem because each sequence captures multiple samples of the objective pedestrian. The original center loss [35] can be expressed as follows:

$$\mathcal{L}_C = \frac{1}{2M} \sum_{i=1}^{M} \frac{1}{K_i} \sum_{j=1}^{K_i} \parallel \boldsymbol{x}_{ij} - \boldsymbol{C}_i \parallel_2^2, \tag{4}$$

where $\boldsymbol{x}_{ij}$ denotes the jth deep feature of the ith person, K_i refers to the number of the ith identity, and $\boldsymbol{C}_i$ indicates the ith class center.

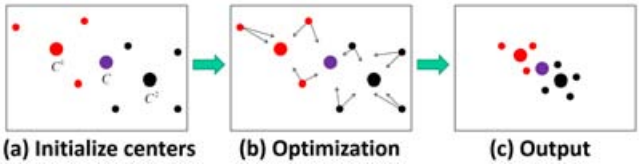

Fig. 3. Cross-view center loss: (a) Initialize the centers for all the views and all the samples; (b) Optimization process that narrows the gaps between the samples and their corresponding centers; (c) Output result.

View information is an intrinsic characteristic of re-id. When applying center loss to re-id, view information is useful for improving the performance. In this study, we extend center loss to a cross-view version (CV-CL). CV-CL is proposed to penalize the distances between the deep features and their corresponding view-specific centers. With CV-CL, we can narrow the gaps between different camera views. Figure 3 illustrates the idea of CV-CL.

Given training features $\{\boldsymbol{x}_{ip}^1, \boldsymbol{x}_{iq}^2\}$, we can easily compute the centers and obtain $\{\boldsymbol{C}_i^1, \boldsymbol{C}_i^2, \boldsymbol{C}_i\}$, where $\boldsymbol{C}_i$ denotes the center of the ith class of all samples, and $\boldsymbol{C}_i^1$ and $\boldsymbol{C}_i^2$ represent the centers of the ith class of samples from the corresponding

view. The formula for CV-CL is as follows:

$$\mathcal{L}_{cv-cl} = \frac{1}{2M}\sum_{i=1}^{M}(\frac{1}{K_i^1}\sum_{p=1}^{K_i^1}(\parallel \boldsymbol{x}_{ip}^1 - \boldsymbol{C}_i^1 \parallel_2^2 + \parallel \boldsymbol{x}_{ip}^1 - \boldsymbol{C}_i \parallel_2^2) + \frac{1}{K_i^2}\sum_{q=1}^{K_i^2}(\parallel \boldsymbol{x}_{iq}^2 - \boldsymbol{C}_i^2 \parallel_2^2 + \parallel \boldsymbol{x}_{iq}^2 - \boldsymbol{C}_i \parallel_2^2)). \tag{5}$$

We combine CV-CL with the softmax loss. Therefore, the objective function can be written as follows:

$$\mathcal{L}_2 = \sum_{v=1}^{2}\mathcal{L}_s^v + \lambda_2\mathcal{L}_{cv-cl}. \tag{6}$$

D. Optimization of CV-EC and CV-CL

We adopt the standard stochastic gradient-based optimization algorithm to optimize the parameters of the view-specific networks by using CV-EC and CV-CL. For CV-EC, the gradients of $\mathcal{L}_{cv-ec}$ with respect to $\boldsymbol{x}_{ip}^1$ and $\boldsymbol{x}_{iq}^2$ are computed as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_{cv-ec}}{\partial\boldsymbol{x}_{ip}^1} &= \boldsymbol{x}_{ip}^1 - \boldsymbol{x}_{iq}^2,\\ \frac{\partial\mathcal{L}_{cv-ec}}{\partial\boldsymbol{x}_{iq}^2} &= \boldsymbol{x}_{iq}^2 - \boldsymbol{x}_{ip}^1.\end{aligned} \tag{7}$$

By denoting the parameters of view-specific networks as $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$, we obtain the following:

$$\begin{aligned}\frac{\partial\mathcal{L}_1}{\partial\boldsymbol{\theta}^1} &= \frac{\partial\mathcal{L}_s^1}{\partial\boldsymbol{\theta}^1} + \frac{\partial\mathcal{L}_{cv-ec}}{\partial\boldsymbol{x}_{ip}^1}\cdot\frac{\partial\boldsymbol{x}_{ip}^1}{\partial\boldsymbol{\theta}^1},\\ \frac{\partial\mathcal{L}_1}{\partial\boldsymbol{\theta}^2} &= \frac{\partial\mathcal{L}_s^2}{\partial\boldsymbol{\theta}^2} + \frac{\partial\mathcal{L}_{cv-ec}}{\partial\boldsymbol{x}_{iq}^2}\cdot\frac{\partial\boldsymbol{x}_{iq}^2}{\partial\boldsymbol{\theta}^2}.\end{aligned} \tag{8}$$

Finally, we update $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$ with the learning rate μ as follows:

$$\begin{aligned}\boldsymbol{\theta}^1 &:= \boldsymbol{\theta}^1 - \mu\cdot\frac{\partial\mathcal{L}_1}{\partial\boldsymbol{\theta}^1},\\ \boldsymbol{\theta}^2 &:= \boldsymbol{\theta}^2 - \mu\cdot\frac{\partial\mathcal{L}_1}{\partial\boldsymbol{\theta}^2}.\end{aligned} \tag{9}$$

For CV-CL, the gradients of $\mathcal{L}_{cv-cl}$ with respect to $\boldsymbol{x}_{ip}^1$ and $\boldsymbol{x}_{iq}^2$ can be obtained as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{x}_{ip}^1} &= (\boldsymbol{x}_{ip}^1 - \boldsymbol{C}_i^1) + (\boldsymbol{x}_{ip}^1 - \boldsymbol{C}_i),\\ \frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{x}_{iq}^2} &= (\boldsymbol{x}_{iq}^2 - \boldsymbol{C}_i^2) + (\boldsymbol{x}_{iq}^2 - \boldsymbol{C}_i),\end{aligned} \tag{10}$$

The gradients of $\boldsymbol{C}_i^1$, $\boldsymbol{C}_i^2$, and $\boldsymbol{C}_i$ can be obtained as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{C}_i^1} &= \boldsymbol{C}_i^1 - \boldsymbol{x}_{ip}^1,\\ \frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{C}_i^2} &= \boldsymbol{C}_i^2 - \boldsymbol{x}_{iq}^2,\\ \frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{C}_i} &= (\boldsymbol{C}_i - \boldsymbol{x}_{ip}^1) + (\boldsymbol{C}_i - \boldsymbol{x}_{iq}^2).\end{aligned} \tag{11}$$

The parameters $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$ are updated as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_2}{\partial\boldsymbol{\theta}^1} &= \frac{\partial\mathcal{L}_s^1}{\partial\boldsymbol{\theta}^1} + \frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{x}_{ip}^1}\cdot\frac{\partial\boldsymbol{x}_{ip}^1}{\partial\boldsymbol{\theta}^1},\\ \frac{\partial\mathcal{L}_2}{\partial\boldsymbol{\theta}^2} &= \frac{\partial\mathcal{L}_s^2}{\partial\boldsymbol{\theta}^2} + \frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{x}_{iq}^2}\cdot\frac{\partial\boldsymbol{x}_{iq}^2}{\partial\boldsymbol{\theta}^2},\\ \boldsymbol{\theta}^1 &:= \boldsymbol{\theta}^1 - \mu\cdot\frac{\partial\mathcal{L}_2}{\partial\boldsymbol{\theta}^1},\\ \boldsymbol{\theta}^2 &:= \boldsymbol{\theta}^2 - \mu\cdot\frac{\partial\mathcal{L}_2}{\partial\boldsymbol{\theta}^2}.\end{aligned} \tag{12}$$

Finally, we update $\{\boldsymbol{C}_i^1, \boldsymbol{C}_i^2, \boldsymbol{C}_i\}$ with the hyperparameter α as follows:

$$\begin{aligned}\boldsymbol{C}_i^1 &:= \boldsymbol{C}_i^1 - \alpha\cdot\frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{C}_i^1},\\ \boldsymbol{C}_i^2 &:= \boldsymbol{C}_i^2 - \alpha\cdot\frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{C}_i^2},\\ \boldsymbol{C}_i &:= \boldsymbol{C}_i - \alpha\cdot\frac{\partial\mathcal{L}_{cv-cl}}{\partial\boldsymbol{C}_i}.\end{aligned} \tag{13}$$

E. Iterative Optimization

Simultaneously learning CV-EC and CV-CL in a united framework is difficult. The optimization processes of CV-EC and CV-CL affect each other because of different convergence speeds. As the deep features from changing viewpoints are disparate, the joint optimization may become diverse with a large learning rate or result in a locally optimal output with a small learning rate. Further, balancing the view-specific networks and the proposed cross-view constraints in a united framework will be difficult because the search space of the regularization coefficients will increase quadratically.

Training view-specific networks with either CV-EC or CV-CL separately is relatively easy. We can extract the discriminative features from the updated view-specific networks. With a substantially reduced cross-view intra-class distance, the view-specific deep networks optimized by CV-EC or CV-CL will be a better initialization for the other constraint. On the basis of the above observations, we propose an iterative algorithm to update the parameters of the view-specific networks.

The iterative optimization algorithm is presented in Algorithm 1. In particular, the parameters of view-specific networks are first optimized by either CV-EC or CV-CL, and then the updated models are utilized as the initialization of the other cross-view constraint. Through iterative optimization, we can optimize the view-specific deep networks from coarse to fine.

F. Multi-View CV-EC and CV-CL

In real surveillance systems, a large number of cameras are placed in different positions. We need to develop an effective framework for multi-view re-id tasks. In the case of more than two camera views, a natural solution is to generate multiple view-specific networks. Then, we can learn a one-to-one cross-view constraint for each pair of the view-specific networks. However, the method mentioned above has two drawbacks. First, the simultaneous training of multiple deep

Algorithm 1 Iterative optimization algorithm

Require: Training samples $\{\boldsymbol{x}_{ip}^1, \boldsymbol{x}_{iq}^2\}$, hyperparameter α, λ_1, and λ_2, learning rate μ;
Ensure: Optimal view-specific parameters $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$;

```
Initialize the parameters of the view-specific networks with the pre-trained model on the re-id dataset;
repeat
    repeat
        Compute joint loss: L_1 = sum_{v=1}^{2} L_s^v + λ_1 L_{cv-ec};
        Compute (∂L_{cv-ec}/∂x_{ip}^1, ∂L_{cv-ec}/∂x_{iq}^2) for each i by Eq.(7);
        Update (θ^1, θ^2) using Eq.(8)-Eq.(9);
    until L_1 < ε_1;
    Compute (C_i^1, C_i^2, C_i) for each i with (θ^1, θ^2);
    repeat
        Compute joint loss: L_2 = sum_{v=1}^{2} L_s^v + λ_2 L_{cv-cl};
        Compute (∂L_{cv-cl}/∂x_{ip}^1, ∂L_{cv-cl}/∂x_{iq}^2) for each i by Eq.(10);
        Update (θ^1, θ^2) using Eq.(12);
        Get (∂L_{cv-cl}/∂C_i^1, ∂L_{cv-cl}/∂C_i^2, ∂L_{cv-cl}/∂C_i) for each i by Eq.(11);
        Update (C_i^1, C_i^2, C_i) for each i using Eq.(13);
    until L_2 < ε_2;
    L = L_1 + L_2;
until L < ε;
```

networks requires a large amount of computational resources. The required resources increase proportionally to the number of cameras. Second, for a system with N views, we need to optimize C_N^2 one-to-one objective loss functions during training. Optimizing a model with so many loss functions is difficult.

To solve this problem, we propose a simple but efficient iterative optimization algorithm. For each camera view, samples from all other views can be considered to be from "the other view". Then, the multi-view re-id problem can be regarded as multiple one-to-others cross-view re-id problems. We can apply Algorithm 1 to solve these problems.

The multi-view optimization method is presented in Algorithm 2. For each view, we obtain two deep networks, one containing the view-generic information and the other containing the general information. We call the above networks the view-specific network and the public network, respectively. Then, the parameters of the public network are used for the initialization of the next iteration. Finally, we obtain all the view-specific networks.

Algorithm 2 Multi-view optimization algorithm

Require: Multi-view deep features;
Ensure: Optimal multi-view parameters $(\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^N, \boldsymbol{\theta})$;

```
Initialize the view-specific networks and the public network with the pre-trained model, set v = 1;
repeat
    repeat
        Update (θ^v) and (θ) through CV-EC, loss L_1;
        Update (θ^v) and (θ) through CV-CL, loss L_2;
        Compute total error: L = L_1 + L_2;
    until L < ε;
    v := v + 1
until v > N;
```

IV. Experiment

We have conducted extensive experiments on several public re-id datasets to validate the effectiveness of the proposed framework. In Section IV-A, we introduce the datasets, the evaluation protocols, and the implementation details. In Section IV-B, we compare the proposed approach with the state-of-the-art methods. In Section IV-C, we evaluate the components of the proposed approach in detail.

A. Experimental Settings

Datasets. In this study, we evaluate the proposed method on five re-id benchmarks: VIPeR [67], CUHK01 [68], CUHK03 [24], SYSU-mReID [69], and Market-1501 [70]. ***VIPeR*** contains 1,464 images of 632 pedestrians captured in two camera views. This dataset is challenging for deep models because of the poor image quality and the lack of training samples. ***CUHK01*** contains 3,884 images of 971 identities from two outdoor cameras. Each identity has two images per view. The image quality of the pedestrians in CUHK01 is better than that in VIPeR. ***CUHK03*** is one of the largest re-id benchmarks in the existing literature. It contains more than 14,000 images of 1,467 people captured from six cameras. The images of each person are from two disjoint cameras. The pedestrians are cropped automatically or manually. ***SYSU-mReid*** is customized for multi-shot re-id. It contains more than 24,000 images of 502 people from two cameras. SYSU-mReid is challenging for the existing methods. ***Market-1501*** is a high-quality multi-view dataset for re-id. It contains 32,668 annotated bounding boxes of 1,501 identities from six cameras in an open system. The images are automatically detected by a deformable parts model detector.

Evaluation protocols. We use the standard protocol to ensure fair comparisons between the proposed method and the state-of-the-art methods. The test protocols are as follows. (**I**) For VIPeR, CUHK01, and SYSU-mReID benchmarks, we randomly split the datasets into two parts. Half the identities are used for training, while the rest identities are used for testing. The cumulative matching characteristic (CMC) is used to evaluate the performance of the compared methods. (**II**) For CUHK03, we follow the standard protocol used by [13]: repeat 20 times to randomly split the samples into 100 people for testing and the remainder for training. We randomly select one image from the gallery for each identity and use all the images in the probe set to obtain the CMC curves. The evaluation process is repeated 100 times[1], and the average value is computed as the final result. (**III**) For Market-1501, the standard protocol is defined by [70]: train the models with the fixed training set (750 identities) and match 3,368 query images with the fixed testing set (751 identities). We conduct comparisons by using both single-query and multi-query (using multiple probe and query images) settings. Rank-1 accuracy and mean average precision (mAP) are computed to evaluate the performance of all the methods.

[1] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

Fig. 4. Examples from various re-id benchmarks. Images in a column correspond to the same identity.

TABLE I
DETAILS OF DEEP NETWORKS

Alexnet		JSTL_DGD	
Layer name	Output size	Layer name	Output size
input	227×227×3	input	144×56×3
conv1	55×55×96	conv1	144×56×32
pool1	27×27×96	conv2	144×56×32
conv2	27×27×256	conv3	144×56×32
pool2	13×13×256	pool3	72×28×32
conv3	13×13×384	inception (4a)	72×28×256
conv4	13×13×384	inception (4b)	72×28×384
conv5	13×13×256	inception (5a)	36×14×512
pool5	6×6×256	inception (5b)	36×14×768
fc6	4096	inception (6a)	36×14×1024
fc7	4096	inception (6b)	36×14×1536
fc8	512	fc7	256
fc9	M	fc8	M

Implementation details. We adopt Alexnet [17] and JSTL_DGD [71] as the baseline models. The implementation details of Alexnet and JSTL_DGD are presented in Table I. For Alexnet, we add an additional full connection layer before the softmax loss layer to form a deeper network. For JSTL_DGD, we implement the baseline model with the source code available online[2]. Note that for the baseline models, a single network is trained and shared for disjoint views. The proposed methods are implemented using the Caffe CNN Library [72]. The parameters are optimized in the stochastic gradient descent manner with a momentum of 0.9, weight decay of 10^{-4}, learning rate μ of 10^{-4}, and hyperparameter α of 10^{-3}. We set the regularization coefficients $\lambda_1 = \lambda_2 = 0.1$ for all the datasets. For Alexnet, the parameters are initialized with the model in [17]. For JSTL_DGD, the parameters are initialized with the model in [71].

B. Comparisons with the State-of-the-Art Models

In this section, we compare the proposed framework with the state-of-the-art approaches for both cross-view and multi-view (>2) re-id tasks. The evaluation was carried out on datasets from two cameras, such as VIPeR, CUHK01, and SYSU-mREID, and datasets from multiple cameras, such as CUHK03 and Market-1501.

(1) Comparisons on VIPeR. We compare the proposed approach with 18 methods on VIPeR, including metric learning-based models, view-specific models, and deep models. The details of the comparisons are presented in Table II. From Table II, we can see that the proposed view-specific deep network-based framework is competitive against other models. The experimental results validate that CV-EC is effective in improving the existing deep networks even without a sufficient number of training samples. The proposed approach manages to increase the rank-1 accuracy of JSTL_DGD from 47.2% to 51.9%. Because only one sample per view is used for each identity, CV-CL and ICV-ECCL cannot be utilized for VIPeR. With an increasing number of training samples, we can implement view-specific networks with ICV-ECCL and achieve better recognition accuracy. The recognition accuracy of the proposed model is lower than that of CRAFT-MFA and SSM because of the lack of training samples. Moreover, both CRAFT-MFA and SSM benefit from the fusion of multiple features. Our work outperforms CRAFT-MFA by 1.6% and SSM by 1.17% in terms of the rank-1 accuracy calculated using the single features. We can further improve the performance of the view-specific networks by using the feature fusion strategy.

TABLE II
COMPARISONS ON VIPER

Rank (%)	1	5	10	20
KISSME [1]	22	-	68	-
rKPCCA [11]	22.3	55.5	72.4	86
LFDA [4]	24.2	52	67.1	82
CPDL [15]	34.0	64.2	77.5	88.6
MLACM [59]	34.9	59.3	70.2	81.8
Gated-SCNN [28]	37.8	66.9	77.4	-
Deep Ranking [29]	38.4	69.2	81.3	90.4
XQDA [6]	40	68.1	80.5	91.1
WARCA [73]	40.2	68.2	80.7	91.1
S-LSTM [27]	42.4	68.7	79.4	-
HVIL-RMEL [74]	42.4	72.6	83	90.4
CVDCA [14]	43.3	72.7	83.5	92.2
Metric Ensemble [7]	45.9	77.5	88.9	95.8
TCP [31]	47.8	74.7	84.8	89.2
FFN Net [75]	51.1	81	91.4	96.9
DNS [76]	51.2	82.1	90.5	95.9
SSM [77]	50.73	-	90.0	95.6
SSM (Fusion) [77]	53.7	-	91.5	96
CRAFT-MFA [13]	50.3	80.0	89.6	95.5
CRAFT-MFA (+LOMO) [13]	**54.2**	**82.4**	**91.5**	**96.9**
JSTL_DGD [71]	47.2	73.4	80.4	86.4
JSTL_DGD+CV-EC	51.9	76.6	85.4	93.4

(2) Comparisons on CUHK01. We manage to learn view-specific deep networks with ICV-ECCL on CUHK01. Table III shows the comparisons between the proposed framework and the other state-of-the-art methods. The proposed framework outperforms the compared methods. We notably improve the highest rank-1 accuracy from 78.8% (by CRAFT-MFA) to 83.5%. Further, our work outperforms CRAFT-MFA by 9% in terms of the rank-1 accuracy calculated using the single features. Compared with the original JSTL_DGD model, a 7% improvement in the rank-1 accuracy is observed. The proposed approach also improves the performance of Alexnet by 13.1% in the rank-1 accuracy.

(3) Comparisons on CUHK03. Deep networks can achieve excellent performance with a sufficient number of training samples from CUHK03. Table IV shows the comparisons of the proposed method and the other approaches. ICV-ECCL achieves the highest rank-1 accuracy and surpasses the best alternative (CRAFT-MFA) by 1.1%. In particular, ICV-ECCL

[2]https://github.com/Cysu/dgd_person_reid

TABLE III
COMPARISONS ON CUHK01

Rank (%)	1	5	10	20
LMNN [44]	13.4	31.3	42.3	54.1
ITML [66]	16	35.2	45.6	59.8
rKPCCA [11]	16.7	41	54.1	67.7
Alexnet [17]	18.7	42.5	55.6	66.7
Alexnet+ICV-ECCL	31.8	58.4	70.2	80.6
XQDA [6]	63.2	83.9	90	94.9
CVDCA [14]	47.8	74.2	83.4	89.9
Deep Ranking [29]	50.4	75.9	84	91.3
Metric Ensemble [7]	53.4	76.4	84.4	90.5
WARCA [73]	65.6	85.3	90.5	95
TCP [31]	53.7	84.3	91	96.3
FFN Net [75]	55.5	78.4	83.7	92.6
DNS [76]	69.1	86.9	91.8	95.4
CRAFT-MFA [13]	74.5	91.2	94.8	97.1
CRAFT-MFA (+LOMO) [13]	78.8	92.6	95.3	97.8
JSTL_DGD [71]	76.5	92.4	95.3	97.3
JSTL_DGD+ICV-ECCL	**83.5**	**95.2**	**97.3**	**98.8**

TABLE IV
COMPARISONS ON CUHK03

Rank (%)	1	5	10	20
LMNN [44]	5.1	17.7	28.3	-
ITML [66]	6.3	18.7	29	-
KISSME [1]	11.7	33.3	48	-
BoW [70]	23	42.4	52.4	64.2
XQDA [6]	46.3	78.9	88.6	94.3
HVIL [74]	56.1	64.7	75.7	87.4
CVDCA [14]	47.8	74.2	83.4	89.9
Metric Ensemble [7]	62.1	89.1	94.3	97.8
S-LSTM [27]	57.3	80.1	88.3	-
WARCA [73]	75.4	94.5	97.5	99.1
Alexnet [17]	52.4	84.8	92.9	97.4
Alexnet+ICV-ECCL	69.9	91.8	97.3	99.1
TCP [31]	53.7	84.3	91	96.3
DNS [76]	54.7	84.8	94.8	95.2
Deep Histogram Loss [34]	65.8	92.9	97.6	99.4
SSM (Fusion) [77]	72.7	92.4	96	-
CRAFT-MFA [13]	84.3	97.1	98.3	99.1
CRAFT-MFA (+LOMO) [13]	87.5	97.4	98.7	99.5
JSTL_DGD [71]	83.4	97.1	98.7	99.5
JSTL_DGD+ICV-ECCL	**88.6**	**98.2**	**99.2**	**99.7**

outperforms CRAFT-MFA by 4.3% in terms of the rank-1 accuracy calculated using the single features. A significant improvement is also observed in the rank-1 accuracy of Alexnet from 52.4% to 69.9%. The experimental results prove that exploiting view information during feature extraction is effective for multi-view re-id tasks.

(4) Comparisons on SYSU-mREID. Table V shows the experimental results for SYSU-mREID. The proposed framework remarkably outperforms the state-of-the-art methods with a margin of 23.3% (64.1%-40.8%). With ICV-ECCL, the rank-1 accuracy of Alexnet/JSTL_DGD is notably improved from 35%/58.5% to 42%/65.1%. The experimental results indicate that the proposed framework is also effective for multi-shot re-id tasks.

(5) Comparisons on Market-1501. The experimental results for Market-1501 are presented in Table VI. The proposed framework remains competitive for a realistic protocol. ICV-ECCL achieves the highest rank-1 accuracy and mAP for both single-query and multi-query re-id tasks. The proposed

TABLE V
COMPARISONS ON SYSU-MREID

Rank (%)	1	5	10	20
KISSME [1]	16.9	39.8	54.7	69
rKPCCA [11]	22.3	53	68.1	83.7
LFDA [4]	26.2	55.6	68.8	80.3
MLACM [59]	30.8	55.5	67.4	77
CVDCA [14]	40.8	71.4	82.2	90.6
Alexnet [17]	35	62	73.7	84.3
Alexnet+ICV-ECCL	42	70	80.4	89
JSTL_DGD [71]	58.5	82.4	87.7	93.9
JSTL_DGD+ICV-ECCL	**65.1**	**86.4**	**91.7**	**95.4**

TABLE VI
COMPARISONS ON MARKET-1501

Query	Single Query		Multiple Query	
Evaluation models	Rank-1	mAP	Rank-1	mAP
KISSME [1]	40.5	19	-	-
BoW [70]	34.4	14.1	-	-
XQDA [6]	43.8	22.2	54.1	28.4
WARCA [73]	45.2	-	-	-
S-LSTM [27]	-	-	61.6	35.3
Deep Histogram Loss [34]	59.5	-	-	-
Gated-SCNN [28]	65.9	39.6	76	48.5
DNS [76]	61	35.7	71.6	46
IDE (R) [78]	77.1	63.6	-	-
SSM (Fusion) [77]	82.2	68.8	88.2	76.2
CRAFT-MFA [13]	68.7	42.3	77.0	50.3
CRAFT-MFA (+LOMO) [13]	71.8	45.5	79.7	54.3
JSTL_DGD [71]	81.6	56.6	83.1	68.1
JSTL_DGD+ICV-ECCL	**88.4**	**69.5**	**90.6**	**77.3**

framework outperforms the best model (SSM) in the existing literature by 6.2%/0.7% for rank-1/mAP in the single-query task and 2.4%/1.1% for rank-1/mAP in the multi-query task. For the baseline model (JSTL_DGD), ICV-ECCL exhibits an improvement of 6.8%/13.3% in rank-1/mAP in the single-query task and 7.5%/9.2% in rank-1/mAP in the multi-query task. The experimental results validate the effectiveness of the proposed approach in the case of a realistic testing protocol.

(6) Conclusion of comparisons. From the experimental results, we can conclude the following. First, learning view-specific deep networks with cross-view constraints can significantly improve the performance of the existing deep networks. ICV-ECCL is proved to be effective in enhancing the discriminative abilities of both Alexnet and JSTL_DGD. Second, the proposed framework can be adapted to various re-id benchmarks. Comparisons on VIPeR and CUHK01 validate that view-specific deep networks are competitive even without a sufficient number of training samples. With enough training data, a significant improvement can be observed on CUHK03. The proposed approach is also adaptive to the multi-shot re-id task on SYSU-mREID and the multi-view realistic re-id task on Market-1501. Third, the proposed framework outperforms the state-of-the-art methods with respect to several re-id benchmarks, including the CUHK01, CUHK03, SYSU-mREID, and Market-1501 datasets.

C. In-Depth Analysis of the Proposed Framework

In this section, we conduct a quantitative self-evaluation to verify the effectiveness of the proposed framework in detail.

TABLE VII
EVALUATION OF THE OVERALL FRAMEWORK

Dataset	VIPeR				CUHK01				CUHK03				SYSU-mREID				Market-1501			
Rank (%)	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
Alexnet [17]	-	-	-	-	18.7	42.5	55.6	66.7	52.4	84.8	92.9	97.4	34.9	62.0	73.7	84.3	-	-	-	-
Alexnet+ ICV-ECCL	-	-	-	-	31.8	58.4	70.2	80.6	69.9	91.8	97.3	99.1	42	70	80.4	89	-	-	-	-
JSTL_DGD [71]	47.2	73.4	80.4	86.4	76.5	92.4	95.3	97.3	83.4	97.1	98.7	99.5	58.5	82.4	89.2	93.9	81.6	92.3	95.3	97
JSTL_DGD+ ICV-ECCL	**51.9**	**76.6**	**85.4**	**93.4**	**83.5**	**95.2**	**97.6**	**98.8**	**88.6**	**98.2**	**99.3**	**99.7**	**65.1**	**86.4**	**91.7**	**95.4**	**88.4**	**94.8**	**96.5**	**98**

TABLE VIII
PERFORMANCE WITH RESPECT TO DIFFERENT CONSTRAINTS

Compared methods	Accuracy	Cross-view intra-class distance
Alexnet	52.4	1.27×10^5
Alexnet+CV-EC	66.56	3.74×10^3
Alexnet+CV-CL	64.72	4.81×10^3
Alexnet+Triplet	60.8	-
Alexnet+ICV-ECCL	69.94	1.93×10^3

TABLE IX
PERFORMANCE WITH RESPECT TO ITERATIVE OPTIMIZATION

Iteration step	Accuracy	Cross-view intra-class distance
Initial baseline	52.4	1.27×10^5
1st iter CV-EC	66.56	3.74×10^3
1st iter CV-CL	68.12	2.82×10^3
2nd iter CV-EC	69.33	1.96×10^3
2nd iter CV-CL	69.94	1.93×10^3

We not only evaluate the effect of our overall framework but also assess each component to measure its contribution.

Effectiveness of CV-EC and CV-CL. We train Alexnet with CV-EC and CV-CL on CUHK03 to evaluate their contributions. CV-EC and CV-CL are compared with the baseline model and the triplet loss [79]. The experimental results are presented in Table VIII. We can see that CV-EC and CV-CL surpasses the baseline model by 14.16% and 12.32%, respectively. The cross-view intra-class distance is reduced to 3.74×10^3 by CV-EC and 4.81×10^3 by CV-CL from 1.27×10^5 of the baseline model. Moreover, CV-EC/CV-CL outperforms the triplet loss by 5.75%/3.92%. The above observation indicates that the proposed framework is more effective than the state-of-the-art deep metric learning methods for the re-id task. Note that CV-EC minimizes the gaps between different view-specific features, while CV-CL learns the concentrated clusters around the centers. Therefore, CV-EC outperforms CV-CL in terms of both the recognition accuracy and the cross-view intra-class distance.

Effectiveness of the iterative optimization algorithm. Table IX shows the effect of the iterative learning algorithm. We can see that the recognition accuracy keeps increasing and the cross-view intra-class distance keeps decreasing during training. With iterative optimization, we obtain a better initialization for the view-specific networks after each step. The optimization process is thus from coarse to fine. Finally, we can obtain the desirable view-specific deep networks after several iterations.

Effectiveness of the overall framework. By comparisons with the baseline models, we quantitatively evaluate the effectiveness of the entire framework for the existing deep networks on various re-id benchmarks. The proposed framework is implemented over Alexnet and JSTL_DGD and tested on the VIPeR, CUHK01, CUHK03, SYSU-mREID, and Market-1501 datasets. The experimental results are presented in Table VII. A remarkable improvement compared with the baseline models is observed in the cases of all the datasets. For Alexnet, a performance gain of 17.5%/13.1%/7.02% in the rank-1 accuracy is observed for the CUHK03/CUHK01/SYSU-mREID dataset. For JSTL_DGD, a notable improvement of 4.75%/7.05%/5.23%/6.64%/6.83% in the rank-1 accuracy is obtained for the VIPeR/CUHK01/CUHK03/SYSU-mREID/Market-1501 dataset. Figure 5 shows the CMC curves of the overall framework and each component of the proposed approach on CUHK03. ICV-ECCL is clearly effective in improving the performance of Alexnet and JSTL_DGD. Compared with the baseline model, a significant improvement from each component of the proposed approach is also observed.

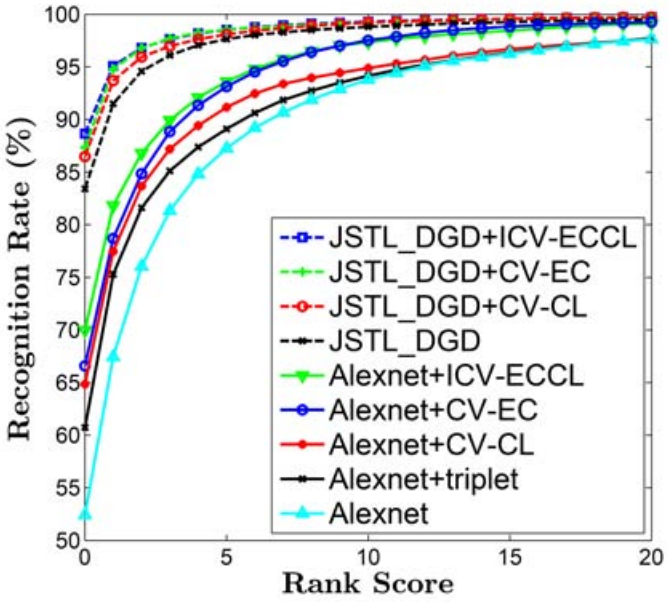

Fig. 5. Evaluation of the proposed framework on CUHK03.

Evaluation of the convergence of CV-EC and CV-CL. Figure 6 validates the convergence of CV-EC and CV-CL. We integrate CV-EC and CV-CL with Alexnet and train the view-specific networks on CUHK03. In Figure 6(a), we observe a fast convergence of CV-EC. The initial training loss from the baseline model is huge and drops dramatically after a

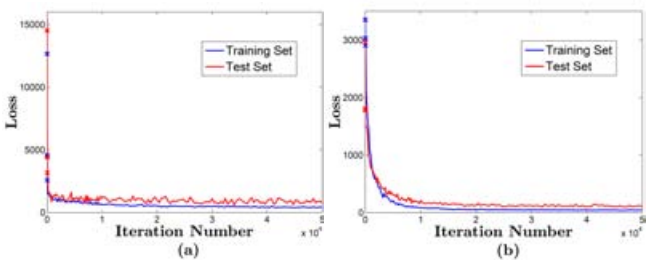

Fig. 6. The convergence of CV-EC and CV-CL. Figure 6(a) shows the training error and the test error of CV-EC, while Figure 6(b) shows the training error and the test error of CV-CL. The training loss and the test loss drop dramatically in the early iterations and keep decreasing for both CV-EC and CV-CL.

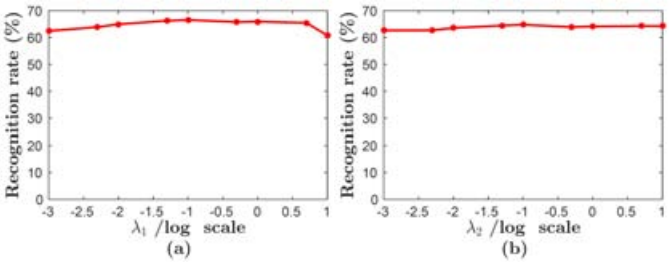

Fig. 7. Influence of λ_1 and λ_2 on the CUHK03 Dataset. Best performance is obtained when $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$.

few iterations. During training, the losses of both the training set and the test set keep decreasing. These decreases imply that the training process is convergent and the view-specific networks have excellent generalizability. Figure 6(b) illustrates the training process of CV-CL. The convergence of CV-CL is similar to that of CV-EC. Further, the training loss keeps decreasing along with the test loss. In the early stage of the optimization, the convergence speed of CV-CL is slower than that of CV-EC because more parameters are involved in CV-CL. Note that in the case of a small test loss, the generalizability of view-specific features generated by CV-CL is excellent.

Discussion on the regularization coefficients λ_1 and λ_2. Figure 7(a) shows the relationship between the rank-1 accuracy of CV-EC and the regularization coefficient λ_1. λ_1 is used to balance CV-EC and softmax. We can see that CV-EC achieves the best recognition accuracy when $\lambda_1 = 0.1$. Therefore, we set $\lambda_1 = 0.1$ for all the datasets. Similarly, Figure 7(b) shows that CV-CL achieves the best recognition accuracy when $\lambda_2 = 0.1$. As with λ_1, we set $\lambda_2 = 0.1$ for all the datasets.

Figure 7 shows that the performance of both CV-EC and CV-CL remains relatively stable across a wide range of regularization coefficients (from 10^{-3} to 10^{1}). We can conclude that implementing view-specific networks with cross-view constraints can improve the performance of deep features for most cases. Note that when $\lambda_1 = \lambda_2 = 0$, the cross-view constraints will be removed. Without the cross-view constraints, the performance of the view-specific networks will decrease significantly. In contrast, when the regularization coefficients are infinitely large, the view-specific softmax losses will be suppressed. The deep features extracted by the view-specific networks may manage to minimize the view discrepancies but fail to distinguish the identities. Finally, we also consider the relationship between the regularization coefficients of CV-EC and CV-CL. The regularization coefficients are independent because the proposed framework adopts an iterative optimization algorithm. This framework can achieve the best performance with the best alternative regularization coefficients of CV-EC and CV-CL.

V. Conclusion

Our research focuses on learning view-specific deep networks for re-id, which has been overlooked in the existing literature. We propose a novel framework called the iterative cross-view Euclidean constraint and center loss (ICV-ECCL). The proposed framework exploits view-specific information during low-level feature extraction to minimize the cross-view intra-class distance. Moreover, ICV-ECCL adopts an iterative optimization algorithm to iteratively optimize the parameters of the view-specific networks by using CV-EC and CV-CL. We further extend ICV-ECCL to a multi-view version to cope with benchmarks captured from multiple camera views.

The proposed framework is implemented over Alexnet and JSTL_DGD. We evaluate the effectiveness of ICV-ECCL on the VIPeR, CUHK01, CUHK03, SYSU-mREID, and Market-1501 benchmarks. The experiments validate that the proposed approach significantly improves the performance of both Alexnet and JSTL_DGD in all the datasets. Moreover, the proposed framework outperforms the state-of-the-art methods on the CUHK01, CUUK03, SYSU-mREID, and Market-1501 datasets. We can conclude that learning view-specific deep networks with ICV-ECCL is effective for the re-id task.

Acknowledgment

This project was supported by the NSFC (U1611461, 61573387, 61672544).

References

[1] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[2] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016, pp. 1268–1277.

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

[4] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013, pp. 3318–3325.

[5] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014, pp. 1–16.

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[7] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015, pp. 1846–1855.
[8] J. Jia and Q. Ruan, "Cross-view analysis by multi-feature fusion for person re-identification," in *ICSP*, 2016, pp. 107–112.
[9] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013, pp. 3594–3601.
[10] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE TMM*, vol. 18, no. 2, pp. 260–272, 2016.
[11] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012, pp. 2666–2672.
[12] W. He, Y. Chen, and J. Lai, "Cross-view transformation based sparse reconstruction for person re-identification," in *ICPR*, 2016, pp. 3410–3415.
[13] Y. Chen, X. Zhu, W. Zheng, and J. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE TPAMI*, vol. 40, no. 2, pp. 392–402, 2017.
[14] Y. Chen, W. Zheng, J. Lai, and P. Yuen, "An asymmetric distance model for cross-view feature mapping in person re-identification," *IEEE TCSVT*, vol. 27, no. 8, pp. 1661–1675, 2016.
[15] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *IJCAI*, 2015, pp. 2155–2161.
[16] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. Zhu, "Cross-view action modeling, learning and recognition," in *CVPR*, 2014, pp. 2649–2656.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[19] J. Hu, J. Lu, and Y. Tan, "Discriminative deep metric learning for face verification in the wild," in *CVPR*, 2014, pp. 1875–1882.
[20] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014, pp. 1891–1898.
[21] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015, pp. 4305–4314.
[22] Y. Kong, Z. Ding, J. Li, and Y. Fu, "Deeply learned view-invariant features for cross-view action recognition," *IEEE TIP*, vol. 26, no. 6, pp. 3028–3037, 2017.
[23] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014, pp. 34–39.
[24] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.
[25] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015, pp. 3908–3916.
[26] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.
[27] 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, pp. 135–153.
[28] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016, pp. 791–808.
[29] S. Chen, C. Guo, and J. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE TIP*, vol. 25, no. 5, pp. 2353–2367, 2016.
[30] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *ECCV*, 2016, pp. 475–491.
[31] 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, pp. 1335–1344.
[32] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE TIP*, pp. 3492–3506, 2017.
[33] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014, pp. 1386–1393.
[34] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *NIPS*, 2016, pp. 4170–4178.
[35] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.
[36] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.
[37] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *ICCV*, 2015, pp. 3200–3208.
[38] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *ICPR*, 2010, pp. 1413–1416.
[39] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, vol. 1, no. 2, 2011, pp. 1–11.
[40] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014, pp. 536–551.
[41] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372.
[42] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE TPAMI*, vol. 24, no. 4, pp. 509–522, 2002.
[43] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *ICCV*, 2007, pp. 1–8.
[44] K. Q. Weinberger and L. K. Saul, "Distance metric

learning for large margin nearest neighbor classification," *JMLR*, vol. 10, no. Feb, pp. 207–244, 2009.

[45] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[46] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *CVPR*, 2014, pp. 3550–3557.

[47] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *CVPR*, 2015, pp. 4184–4193.

[48] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *CVPR*, vol. 2, 2006, pp. 1528–1535.

[49] A. BedagkarGala and S. K. Shah, "Part-based spatio-temporal model for multi-person re-identification," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1908–1915, 2012.

[50] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *SIBGRAPI*, 2009, pp. 322–329.

[51] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *BMVC*, 2012, pp. 1–11.

[52] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[53] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *ICPR*, vol. 3, 2006, pp. 1204–1207.

[54] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *BMVC*, vol. 2, no. 5, 2010, pp. 1–11.

[55] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *ICCV*, 2013, pp. 3567–3574.

[56] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *CVIU*, vol. 117, no. 2, pp. 130–144, 2013.

[57] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," *ACCV*, pp. 501–512, 2011.

[58] R. Kawai, Y. Makihara, C. Hua, H. Iwama, and Y. Yagi, "Person re-identification using view-dependent score-level fusion of gait and color features," in *ICPR*, 2012, pp. 2694–2697.

[59] S. Shi, C. Guo, J. Lai, S. Chen, and X. Hu, "Person re-identification with multi-level adaptive correspondence models," *Neurocomputing*, vol. 168, pp. 550–559, 2015.

[60] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE TIP*, vol. 18, no. 7, pp. 1512–1523, 2009.

[61] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV Workshops Demonstrations*, 2012, pp. 413–422.

[62] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes." in *BMVC*, vol. 2, no. 3, 2012, pp. 1–12.

[63] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, "Attribute-restricted latent topic model for person re-identification," *Pattern Recognition*, vol. 45, no. 12, pp. 4204–4213, 2012.

[64] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *ICCV*, 2015, pp. 3739–3747.

[65] R. Layne, T. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *ECCV Workshops Demonstrations*, 2012, pp. 402–412.

[66] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007, pp. 209–216.

[67] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *VS-PETS Workshop*, vol. 3, no. 5, 2007, pp. 1–7.

[68] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012, pp. 31–44.

[69] C. Guo, S. Chen, J. Lai, X. Hu, and S. Shi, "Multi-shot person re-identification with automatic ambiguity inference and removal," in *ICPR*, 2014, pp. 3540–3545.

[70] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[71] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258.

[72] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACMMM*, 2014, pp. 675–678.

[73] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *ECCV*, 2016, pp. 875–890.

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

[75] S. Wu, Y. Chen, X. Li, A. Wu, J. You, and W. Zheng, "An enhanced deep feature representation for person re-identification," in *WACV*, 2016, pp. 1–8.

[76] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016, pp. 1239–1248.

[77] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, 2017, pp. 2530–2539.

[78] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 3652–3661.

[79] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

Zhanxiang Feng received the B.E. degree in automation from Sun Yet-Sen University, China, in 2012. He is currently pursuing the Ph.D. degree in information and communication engineering with Sun Yat-Sen University, China. His research interests include person re-identification, face recognition, face hallucination, image super-resolution, and visual surveillance.

Jianhuang Lai received his M.Sc. degree in applied mathematics in 1989 and his Ph.D. in mathematics in 1999 from Sun Yat-Sen University, China. He joined Sun Yat-sen University in 1989 as an Assistant Professor, where currently, he is a Professor in School of Data and Computer Science. His current research interests are in the areas of computer vision, pattern recognition and its applications. He has published over 250 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. Prof. Lai serves as a deputy director of the Image and Graphics Association of China and also serves as a standing director of the Image and Graphics Association of Guangdong. He is also the deputy director of Computer Vision Committee, China Computer Federation (CCF).

Xiaohua Xie received the B.S. degree in mathematics and applied mathematics from Shantou University in 2005, the M.S. degree in information and computing science and the Ph.D. degree in applied mathematics from Sun Yat-sen University, China, in 2007 and 2010, respectively. He was an Associate Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He is currently a Research Professor with Sun Yat-sen University. He has authored or co-authored over 30 papers in prestigious international journals and conferences. His current research fields cover image processing, computer vision, pattern recognition, and computer graphics.