IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 29, 2020, accepted September 10, 2020, date of publication September 15, 2020,
date of current version September 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024289

Heterogeneous Distance Learning Based on Kernel Analysis-Synthesis Dictionary for Semi-Supervised Image to Video Person Re-Identification

XIAOKE ZHU[1,2,3], (Member, IEEE), PENGFEI YE[1], XIAO-YUAN JING[3], (Member, IEEE), XINYU ZHANG[3], XIANG CUI[1], XIAOPAN CHEN[1], AND FAN ZHANG[1]

[1]Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475001, China
[2]School of Computer and Information Engineering, Institute of Data and Knowledge Engineering, Henan University, Kaifeng 475001, China
[3]School of Computer Science, Wuhan University, Wuhan 430072, China

Corresponding authors: Xiaopan Chen (xpchen@henu.edu.cn) and Fan Zhang (zhangfan@henu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0202001, in part by the NSFC-Key Project of General Technology Fundamental Research United Fund under Grant U1736211, in part by the National Nature Science Foundation of China under Grant 61672208, in part by the Key Scientific and Technological Project of Henan Province under Grant 192102210277, and in part by the Higher Education Institution Key Research Projects of Henan Province under Grant 19A520001.

ABSTRACT Image to video person re-identification (IVPR), i.e., matching between pedestrian video and image, is an important task in practice. Although several methods have been presented for IVPR, most of these methods investigate the IVPR problem under the supervised setting, and require a large number of labeled image-video pairs for training. In this article, we study the IVPR problem under the semi-supervised setting, and propose a Kernel Analysis-synthesis Dictionary based heterogeneous Distance Learning (KADDL) approach. Specifically, KADDL first learns two pairs of kernel analysis-synthesis dictionaries from the labeled and unlabeled training image-video data in the kernel space. With the learned dictionary pairs, the heterogeneous image and video features can be transformed into coding coefficients of the same representation space, such that the gap between image and video can be bridged. Then, KADDL learns a discriminative distance metric over the transformed coding coefficients, to make the coding coefficients of positive image-video pair become similar, while those of negative image-video pair dissimilar. To make better use of the unlabeled data, we further designed a reliability-based semi-supervised strategy for KADDL. Experiments on several publicly available pedestrian sequence datasets demonstrate the effectiveness of the proposed approach.

INDEX TERMS Semi-supervised image to video person re-identification, distance learning, kernel dictionary learning, coupled dictionary learning.

I. INTRODUCTION

Person re-identification (re-id) [1]–[4] is a key task in many safety-critical applications, such as automated video surveillance and forensics, and has attracted lots of research interests in the machine learning and computer vision communities. Given a pedestrian image/video captured by one camera, person re-id aims to identify the same person from images/videos captured by other non-overlapping cameras [5]–[7]. Due to the existence of factors such as changes in illumination, viewpoint, occlusion and resolution, there usually exist large differences between the images/videos captured by different cameras in practice, which makes the person re-id across cameras a challenging task.

To relieve the difficulties existed in person re-id task, a series of methods have been proposed in recent years [7]–[10]. According to the used pedestrian representation in the matching process, existing person re-id methods can be mainly divided into three categories: image-based person re-id, video-based person re-id, and Image to Video Person

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

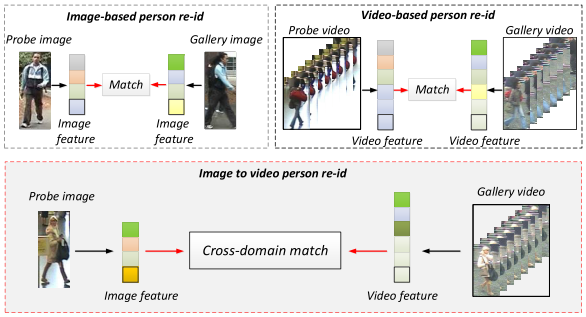


FIGURE 1. Difference between image/video-based person re-id and image to video person re-id.

Re-id (IVPR). Image-based person re-id methods represent each pedestrian with a single image, and focus on the matching between images across cameras. Video-based person re-id methods use the video clip to represent each pedestrian, and focus on the re-identification between videos from different camera views. In either image-based or video-based person re-id methods, the samples to be matched are homogeneous, i.e., image versus image or video versus video.

In many real-world scenarios, the query object may be just one single image, while the gallery set consists of large quantities of surveillance videos [11]–[13]. One instance is locating lost person from the surveillance videos according to the person's image. Another instance is rapidly looking for clues of criminal suspects among a large number of surveillance videos according to one photo of the suspect. In these cases, person re-id has to be conducted between image and video (i.e., IVPR). Therefore, IVPR mainly focuses on the matching between pedestrian image and video clip. In practice, the information contained in image and video are usually inconsistent, which further increases the difficulty of matching between pedestrian image and video. Figure 1 illustrates the differences between image/video-based person re-id and image to video person re-id. In this article, we mainly focus on the problem of image to video person re-id.

A. MOTIVATION

Recently, a series of methods have been presented to investigate the problem of image to video person re-id, and achieved interesting results [11]–[19]. Most of these methods focus on training a discriminative matching model by using a large number of labeled pedestrian image-video pairs. In practice, it's usually time-consuming and expensive to collect large quantities of labeled image-video pairs from non-overlapping cameras, which will limit the application of these methods in real environment. Therefore, we try to investigate the IVPR problem under the semi-supervised setting in this article, to relieve the requirement for the amount of labeled image-video pairs.

Distance learning is an effective technique for identification or verification tasks, such as image/video-based person re-identification [20]–[22], kinship verification [23], and so on. Intuitively, we can use the distance learning technique to solve the difficulty of large within-class variation existed in IVPR. However, a pedestrian video clip usually contains more useful information (e.g., spatial-temporal information) than a single image, which means that the features extracted from video and image are usually heterogeneous (different physical properties and dimensions). This will lead to the result that directly learning distance metric from heterogeneous image and video features usually cannot produce desirable re-identification performance. Therefore, we need to bridge the gap between image and video features before learning distance metric.

Coupled dictionary learning (CDL) is an effective technique to bridge the differences across domains, and has been successfully applied to many machine learning and computer vision tasks [24]–[26]. By learning a pair of dictionaries, CDL can transform the samples from different domains into coding coefficients of the same representation space. Inspired by these works, we can borrow the idea of coupled dictionary learning to bridge the gap between image and video features. However, there is still an important issue to be considered in the learning process. Due to the variations of viewpoint, illumination and occlusion existed in the capturing process, there is no guarantee that the features of images and videos lie in a linear space. Considering that kernel learning is an effective technique to cope with non-linear data, we can combine the coupled dictionary learning and kernel technique to reduce the image-to-video gap more effectively.

Motivated by the above analyses, we intend to investigate the semi-supervised image to video person re-id problem, by combining the distance learning, coupled dictionary learning and kernel learning techniques.

B. CONTRIBUTION

The major contributions of this article are summarized as follows:

- We make the first attempt to investigate the image to video person re-id in the semi-supervised setting, and provide an effective solution.
- We propose a Kernel Analysis-synthesis Dictionary based heterogeneous Distance Learning (KADDL) approach. To solve the matching between image and video, KADDL first transforms the heterogeneous image and video features into coding coefficients of the same representation space by learning a pair of dictionaries from labeled and unlabeled data. Then KADDL learns a discriminative distance metric in the coding coefficient domain to facilitate the matching between image and video. To make better use of unlabeled data, KADDL designs a reliability-based semi-supervised strategy. The solution of KADDL for solving heterogeneous matching is novel.
- We combine the coupled analysis-synthesis dictionary learning and kernel learning techniques for the first time, which can ensure that the transformed

coding coefficients better reflect the intrinsic relationship between image and video.
- We have conducted extensive experiments on three publicly available pedestrian image sequence datasets, including iLIDS-VID, PRID 2011 and MARS. Experimental results have shown that the proposed approach can achieve very competitive or even higher performance than the compared methods by using less labeled data.

II. RELATED WORK

In this section, we briefly review three types of works that are most related to our approach: image to video person re-identification, distance learning and coupled dictionary learning.

A. IMAGE TO VIDEO PERSON RE-IDENTIFICATION

To relieve the difficulties existed in the matching between pedestrian image and video, several image to video person re-id methods have been presented [11], [14]–[18]. For example, Zhu *et al.* [11] proposed a joint feature projection matrix and heterogeneous dictionary pair learning approach, which jointly learns a pair of heterogeneous image and video dictionaries as well as an intra-video projection matrix to facilitate the matching between image and video. In [12], a Temporal Knowledge Propagation (TKP) method is presented, which can propagate the temporal knowledge learned by the video representation network to the image representation network, such that the in- formation asymmetry problem can be alleviated. In [13], Yu *et al.* presented a Cross-media Body-part Attention Network (CBAN), which employs CNN/LSTM to extract body part attention features from images/videos, and uses a media-pulling constraint term to alleviate the inherent cross-media gap. In [18], Xie *et al.* presented an end-to-end neural network, which employs the image captioning and video captioning models to project the learned features into a coordinated space, such that the similarity of image and video can be calculated. In [17], Li *et al.* used the mean shift to extract the salient region from each person image, and clustered all salient regions by least-squares log-density gradient clustering. Finally, the distance between the probe salient region and the gallery clustered salient regions are computed and used for re-identification.

Although these methods have relieved some difficulties existed in image to video person re-id to some extent, they all need a large number of labeled image-video pairs in the training process. In practice, collecting large quantities of labeled image-video pairs from non-overlapping cameras is usually time-consuming and expensive.

To relieve the requirement for the quantity of labeled image-video pairs, Zhang *et al.* [19] presented a cross-modal feature generating and target information preserving transfer network, which transforms the features of unlabeled target sample into the source domain feature space while preserving target identity information, and uses a cross-modal loss term to eliminate the gap between pedestrian images and videos. By leveraging the labeled source dataset, this work reduced the requirement for labeled target data to some extent. However, the performance of this method is rather poor comparing to the supervised image to video person re-identification methods. The possible reason may be that: this work cannot effectively capture the intrinsic characteristic of target data without using the label information.

Different from the existing supervised or unsupervised image to video person re-id methods, our approach tries to solve this problem under the semi-supervised settings. Specifically, our approach trains the learning model by employing a small amount of labeled data as well as a large number of unlabeled data, and uncovers the intrinsic relationship between image and video data in the kernel space.

B. DISTANCE LEARNING

Distance learning technique has been successfully applied in many computer vision tasks [27]–[30]. Distance learning based person re-identification methods mainly focus on seeking an optimal distance metric, under which the distance between truly matching images is small, but the distance between wrong matching images is large [31]–[34]. For example, in [35], Davis *et al.* formulated distance learning as a LogDet optimization problem, which enforces the positive semidefinite constraint automatically to avoid the projection onto the positive semidefinite cone. In [36], Weinberger *et al.* presented the large-margin nearest neighbor (LMNN) method, which aims to pull the neighbors of the same class together while push the neighbors from different classes far away. In [37], Hirzer *et al.* relaxed the positive semi-definiteness constraint to dramatically simplify the problem of learning a Mahalanobis distance metric. In [38], a probabilistic relative distance comparison (PRDC) method is presented to maximize the probability of a truly matching pair having a smaller distance than a wrong matching pair. In [39], Liao *et al.* used the generalized Rayleigh quotient to search a discriminating low-dimensional subspace, where the distance learning can be executed more efficiently. In [22], Nguyen *et al.* presented a distance metric learning method by incorporating kernels into the KISSME [40] method, which allows the distance metric to be learned in a nonlinear feature space induced by a kernel function.

The major differences between our approach and the mentioned distance learning methods are as follows. (i) These methods are designed for homogeneous matching (i.e., the objects participating in the matching are described with the same feature descriptor), while our approach is designed for heterogeneous matching. (ii) These methods learn the distance metric in the original feature space. Different from them, our approach first learns a pair of dictionaries. With which, heterogeneous image and video features can be transformed into coding coefficients in the same representation space, and then our approach learns a discriminative distance metric over the coding coefficients.

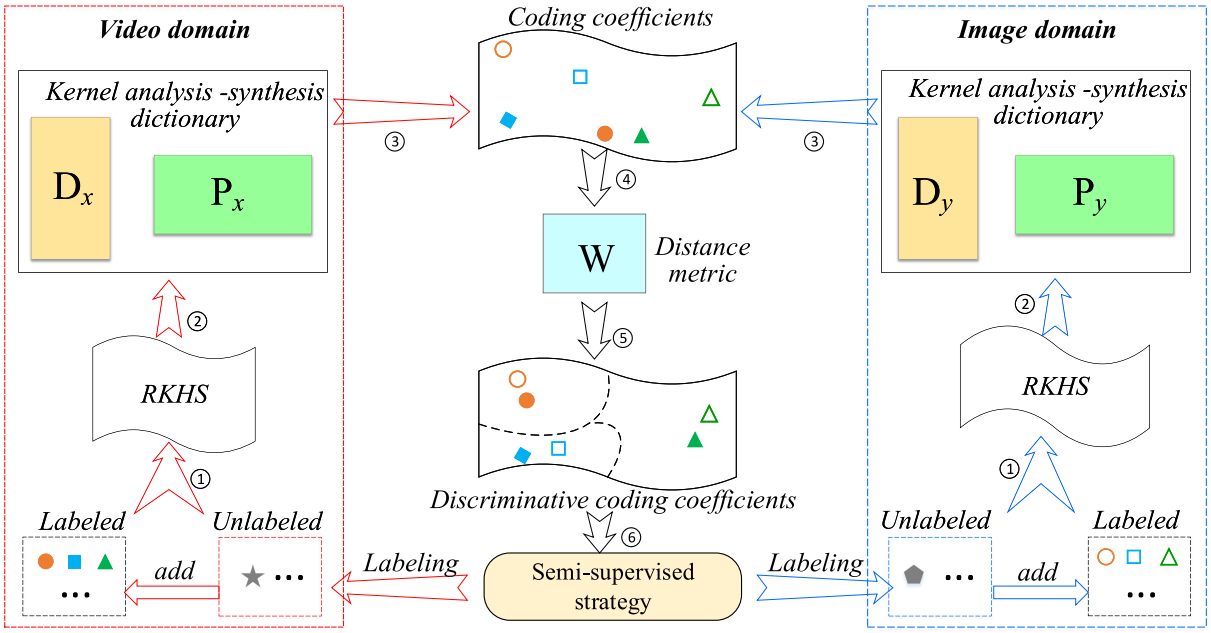


FIGURE 2. Illustration of the basic idea of our approach.

C. COUPLED DICTIONARY LEARNING

As an effective technique for bridging the gap between different modalities, coupled dictionary learning has attracted lots of attention in recent years [41]–[43]. Recently, coupled dictionary learning technique has also been successfully applied to person re-identification. In [44], a semi-supervised coupled dictionary learning (SSCDL) approach is presented for person re-identification, which learns a pair of dictionaries for two camera views. In [45], a cross-view projective dictionary learning (CPDL) approach is presented, which learns effective features for persons across different views. In [46], Jing *et al.* proposed a semi-coupled low-rank discriminant dictionary learning (SLD^{2}L) approach for super-resolution person re-identification, which learns a pair of dictionaries from the training HR and LR images. In [25], a discriminative semi-coupled projective dictionary learning (DSPDL) model is presented, which jointly learns a pair of dictionaries and a mapping to bridge the gap across low and high resolution person images.

The existing coupled dictionary learning methods have relieved the difficulties existed in person re-identification across camera views to some extent. However, all these methods learn dictionaries from training data in the linear manner, but ignore the fact that pedestrian samples usually lie in a non-linear feature space, leading to that the learned dictionaries may not be able to reflect the intrinsic relationship between different camera views. Different from the existing coupled dictionary learning methods, our approach employs the kernel technology to cope with the non-linearity issue, and learns kernel dictionaries from the training data. In addition, our approach learns a discriminative distance metric over the coding coefficients of image and video.

III. THE PROPOSED APPROACH

A. PROBLEM FORMULATION

From the analysis in the motivation part of Section I, we know that there are four intrinsic difficulties existed in image and video data. (i) The features extracted from video and image are usually inconsistent (e.g., physical property and feature dimension); (ii) The extracted image/video features may not lie in a linear feature space. (iii) There usually exist large variations between the image and video of the same person. (iv) Labeling image-video pairs across cameras is time-consuming and expensive. All these difficulties will directly influence the performance of matching between image and video.

To deal with the above difficulties, we propose a Kernel Analysis-synthesis Dictionary based heterogeneous Distance Learning (KADDL) approach. Figure 2 illustrates the basic idea of our approach. (i) To deal with the inconsistency between image and video features, KADDL borrows the idea of coupled dictionary learning and jointly learns view-specific dictionaries for image and video data. Over the learned dictionaries, the heterogeneous image and video data can be transformed into coding coefficients in the same representation space. In this way, the inconsistency between image and video data can be reduced to some extent. Since analysis-synthesis dictionary could provide a more complete view of data representation than analysis dictionary or synthesis dictionary [47], [48], the proposed KADDL approach utilizes the analysis-synthesis dictionary learning technique as the basis in the coupled dictionary learning process. (ii) To cope with the non-linearity issue, KADDL combines the coupled dictionary learning and kernel technology together, such that the learned dictionaries can better characterize

TABLE 1. Notations used in our approach.

Notation	Description
X and Y	the feature sets of training pedestrian images and videos
X_L and Y_L	the labeled training data in X and Y
X_U and Y_U	the unlabeled training data in X and Y
$x_{l,i}$ and $x_{u,i}$	the feature vector of the i^{th} labeled or unlabeled pedestrian image
$y_{l,i}$ and $y_{u,i}$	the feature vector of the i^{th} labeled or unlabeled pedestrian video
$\Phi_x(\cdot)$ and $\Phi_y(\cdot)$	the kernel mapping functions for the image and video features
$K_x(\mathbf{X},\mathbf{X})$, $K_y(\mathbf{Y},\mathbf{Y})$	the kernel Gram matrices of the image and video features
$\Phi_x(\mathbf{X})\mathbf{D}_x$, $\Phi_y(\mathbf{Y})\mathbf{D}_y$	the learned kernel synthesis dictionaries for image and video features
$\mathbf{P}_x\Phi_x(\mathbf{X})^T$, $\mathbf{P}_y\Phi_y(\mathbf{Y})^T$	the learned analysis dictionaries for images and videos
$\mathbf{W}$	the learned distance metric over the representation space

the non-linear feature spaces of image and video. (iii) To deal with the large within-class variations, KADDL learns a discriminative distance metric, under which the distance between image and video of the same person gets close, while that between image and video from different persons becomes far apart. (iv) To relieve the issue of labeling image-video pairs, KADDL designs a simple but efficient semi-supervised strategy, which enables KADDL to be trained by using little labeled data together with large amount of unlabeled data.

Let $\mathbf{X} = [\mathbf{X}_L, \mathbf{X}_U]$ and $\mathbf{Y} = [\mathbf{Y}_L, \mathbf{Y}_U]$ represent the features of training pedestrian images and videos, respectively, where $\mathbf{X}_L \in \mathbb{R}^{p\times n1}$ and $\mathbf{Y}_L \in \mathbb{R}^{q\times n2}$ are labeled, while $\mathbf{X}_U \in \mathbb{R}^{p\times n3}$ and $\mathbf{Y}_U \in \mathbb{R}^{q\times n4}$ are unlabeled. Here, $n1$, $n2$, $n3$ and $n4$ are the sample numbers of the corresponding subsets, p and q are the dimensions of image and video features, respectively. Denote by $x_{l,i} \in \mathbb{R}^p$ ($y_{l,i} \in \mathbb{R}^q$) the feature vector of the i^{th} labeled pedestrian image (video), and denote by $x_{u,i} \in \mathbb{R}^p$ ($y_{u,i} \in \mathbb{R}^q$) the feature vector of the i^{th} unlabeled pedestrian image (video). By embedding image and video features into Reproducing Kernel Hilbert Space (RKHS), coupled dictionary learning can be conducted in RKHS. Let $\Phi_x(\cdot)$ and $\Phi_y(\cdot)$ represent the kernel mapping functions for the image and video features, respectively. Let $K_x(\mathbf{X},\mathbf{X}) = \Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X})$ and $K_y(\mathbf{Y},\mathbf{Y}) = \Phi_y(\mathbf{Y})^T\Phi_y(\mathbf{Y})$ separately denote the kernel Gram matrices of the image and video features. Similar to [48], we use $\Phi_x(\mathbf{X})\mathbf{D}_x$ and $\Phi_y(\mathbf{Y})\mathbf{D}_y$ to separately represent the learned kernel synthesis dictionaries for image and video features, and use $\mathbf{P}_x\Phi_x(\mathbf{X})^T$ and $\mathbf{P}_y\Phi_y(\mathbf{Y})^T$ to denote the learned analysis dictionaries for images and videos, respectively. Here, $\mathbf{D}_x \in \mathbb{R}^{N_1\times m}$, $\mathbf{D}_y \in \mathbb{R}^{N_2\times m}$, $\mathbf{P}_x \in \mathbb{R}^{m\times N_1}$ and $\mathbf{P}_y \in \mathbb{R}^{m\times N_2}$, where $N_1 = n_1 + n_3$, $N_2 = n_2 + n_4$, and m represents the dictionary size. In addition, we denote $\mathbf{W}$ as the learned distance metric over the representation space spanned by the coding coefficients. Table 1 summarizes the notations used in our approach.

Based on these symbols, we design the objective function of our KADDL approach as follows:

$$\begin{aligned}\min_{\mathbf{D}_x,\mathbf{D}_y,\mathbf{P}_x,\mathbf{P}_y,\mathbf{W}} & f(\mathbf{D}_x,\mathbf{D}_y,\mathbf{P}_x,\mathbf{P}_y,\mathbf{X},\mathbf{Y}) \\ & + \alpha g(\mathbf{W},\mathbf{P}_x,\mathbf{P}_y,\mathbf{X}_L,\mathbf{Y}_L) \\ s.t.\ & \|d_{x,i}\|_2^2 \le 1,\ \|d_{y,i}\|_2^2 \le 1,\ \|w_i\|_2^2 \le 1, \quad \forall i, \end{aligned} \tag{1}$$

where α is a balancing factor, and $d_{x,i}$ ($d_{y,i}$, w_i) denotes the i^{th} column of $\mathbf{D}_x$ ($\mathbf{D}_y$, $\mathbf{W}$). The constraint is used to restrict the energy of each column vector in $\mathbf{D}_x$, $\mathbf{D}_y$ and W, such that the updating process will become more stable. Details of $f(\cdot)$ and $g(\cdot)$ are as follows.

■ $f(\cdot)$ is the coupled kernel analysis-synthesis dictionary learning term, which aims to transform the heterogeneous image and video features into coding coefficients of the same representation space. The definition of $f(\cdot)$ is as follows:

$$f(\mathbf{D}_x,\mathbf{D}_y,\mathbf{P}_x,\mathbf{P}_y,\mathbf{X},\mathbf{Y}) = f_x(\mathbf{D}_x,\mathbf{P}_x,\mathbf{X}) + f_y(\mathbf{D}_y,\mathbf{P}_y,\mathbf{Y}), \tag{2}$$

where $f_x(\cdot)$ and $f_y(\cdot)$ are the reconstruction fidelity terms for image and video data, respectively.

$$\begin{aligned} & f_x(\mathbf{D}_x,\mathbf{P}_x,\mathbf{X}) \\ & = \|\Phi_x(\mathbf{X}_L) - \Phi_x(\mathbf{X})\mathbf{D}_x\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}_L)\|_F^2 \\ & \quad + \|\Phi_x(\mathbf{X}_U) - \Phi_x(\mathbf{X})\mathbf{D}_x\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}_U)\|_F^2, \end{aligned} \tag{3}$$

$$\begin{aligned} & f_y(\mathbf{D}_y,\mathbf{P}_y,\mathbf{Y}) \\ & = \|\Phi_y(\mathbf{Y}_L) - \Phi_y(\mathbf{Y})\mathbf{D}_y\mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(\mathbf{Y}_L)\|_F^2 \\ & \quad + \|\Phi_y(\mathbf{Y}_U) - \Phi_y(\mathbf{Y})\mathbf{D}_y\mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(\mathbf{Y}_U)\|_F^2. \end{aligned} \tag{4}$$

Here, $f_x(\mathbf{D}_x,\mathbf{P}_x,\mathbf{X})$ ($f_y(\mathbf{D}_y,\mathbf{P}_y,\mathbf{Y})$) is used to ensure that the learned kernel analysis-synthesis dictionary pair can well reconstruct the features of labeled and unlabeled pedestrian images (videos) in the kernel space. With the learned coupled dictionaries, the image and video features can be transformed into coding coefficients of the same representation space.

■ $g(\cdot)$ is the coding coefficient based distance learning term, which aims to facilitate the matching between the coding coefficients of image and video by learning a distance metric, i.e., the distance between the coefficients of truly matching image-video pair is smaller than that of wrong matching image-video pair. Definition of $g(\cdot)$ is shown in Eq. (5), where $< i, j, k >$ represents a triplet, which consists of a truly matching image-video pair ($x_{l,i}$, $y_{l,j}$) and a negative video $y_{l,k}$. G is the collection of constructed triplets, and λ is a balancing factor.

$$\begin{aligned} & g(\mathbf{W},\mathbf{P}_x,\mathbf{P}_y,\mathbf{X}_L,\mathbf{Y}_L) \\ & = \frac{1}{|G|}\sum_{<i,j,k>\in G}(\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(x_{l,i}) \\ & \quad - \mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(y_{l,j}))\|_2^2 \\ & \quad - \lambda\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(x_{l,i}) - \mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(y_{l,k}))\|_2^2). \end{aligned} \tag{5}$$

Semi-supervised strategy. In Eq. (1), unlabeled data is used to improve the representation ability of the learned dictionaries. In fact, there exists much more useful information

in unlabeled data. To make better use of the information contained in unlabeled data, we label the unlabeled data after each training epoch, and add the reliable newly labeled data into $\mathbf{X}_L$ and $\mathbf{Y}_L$. The definition of reliable newly labeled data is as follows. For unlabeled image $x_{u,i}$ and video $y_{u,j}$, they can be regarded as a reliable image-video pair only when the following two criteria are satisfied. (i) The matching probability between $x_{u,i}$ and $y_{u,j}$ is larger than the threshold. (ii) $x_{u,i}$ and $y_{u,j}$ are the top-k nearest neighbor of each other. Then the updated $\mathbf{X}_L$ and $\mathbf{Y}_L$ are used for the next epoch. Repeat this process until no unlabeled image-video pairs satisfy the above criteria.

However, since pseudo labels are estimated through the distance metric learned with a relatively small amount of labeled data, the accuracy of pseudo labels can be inferior due to the lack of enough labeled data. Researches in [49] also confirm this. Therefore, we adjust the effect of newly added pseudo-label samples in the objective function by measuring the reliability of the pseudo label. Eq. (5) is rewritten as:

$$\begin{aligned} &g(\mathbf{W}, \mathbf{P}_x, \mathbf{P}_y, \mathbf{X}_L, \mathbf{Y}_L) \\ &= \frac{1}{|G|} \sum_{<i,j,k>\in G} (\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(x_{l,i}) \\ &\quad - \mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(y_{l,j}))\|_2^2 \\ &\quad - \lambda\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(x_{l,i}) - \mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(y_{l,k}))\|_2^2) \\ &\quad + \frac{1}{|G_1|} \sum_{<i,j,k>\in G_1} r_{ij}(\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(x_{l,i}) \\ &\quad - \mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(y_{l,j}))\|_2^2 \\ &\quad - \lambda\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(x_{l,i}) - \mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(y_{l,k}))\|_2^2). \end{aligned} \tag{6}$$

where G_1 is the collection of newly constructed triplets, in which the truly matching image-video pair ($x_{l,i}$, $y_{l,j}$) is from the newly added pseudo-label sample pairs. r_{ij} represents the reliability of the newly added pseudo-label sample pair, which can be computed as $r_{ij} = exp^{-d_{ij}}$. Here, d_{ij} is the distance between the coefficients of $x_{l,i}$ and $y_{l,j}$ under the distance metric learned at the previous epoch. The smaller d_{ij} is, the larger the reliability r_{ij} will be. For the convenience of optimization, we simplify Eq. (6) to the following form:

$$\begin{aligned} g(.) &= \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_1^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_1^y)\|_F^2 \\ &\quad - \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_2^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_2^y)\|_F^2 \\ &\quad + \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_3^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_3^y)\|_F^2 \\ &\quad - \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_4^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_4^y)\|_F^2 \\ &= \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y)\|_F^2 \\ &\quad - \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y)\|_F^2 \end{aligned} \tag{7}$$

where $M_1^x(M_1^y, M_2^x, M_2^y, M_3^x, M_3^y, M_4^x, M_4^y)$ is a matrix with each column being $\frac{1}{\sqrt{|G|}}\Phi_x(x_{l,i})$ ($\frac{1}{\sqrt{|G|}}\Phi_y(y_{l,j})$, $\sqrt{\frac{\lambda}{|G|}}\Phi_x(x_{l,i})$, $\sqrt{\frac{\lambda}{|G|}}\Phi_y(y_{l,k})$, $\sqrt{\frac{r_{ij}}{|G_1|}}\Phi_x(x_{l,i})$, $\sqrt{\frac{r_{ij}}{|G_1|}}\Phi_y(y_{l,j})$, $\sqrt{\frac{\lambda r_{ij}}{|G_1|}}\Phi_x(x_{l,i})$, $\sqrt{\frac{\lambda r_{ij}}{|G_1|}}\Phi_y(y_{l,k})$) corresponding to the triplet $< i, j, k >$ in the collection G (G,G,G,G_1,G_1,G_1,G_1). $M_S^x = [M_1^x, M_3^x]$, $M_S^y = [M_1^y, M_3^y]$, $M_G^x = [M_2^x, M_4^x]$, $M_G^y = [M_2^y, M_4^y]$.

B. THE OPTIMIZATION ALGORITHM

Generally, the objective function in Eq. (1) is not convex. To optimize the problem in Eq. (1), we introduce two variables A and B, and relax Eq. (1) into the following problem:

$$\begin{aligned} \min_{\substack{\mathbf{D}_x,\mathbf{D}_y,\mathbf{P}_x,\mathbf{P}_y \\ \mathbf{W},A,B}} &\|\Phi_x(\mathbf{X}) - \Phi_x(\mathbf{X})\mathbf{D}_x A\|_F^2 + \|\Phi_y(\mathbf{Y}) - \Phi_y(\mathbf{Y})\mathbf{D}_y B\|_F^2 \\ &+ \alpha(\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y)\|_F^2 \\ &- \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y)\|_F^2) \\ &+ \tau(\|\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}) - A\|_F^2 \\ &+ \|\mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(\mathbf{Y}) - B\|_F^2) \\ s.t.\ &\|d_{x,i}\|_2^2 \le 1,\ \|d_{y,i}\|_2^2 \le 1,\ \|w_i\|_2^2 \le 1, \quad \forall i, \end{aligned} \tag{8}$$

where τ is a scalar constant.

The variables in Eq. (8) can be alternatively optimized by fixing the others when optimizing one of them. Specifically, we divide the objective function into four sub-problems, including updating $\{\mathbf{D}_x, \mathbf{D}_y\}$, updating $\{\mathbf{P}_x, \mathbf{P}_y\}$, updating $\{\mathbf{A}, \mathbf{B}\}$, and updating $\mathbf{W}$. The step-by-step optimization procedures are as follows.

Step 1: Update A and B. We first initialize $\mathbf{D}_x$, $\mathbf{D}_y$, $\mathbf{P}_x$, $\mathbf{P}_y$ and $\mathbf{W}$ as random matrices with unit Frobenius norm for each column vector. When other variables are fixed, A and B can be updated by:

$$\min_A \|\Phi_x(\mathbf{X}) - \Phi_x(\mathbf{X})\mathbf{D}_x A\|_F^2 + \tau\|\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}) - A\|_F^2, \tag{9}$$

$$\min_B \|\Phi_y(\mathbf{Y}) - \Phi_y(\mathbf{Y})\mathbf{D}_y B\|_F^2 + \tau\|\mathbf{P}_y\Phi_y(\mathbf{Y})^T\Phi_y(\mathbf{Y}) - B\|_F^2. \tag{10}$$

Let the derivative of Eq. (9) with respect to A be zero, the closed-form solution of Eq. (9) can be easily obtained.

$$\begin{aligned} A = &(\mathbf{D}_x^T\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X})\mathbf{D}_x + \tau I)^{-1} \\ &\times (\mathbf{D}_x^T\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}) + \tau\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X})) \end{aligned} \tag{11}$$

By substituting $K_x(\mathbf{X}, \mathbf{X}) = \Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X})$ into 11, the solution of A can be rewritten as:

$$\begin{aligned} A = &(\mathbf{D}_x^T K_x(\mathbf{X}, \mathbf{X})\mathbf{D}_x + \tau I)^{-1} \\ &\times (\mathbf{D}_x^T K_x(\mathbf{X}, \mathbf{X}) + \tau\mathbf{P}_x K_x(\mathbf{X}, \mathbf{X})) \end{aligned} \tag{12}$$

Similarly, the closed-form solution of Eq. (10) can be written as:

$$\begin{aligned} B = &(\mathbf{D}_y^T K_y(\mathbf{Y}, \mathbf{Y})\mathbf{D}_y + \tau I)^{-1} \\ &\times (\mathbf{D}_y^T K_y(\mathbf{Y}, \mathbf{Y}) + \tau\mathbf{P}_y K_y(\mathbf{Y}, \mathbf{Y})) \end{aligned} \tag{13}$$

Step 2: Update $\mathbf{D}_x$ and $\mathbf{D}_y$. By fixing $\mathbf{P}_x$, $\mathbf{P}_y$, $\mathbf{W}$, A and B, the objective function regarding to $\mathbf{D}_x$ can be written as follows:

$$\min_{\mathbf{D}_x} \|\Phi_x(\mathbf{X}) - \Phi_x(\mathbf{X})\mathbf{D}_x A\|_F^2,\ s.t.\ \|d_{x,i}\|_2^2 \le 1, \quad \forall i. \tag{14}$$

By introducing a relaxation variable Q, Eq. (14) can be rewritten as:

$$\min_{\mathbf{D}_x,Q} \|\Phi_x(\mathbf{X}) - \Phi_x(\mathbf{X})Q\|_F^2 + \tau_1\|Q - \mathbf{D}_x A\|_F^2, \quad s.t.\ \|d_{x,i}\|_2^2 \le 1, \quad \forall i, \tag{15}$$

where τ_1 is a scalar constant. Eq. (15) can be solved by updating $\mathbf{D}_x$ and Q alternatively. By keeping only the terms relevant to Q, we can obtain $\min_Q \|\Phi_x(\mathbf{X}) - \Phi_x(\mathbf{X})Q)\|_F^2 + \tau_1\|Q - \mathbf{D}_x A\|_F^2$. By setting the derivative to zero, the solution of Q can be obtained as:

$$Q = (K_x(\mathbf{X},\mathbf{X}) + \tau_1 I)^{-1}(K_x(\mathbf{X},\mathbf{X}) + \tau_1 \mathbf{D}_x A). \tag{16}$$

By ignoring irrelevant terms with respect to $\mathbf{D}_x$, Eq. (15) reduces to the following form:

$$\min_{\mathbf{D}_x} \tau_1\|Q - \mathbf{D}_x A\|_F^2,\ s.t.\ \|d_{x,i}\|_2^2 \le 1, \quad \forall i. \tag{17}$$

$\mathbf{D}_x$ can be updated by solving the problem in Eq. 17. Here, we use the similar way as [47] to solve Eq. 17, i.e., introducing a variable S:

$$\min_{\mathbf{D}_x} \tau_1\|Q - \mathbf{D}_x A\|_F^2,\ \ s.t.\ \mathbf{D}_x = S,\ \|s_i\|_2^2 \le 1 \quad \forall i. \tag{18}$$

The problem in (18) can be solved by using the ADMM [50] algorithm:

$$\begin{cases} \mathbf{D}_x = \arg\min_{\mathbf{D}_x} \tau_1\|Q - \mathbf{D}_x A\|_F^2 + \rho\|\mathbf{D}_x - S + T\|_F^2 \\ S = \arg\min_S \rho\,\|\mathbf{D}_x - S + T\|_F^2\,,\ s.t.\ \|s_i\|_2^2 \le 1 \\ T = T + \mathbf{D}_x - S,\ \ \textit{update } \rho \textit{ if appropriate}. \end{cases} \tag{19}$$

where the initial value of T is a zero matrix. $\mathbf{D}_y$ can be updated in the similar way as $\mathbf{D}_x$.

Step 3: Update $\mathbf{P}_x$ and $\mathbf{P}_y$. By removing irrelevant terms with respect to $\mathbf{P}_x$, the objective function can be written as follows:

$$\begin{aligned} \min_{\mathbf{P}_x} \alpha(&\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y)\|_F^2 \\ &- \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y)\|_F^2) \\ &+ \tau\|\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}) - A\|_F^2. \end{aligned} \tag{20}$$

By introducing two relaxation variables V_1 and V_2, Eq. (20) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{P}_x,V_1,V_2} \alpha(&\|\mathbf{W}^T(V_1 - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y)\|_F^2 \\ &- \|\mathbf{W}^T(V_2 - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y)\|_F^2) \\ &+ \tau\|\mathbf{P}_x\Phi_x(\mathbf{X})^T\Phi_x(\mathbf{X}) - A\|_F^2 \\ &+ \tau_2(\|V_1 - \mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x\|_F^2 \\ &+ \|V_2 - \mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x\|_F^2), \end{aligned} \tag{21}$$

where τ_2 is a scalar constant. By setting the derivative w.r.t. each variable to zero, $\mathbf{P}_x$, V_1 and V_2 can be updated alternatively as follows:

$$\begin{aligned} \mathbf{P}_x =\ & (\tau K_x(\mathbf{X},\mathbf{X})K_x(\mathbf{X},\mathbf{X})^T + \tau_2\Phi_x(\mathbf{X})^T M_S^x M_S^{xT}\Phi_x(\mathbf{X}) \\ & + \tau_2\Phi_x(\mathbf{X})^T M_G^x M_G^{xT}\Phi_x(\mathbf{X}))^{-1}(\tau A K_x(\mathbf{X},\mathbf{X})^T \\ & + \tau_2 V_1 M_S^{xT}\Phi_x(\mathbf{X}) + \tau_2 V_2 M_G^{xT}\Phi_x(\mathbf{X})) \end{aligned} \tag{22}$$

$$\begin{aligned} V_1 =\ & (\alpha\mathbf{W}\mathbf{W}^T + \tau_2 I)^{-1}(\alpha\mathbf{W}\mathbf{W}^T\mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y \\ & + \tau_2\mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x) \end{aligned} \tag{23}$$

$$\begin{aligned} V_2 =\ & (\alpha\mathbf{W}\mathbf{W}^T + \tau_2 I)^{-1}(\alpha\mathbf{W}\mathbf{W}^T\mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y \\ & + \tau_2\mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x) \end{aligned} \tag{24}$$

We update $\mathbf{P}_y$ in the similar way.

Step 4: Update W. By keeping only the terms relevant to $\mathbf{W}$, we can obtain:

$$\begin{aligned} \min_{\mathbf{W}}\ & \alpha(\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y)\|_F^2 \\ & - \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y)\|_F^2) \\ s.t.\ & \|w_i\|_2^2 \le 1, \quad \forall i. \end{aligned} \tag{25}$$

By introducing a variable S, problem (25) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{W}}\ & \alpha(\|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y)\|_F^2 \\ & - \|\mathbf{W}^T(\mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y)\|_F^2) \\ s.t.\ & \mathbf{W} = S,\ \|s_i\|_2^2 \le 1, \quad \forall i. \end{aligned} \tag{26}$$

Then, the ADMM algorithm can be employed to effectively solve this problem.

$$\begin{cases} \mathbf{W} = \arg\min_{\mathbf{W}} \alpha(\|\mathbf{W}^T M_1\|_F^2 - \|\mathbf{W}^T M_2\|_F^2) + \rho\|\mathbf{W} \\ \quad - S + T\|_F^2 \\ S = \arg\min_S \rho\,\|\mathbf{W} - S + T\|_F^2\,,\ s.t.\ \|s_i\|_2^2 \le 1 \\ T = T + \mathbf{W} - S,\ \ \textit{update } \rho \textit{ if appropriate}, \end{cases} \tag{27}$$

where $M_1 = \mathbf{P}_x\Phi_x(\mathbf{X})^T M_S^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_S^y$, $M_2 = \mathbf{P}_x\Phi_x(\mathbf{X})^T M_G^x - \mathbf{P}_y\Phi_y(\mathbf{Y})^T M_G^y$.

We repeat the above procedure until convergence. Algorithm 1 summarizes the optimization process of our approach.

Algorithm 1 The Proposed Kernel Analysis-Synthesis Dictionary Based Heterogeneous Distance Learning Approach

Require: Training image and video sets $\mathbf{X}$ and $\mathbf{Y}$
Ensure: $\mathbf{D}_x, \mathbf{D}_y, \mathbf{P}_x, \mathbf{P}_y, \mathbf{W}$
1: Initialize $\mathbf{D}_x, \mathbf{D}_y, \mathbf{P}_x, \mathbf{P}_y, \mathbf{W}, \alpha, \lambda$, and τ
2: **while** not converge **do**
3: Update $\mathbf{A}$ and $\mathbf{B}$ by (9) and (10), respectively;
4: Update $\mathbf{D}_x$ and $\mathbf{D}_y$ according to (15);
5: Update $\mathbf{P}_x$ and $\mathbf{P}_y$ according to (20);
6: Update $\mathbf{W}$ according to (25);
7: **end while**
8: **return** $\mathbf{D}_x, \mathbf{D}_y, \mathbf{P}_x, \mathbf{P}_y$ and $\mathbf{W}$;

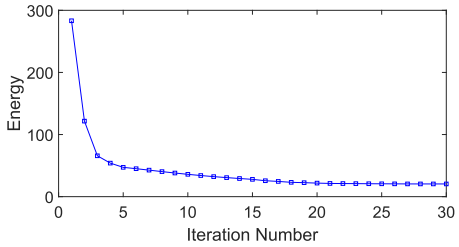


FIGURE 3. Convergence curve of the proposed KADDL approach on the iLIDS-VID dataset.

C. MATCHING BETWEEN IMAGE AND VIDEO FEATURES USING KADDL

In the testing phase, the matching between image and video features can be realized by using the learned kernel analysis-synthesis dictionary pairs $\{\mathbf{D}_x, \mathbf{P}_x\}$ and $\{\mathbf{D}_y, \mathbf{P}_y\}$) as well as the metric matrix $\mathbf{W}$. Let $x_i \in \mathbb{R}^p$ be the feature vector of a testing image, and $y_j \in \mathbb{R}^q$ be the feature vector of a video. Detailed steps of matching between x_i and y_j are as follows.

(1) Transform the heterogeneous image and video features into coding coefficients of the same dimension by using the learned dictionaries. Specifically, the coefficient of x_i is obtained by using $a_i = \mathbf{P}_x \Phi_x(\mathbf{X})^T \Phi_x(x_i)$, and the coefficient of y_j is obtained by using $b_j = \mathbf{P}_y \Phi_y(\mathbf{Y})^T \Phi_y(y_j)$.

(2) Calculate the distance between the coefficients of image and video by using the learned distance metric $\mathbf{W}$. The distance between a_i and b_j can be calculated as $d(a_i, b_j) = \|\mathbf{W}^T(a_i - b_j)\|_2^2$.

D. COMPLEXITY AND CONVERGENCE

In the optimization process of our KADDL approach, all variables are updated iteratively. In each iteration, the time complexity of updating $\mathbf{A}$ or $\mathbf{B}$ is $O(m^3)$, where m is the dictionary size; updating $\mathbf{D}_x$ and $\mathbf{D}_y$ costs $Ok(m^3)$, where k is the iteration number in the ADMM algorithm; the time complexities for updating $\mathbf{P}_x$ and $\mathbf{P}_y$ are $Ok(N_1^3)$ and $Ok(N_2^3)$, respectively; updating $\mathbf{W}$ costs $Ok(m^3)$. In our experiments, k is usually smaller than 10, and the dictionary size m is usually much smaller than the sample numbers of $\mathbf{X}$ and $\mathbf{Y}$ (i.e., N_1 and N_2). Therefore, the major computational burden in the training phase of KADDL is on updating $\mathbf{P}_x$ and $\mathbf{P}_y$. Fortunately, the operations that cost $O(N_1^3)$ and $O(N_2^3)$ will not change in iteration process, and thus can be pre-computed. This greatly accelerates the training process.

The proposed optimization algorithm for KADDL is an alternate iterative optimization algorithm. In each iteration, $\{\mathbf{A}, \mathbf{B}\}$, $\{\mathbf{D}_x, \mathbf{D}_y\}$, $\{\mathbf{P}_x, \mathbf{P}_y\}$ and $\mathbf{W}$ are updated alternatively, and each sub-problem is convex. The convergence of such a problem has already been intensively studied in [51]. Figure 3 shows the convergence curve of our algorithm on the iLIDS-VID dataset. We can see that the energy drops quickly and begins to stabilize after 20 iterations. In most of our experiments, our algorithm will converge in less than 30 iterations.

IV. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed approach, we conduct extensive experiments on three publicly available person sequence datasets, including iLIDS-VID [52], PRID 2011 [53] and MARS [54].

A. EXPERIMENTAL SETTINGS

1) FEATURE EXTRACTION

For the image set, we employ Part-based Convolutional Baseline (PCB) method [55] to extract features from each image, and use the obtained feature vector to represent the image. For the video sequence, two kinds of features are employed. Specifically, we first extract HOG3D feature from each sequence by using the same way as [54]. Then, we extract PCB feature from each frame of the sequence, and perform max-pooling operation on the obtained feature vectors to generate a feature vector for the sequence. Finally, the extracted HOG3D and PCB features are concatenated to represent the video sequence.

2) EVALUATION SETTING

In this article, we conduct experiments by following the similar evaluation protocol as [15]. For the experiments on iLIDS-VID and PRID 2011, all persons are randomly split into two sets of equal size, with one for training and the other for testing. Then, we select the first image from each sequence of the probe camera to construct the image set, and the sequences from the gallery camera are used as the video set. For the MARS dataset, the presetting of training/test split is used. For each person in the training set, we select the first frame from half of his/her video sequences to form the image set, and the remaining sequences are used to construct the video set. For the test set, the first frame of each query video sequence is used as the probe image. For the experiment on each dataset, we repeat each experiment 10 times and report the average rank-r matching rates.

3) PARAMETER SETTING

There are three parameters in our KADDL approach, including α, λ, and τ. In experiments, we set their values by using the cross validation technique on the training data. In addition, the size of image and video dictionaries is set as 240, 220 and 300 for iLIDS-VID, PRID 2011 and MARS, respectively; the number of columns in metric matrix $\mathbf{W}$ is set as 180, 170 and 210 for iLIDS-VID, PRID 2011 and MARS, respectively. The kernel function $k(x, y) = exp(-\|x - y\|^2/s)$ is used for image and video features, and the kernel parameter is set as the mean of the pairwise distances of samples [48].

B. RESULTS AND ANALYSIS

To demonstrate the effectiveness of our approach, we compare KADDL with several state-of-the-art methods on the iLIDS-VID, PRID 2011 and MARS datasets, and report the detailed rank 1, 5, 10, 20 matching rates of all methods. **For our approach, the results are obtained in the case that only**

TABLE 2. Top *r* ranked matching rates (%) on the iLIDS-VID dataset. The results of KADDL are obtained by using only half of the labeled training samples.

Method	*r*=1	*r*=5	*r*=10	*r*=20
PHDL [11]	28.2	50.4	65.9	80.4
MPHDL [11]	32.6	55.8	69.3	83.2
RCME [18]	40.1	67.2	79.7	86.7
TMSL [16]	39.5	66.9	79.6	86.6
P2SNet [15]	40.0	68.5	78.1	90.0
TKP [12]	54.6	79.4	86.9	93.5
KADDL	56.3	81.9	88.7	95.1

TABLE 3. Top *r* ranked matching rates (%) on the PRID 2011 dataset. The results of KADDL are obtained by using only half of the labeled training samples.

Method	*r*=1	*r*=5	*r*=10	*r*=20
PHDL [11]	41.9	67.3	85.5	92.4
MPHDL [11]	46.3	72.6	87.4	93.3
SRLLGC [17]	52.3	72.1	82.6	85.2
RCME [18]	67.5	83.6	93.6	97.6
TMSL [16]	68.5	84.7	93.1	97.3
P2SNet [15]	73.3	90.5	94.7	97.8
KADDL	80.4	95.1	97.2	98.8

TABLE 4. Top *r* ranked matching rates (%) on the MARS dataset. The results of KADDL are obtained by using only half of the labeled training samples.

Method	*r*=1	*r*=5	*r*=10	*r*=20
PHDL [11]	35.7	51.5	60.9	67.3
MPHDL [11]	39.1	53.8	63.3	69.0
SRLLGC [17]	62.1	75.3	-	83.7
TMSL [16]	56.5	70.6	-	83.5
P2SNet [15]	55.3	72.9	78.7	83.7
TKP [12]	75.6	87.6	90.9	-
KADDL	74.3	87.8	91.5	93.2

half of the training samples are labeled. For the competing methods, the reported results are from their own papers.

Tables 2 - 4 report the matching rates of all methods on the iLIDS-VID, PRID 2011 and MARS datasets. We can observe that: (1) our KADDL approach achieves the best matching results on the iLIDS-VID dataset. In particular, our approach improves the rank 1 matching rate at least by 1.7%(=56.3%-54.6%). (2) Our approach achieves much higher matching rates than other methods on PRID 2011. Taking the rank 1 matching rate as an example, KADDL improves the average matching rate at least by 7.1% (=80.4%-73.3%). (3) Our approach achieves very competitive results on MARS, and even outperforms the best compared method TKP from rank 5. Note that the results of our approach are obtained in the case that only half of the training data is labeled.

The major reasons why our approach can achieve very competitive or even better results with fewer labeled data are three-fold: (1) Considering that the image and video data lies in non-linear feature space, our approach learns coupled dictionaries in the Reproducing Kernel Hilbert Space, such that the learned dictionaries can better characterize the intrinsic relationship between image and video data, and reduce the gap between image and video. (2) Our approach learns discriminative distance metric over the coding coefficients of the images and videos, which can improve the performance of the learned re-identification model. (3) KADDL can exploit the discriminant information contained in the unlabeled data by using the designed reliability-based semi-supervised strategy, which further improves the discriminability of our approach.

TABLE 5. Top *r* ranked matching rates (%) of KADDL and KADDL_base on three datasets. The results are obtained by using half of the labeled training samples.

iLIDS-VID	*r*=1	*r*=5	*r*=10	*r*=20
KADDL_base	51.8	78.5	86.1	93.7
KADDL	56.3	81.9	88.7	95.1
PRID 2011	*r*=1	*r*=5	*r*=10	*r*=20
KADDL_base	76.2	92.4	95.7	98.3
KADDL	80.4	95.1	97.2	98.8
MARS	*r*=1	*r*=5	*r*=10	*r*=20
KADDL_base	68.5	83.7	89.2	91.1
KADDL	74.3	87.8	91.5	93.2

V. IN-DEPTH DISCUSSION AND ANALYSIS

A. EFFECT OF THE DESIGNED SEMI-SUPERVISED STRATEGY

The proposed KADDL approach utilizes the designed reliability-based semi-supervised strategy to exploit the useful information contained in unlabeled data. In this experiment, we evaluate the effectiveness of the designed semi-supervised strategy. To this end, we compare reliability-based semi-supervised strategy with the baseline semi-supervised strategy (i.e., regarding the newly added pseudo-label samples as the truly labeled samples, and making use of them with the same manner). Here, we call the modified version of KADDL with the baseline semi-supervised strategy as KADDL_base. Table 5 reports the rank 1-20 matching rates of KADDL and KADDL_base on the iLIDS-VID, PRID 2011 and MARS datasets. We can observe that the rank 1 matching rate of our approach is improved at least by 4.2% (80.4%-76.2%) by using the designed reliability-based semi-supervised strategy. These results indicate that the designed semi-supervised strategy is beneficial to making better use of the useful information contained in unlabeled data.

B. EFFECT OF USING NON-LINEAR TECHNIQUE

In this experiment, we evaluate the effect of using non-linear technique. Specifically, we compare our approach with the modified version that directly learns linear synthesis-analysis dictionary pair from the original feature data, and report the performance improvement from the linear feature space to non-linear feature space. We call the linear version of

TABLE 6. Rank 1 matching rates (%) of KADDL and KADDL_*linear* on three datasets.

Method	iLIDS-VID	PRID 2011	MARS
KADDL_*linear*	51.6	75.6	69.0
KADDL	56.3	80.4	74.3

TABLE 7. Rank 1 matching rates (%) of KADDL and KADDL-W on three datasets.

Method	iLIDS-VID	PRID 2011	MARS
KADDL-W	54.1	78.2	70.5
KADDL	56.3	80.4	74.3

our KADDL approach as KADDL_*linear*, which learns the dictionaries without using kernel technique. Table 6 shows the rank 1 matching rates of KADDL and KADDL_*linear* on the iLIDS-VID, PRID 2011 and MARS datasets. One can easily observe that our approach achieves better performance by using non-linear technique (kernel technique). The major reason is that samples of the image/video set usually lie in non-linear feature space, and learning dictionaries in RKHS space can better capture the intrinsic distribution of image/video data.

C. EFFECT OF THE CODING COEFFICIENT BASED DISTANCE LEARNING TERM

To evaluate the effect of the designed coding coefficient based distance learning term, we generate the modified version of KADDL by removing $\mathbf{W}$ from Eq. (5), and compare the performance of KADDL and its modified version. Here, we name the modified version of KADDL without using $\mathbf{W}$ as KADDL-W. Table 7 reports the rank 1 matching rates of KADDL and KADDL-W on three datasets. We can see that the performance of KADDL decreases without using $\mathbf{W}$, which means that learning the metric matrix $\mathbf{W}$ is beneficial to further enhancing the discriminability of the obtained re-identification model.

D. EFFECT OF THE QUANTITY OF LABELED SAMPLES

In our KADDL approach, the distance learning process (especially for the first epoch) depends on the labeled data and pseudo-label data. Thus, the quantity of labeled data will have an impact on the discriminability of the learned model. To investigate the influence of the quantity of labeled data to the performance of KADDL, we perform KADDL by changing the percentage of labeled data in the training set from 5% to 100%, and observe the performance changes. Figure 4 plots the rank 1 matching rates of our approach versus different percentages of labeled data on the iLIDS-VID dataset. We can observe that: (1) the rank 1 matching rate of our approach grows along with the increase of labeled training data, and reaches over 65% when all the images and videos in the training set are labeled; (2) When only 10% or even 5% training samples are labeled, the rank 1 performance of our approach is still higher than 40%. The results indicate that our approach can leverage the information contained in unlabeled data to learn more effective re-identification model. Similar effects can be observed on the other datasets.

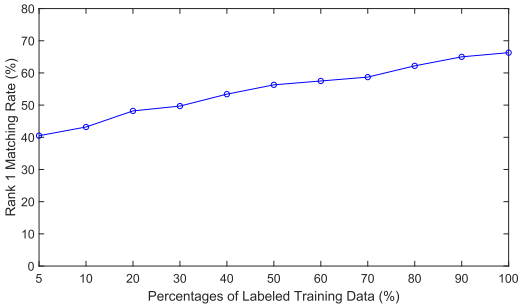


FIGURE 4. Rank 1 matching rates of KADDL versus different percentages of labeled data in the training data on the iLIDS-VID dataset.

TABLE 8. Rank 1 matching rates (%) of KADDL by using different kernel functions (including Linear kernel, Polynomial kernel and Gaussian kernel) on three datasets.

Method	iLIDS-VID	PRID 2011	MARS
KADDL_L	53.9	78.5	73.0
KADDL_P	55.7	79.1	74.6
KADDL	56.3	80.4	74.3

E. EFFECT OF DIFFERENT KERNEL FUNCTIONS

In the proposed KADDL approach, Gaussian kernel is utilized as the default kernel function. In this experiment, we will evaluate the influence of different kernel functions to the performance of our approach, including Linear kernel, Polynomial kernel and Gaussian kernel. We name the modified version of our approach using Linear kernel as KADDL_L, and name the version using Polynomial kernel as KADDL_P. Table 8 reports the rank 1 matching rates of KADDL, KADDL_L and KADDL_P on the iLIDS-VID, PRID 2011 and MARS datasets. We can observe that: for the iLIDS-VID and PRID 2011 datasets, using Gaussian kernel (i.e., KADDL) can bring better performance; For the MARS dataset, using Polynomial kernel is a better choice. The experimental results indicate that selecting more appropriate kernel functions can induce better performance for our approach. In practice, we can realize this by using multi-kernel technique.

F. COMPUTATION TIME

In this experiment, we investigate the training and testing time of our KADDL approach. We run KADDL on a computer with an Intel I9 eight-core 3.6GHZ CPU and 32GB memory. In the training phase, the computation time of learning dictionaries and distance metric on the iLIDS-VID, PRID 2011 and MARS datasets is 8.6, 4.2 and 15.3 minutes, respectively. In the testing phase, the testing time for one query image is

less than 0.1 seconds. In practice, the training phase is usually off-line, thus our approach is suitable for practical use.

VI. CONCLUSION

In this article, we investigate the problem of image to video person re-id under the semi-supervised setting, and propose a Kernel Analysis-synthesis Dictionary based heterogeneous Distance Learning (KADDL) approach. By learning coupled kernel analysis-synthesis dictionaries, our approach can transform the heterogeneous image and video features into coding coefficients of the same representation space, such that the coupled inconsistency can be reduced. By designing the reliability-based semi-supervised strategy, our approach can exploit the discriminant information contained in unlabeled data effectively. By learning distance metric over the coding coefficients in the representation space, the discriminability of our approach can be further improved.

Experimental results on three widely used person sequence datasets demonstrate that: (1) The proposed KADDL approach can achieve competitive or even better results than the compared methods by using only half of the labeled data. (2) By introducing kernel technique into coupled analysis-synthesis dictionary learning, the learned dictionaries can better reveal the intrinsic relationship between two non-linear distributions. (3) The designed reliability-based semi-supervised strategy is beneficial to making better use of the information contained in unlabeled data.

REFERENCES

[1] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[2] J. García, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni, "Discriminant context information analysis for post-ranking person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1650–1665, Apr. 2017.

[3] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7202–7211.

[4] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3633–3642.

[5] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 591–606, Mar. 2016.

[6] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *Proc. IJCAI*, 2016, pp. 3552–3559.

[7] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Trans. Image Process.*, vol. 29, pp. 5481–5490, 2020.

[8] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 579–590, 2020.

[9] Y. Tang, X. Yang, N. Wang, B. Song, and X. Gao, "CGAN-TM: A novel domain-to-domain transferring method for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 5641–5651, 2020.

[10] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2013–2025, 2020.

[11] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 717–732, Mar. 2018.

[12] X. Gu, B. Ma, H. Chang, S. Shan, and X. Chen, "Temporal knowledge propagation for Image-to-Video person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9646–9655.

[13] B. Yu, N. Xu, and J. Zhou, "Cross-media body-part attention network for Image-to-Video person re-identification," *IEEE Access*, vol. 7, pp. 94966–94976, 2019.

[14] X. Zhu, X. Jing, F. Wu, Y. Wang, W. Zuo, and W. Zheng, "Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image," in *Proc. AAAI*, 2017, pp. 4341–4348.

[15] G. Wang, J. Lai, and X. Xie, "P2SNet: Can an image match a video for person re-identification in an end-to-end way?" *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2777–2787, Oct. 2018.

[16] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai, "Image-to-video person re-identification with temporally memorized similarity learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2622–2632, Oct. 2018.

[17] T. Li, L. Sun, C. Han, and J. Guo, "Salient region-based least-squares log-density gradient clustering for image-to-video person re-identification," *IEEE Access*, vol. 6, pp. 8638–8648, 2018.

[18] Z. Xie, L. Li, X. Zhong, and L. Zhong, "Image-to-video person re-identification by reusing cross-modal embeddings," 2018, *arXiv:1810.03989*. [Online]. Available: https://arxiv.org/abs/1810.03989

[19] X. Zhang, S. Li, X.-Y. Jing, F. Ma, and C. Zhu, "Unsupervised domain adaption for image-to-video person re-identication," *Multimed Tools Appl.*, 2020, doi: 10.1007/s11042-019-08550-9.

[20] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5683–5695, Nov. 2018.

[21] X. Zhu, X.-Y. Jing, F. Zhang, X. Zhang, X. You, and X. Cui, "Distance learning by mining hard and easy negative samples for person re-identification," *Pattern Recognit.*, vol. 95, pp. 211–222, Nov. 2019.

[22] B. Nguyen and B. De Baets, "Kernel distance metric learning using pairwise constraints for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 589–600, Feb. 2019.

[23] H. Yan and J. Hu, "Video-based kinship verification using distance metric learning," *Pattern Recognit.*, vol. 75, pp. 15–24, Mar. 2018.

[24] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[25] K. Li, Z. Ding, S. Li, and Y. Fu, "Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification," in *Proc. AAAI*, 2018, pp. 2331–2338.

[26] P. Song, L. Weizman, J. F. C. Mota, Y. C. Eldar, and M. R. D. Rodrigues, "Coupled dictionary learning for multi-contrast MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 621–633, Mar. 2020.

[27] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.

[28] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 498–505.

[29] D. Tao, Y. Guo, Y. Li, and X. Gao, "Tensor rank preserving discriminant analysis for facial recognition," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 325–334, Jan. 2018.

[30] Z. Ding and Y. Fu, "Robust transfer metric learning for image classification," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 660–670, Feb. 2017.

[31] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.

[32] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.

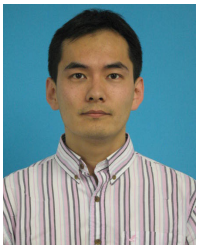
[33] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 203–208.
[34] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing KISS metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.
[35] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn. - ICML*, 2007, pp. 209–216.
[36] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[37] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. ECCV*, vol. 7577, 2012, pp. 780–793.
[38] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.
[39] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[40] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[41] L. Sun, Y. Zhou, Z. Jiang, and A. Men, "Coupled analysis-synthesis dictionary learning for person re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 365–369.
[42] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen, "Online data organizer: Micro-video categorization by structure-guided multimodal dictionary learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1235–1247, Mar. 2019.
[43] V. Singhal and A. Majumdar, "A domain adaptation approach to solve inverse problems in imaging via coupled deep dictionary learning," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107163.
[44] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3550–3557.
[45] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proc. IJCAI*, 2015, pp. 2155–2161.
[46] X.-Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, and J.-Y. Yang, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, Mar. 2017.
[47] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in *Proc. NIPS*, vol. 2014, pp. 793–801.
[48] X. Zhu, X. Jing, F. Wu, D. Wu, L. Cheng, S. Li, and R. Hu, "Multi-kernel low-rank dictionary pair learning for multiple features based image classification," in *Proc. AAAI*, 2017, pp. 2970–2976.
[49] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.
[50] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[51] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, Nov. 2007.
[52] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. ECCV*, vol. 2014, pp. 688–703.
[53] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Berlin, Germany: Springer-Verlag, 2011, pp. 91–102.
[54] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, 2016, pp. 868–884.
[55] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)," in *Proc. ECCV*, 2018, pp. 501–518.

XIAOKE ZHU (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence system from Wuhan University, Wuhan, China, in 2017. He is currently an Associate Professor with the School of Computer and Information Engineering, Henan University, China. He has published more than 30 scientific articles, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), the IEEE TRANSACTIONS ON SERVICES COMPUTING (TSC), PR, CVPR, AAAI, and IJCAI. His current research interests include person re-identification, kinship verification, and image classification.

PENGFEI YE received the B.S. degree from Henan University, China, in 2019, where he is currently pursuing the M.S. degree with the School of Computer and Information Engineering. His research interests include dictionary learning and deep learning.

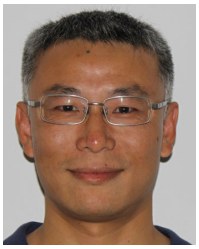
XIAO-YUAN JING (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, in 1998. He was a Professor with the Department of Computer Science, Shenzhen Research Student School, Harbin Institute of Technology, in 2005. He is currently a Professor with the School of Computer Science, Wuhan University, China. He has published more than 100 scientific articles, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), *Trends in Cell Biology* (TCB), the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), CVPR, AAAI, IJCAI, WWW, and PR.

XINYU ZHANG received the M.S. degree in computer application technology from the China University of Mining and Technology, Xuzhou, China, in 2013. He is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.



XIANG CUI received the Ph.D. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2012. He is currently an Associate Professor with the School of Computer and Information Engineering, Henan University, China. His current research interests include image processing and machine learning.

XIAOPAN CHEN received the Ph.D. degree from Henan University, Kaifeng, China, in 2015. He is currently an Associate Professor with the School of Computer and Information Engineering, Henan University. His research interests include face recognition, kinship verification, and deep learning.



FAN ZHANG received the Ph.D. degree in computer science from the Beijing University of Technology, China, in 2005. He is currently a Professor with the School of Computer and Information Engineering, Henan University, China. He has published more than 50 scientific articles. His research interests include pattern recognition and image processing.