# Robust Depth-based Person Re-identification

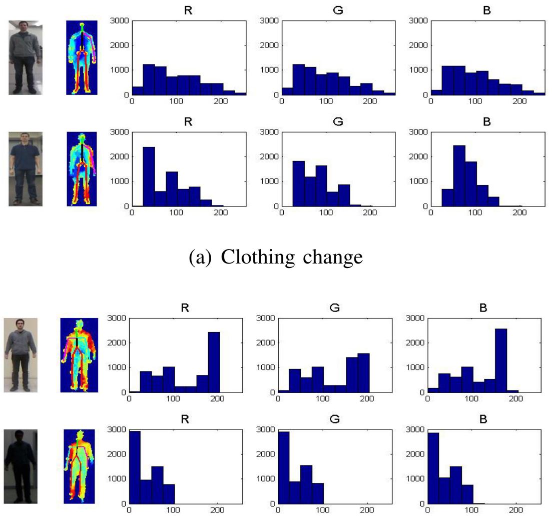
Ancong Wu, Wei-Shi Zheng, Jianhuang Lai

***Abstract*—Person re-identification (re-id) aims to match people across non-overlapping camera views. So far the RGB-based appearance is widely used in most existing works. However, when people appeared in extreme illumination or changed clothes, the RGB appearance-based re-id methods tended to fail. To overcome this problem, we propose to exploit depth information to provide more invariant body shape and skeleton information regardless of illumination and color change. More specifically, we exploit depth voxel covariance descriptor and further propose a locally rotation invariant depth shape descriptor called Eigen-depth feature to describe pedestrian body shape. We prove that the distance between any two covariance matrices on the Riemannian manifold is equivalent to the Euclidean distance between the corresponding Eigen-depth features. Furthermore, we propose a kernelized implicit feature transfer scheme to estimate Eigen-depth feature implicitly from RGB image when depth information is not available. We find that combining the estimated depth features with RGB-based appearance features can sometimes help to better reduce visual ambiguities of appearance features caused by illumination and similar clothes. The effectiveness of our models was validated on publicly available depth pedestrian datasets as compared to related methods for person re-identification.**

***Index Terms*—person re-identification, depth information.**

## I. Introduction

The task of person re-identification (re-id) is to match people in a distributed multi-camera surveillance system at different time and locations, with wide applications to forensic search, multi-camera tracking and access control, etc. In most short-term applications, low-level features such as color and textures are important appearance cues used to match. It is apparent that lighting will significantly affect the performance of these low-level features. In more extreme cases, when lighting condition changes greatly (e.g., with v.s. without lighting), color information of clothes becomes unreachable. Moreover, when people change clothes, color and textures become unreliable. For example, Figure 1 shows how color histograms change when people change clothes or appear in extreme illumination. In these cases, most existing re-id systems are not workable, since they are RGB-based.



(a) Clothing change

(b) Illumination change

Fig. 1. Illustration of change of color histograms and invariance of depth and skeletons. From left to right, the first column shows RGB images, the second column shows depth images (shown by pseudo-color images) and skeletons, and the remaining columns show histograms of R, G, B channels, respectively.

In comparison to RGB information, depth information can maintain more invariant even when suffering from clothing change and extreme illumination. As shown in Figure 1, shape and skeleton of body are likely more invariant under extreme lighting and clothing change. Nowadays, extracting depth and skeleton information with depth cameras (e.g., Microsoft Kinect) is not difficult in an indoor environment. Kinect sensor obtains depth value (distance to the camera) of each pixel by infrared, regardless of object color and illumination in indoor applications. With depth information, the life-size point cloud and skeleton of a person can be extracted, providing shape and physical information of his/her body. Moreover, with depth value of each pixel, pedestrians can be more easily segmented from background, so that background influence can be largely eliminated. Hence, using depth information could overcome some difficulties in RGB appearance-based methods, such as color change, illumination change and background clutters.

Although there are some advantages for depth-based re-identification as compared to the RGB appearance-based methods, challenges and limitations also come along with depth information. Firstly, the depth images captured by depth device change significantly when a person's viewpoint changes. Secondly, noises from devices exist in the captured depth images. These two aspects will seriously affect the use of depth

This work was supported partially by NSFC (No.61522115, 61472456, 61573387, 61661130157, 61628212), Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, the GuangDong Program (No.2015B010105005), the Guangdong Science and Technology Planning Project (No.2016A010102012, 2014B010118003), and Guangdong Program for Support of Top-notch Young Professionals (No.2014TQ01X779). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wen Gao. (Corresponding author: Wei-Shi Zheng)

Ancong Wu is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China; and is also with the Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China. Email: wuancong@mail2.sysu.edu.cn.

Wei-Shi Zheng is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China; and is also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China. E-mail: wszheng@ieee.org.

Jianhuang Lai is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China; and is also with Guangdong Province Key Laboratory of Information Security, P. R. China. E-mail: stsljh@mail.sysu.edu.cn.

information for person re-identification. So far, a few methods [1]–[4] have been developed to exploit depth information for person re-identification, but the above two problems are still not well solved in existing methods. [1] uses only skeletons to extract feature. In [2], [3], besides using skeleton to extract physical information, applying point clouds converted from depth images for 3D body shape matching is also considered, but alignment errors and noises of point clouds are the problems remained unsolved. In [4], a deep model is applied to classify the person point cloud sequences, in which feature extraction and classification are jointly modeled and body shape is not explicitly described. Therefore, body shape description is still an important biometric cue which needs further study for person re-identification.

In this work, we aim to design a depth shape descriptor which is locally invariant to rotation[1] and insensitive to noise. We propose two depth shape descriptors: depth voxel covariance descriptor and Eigen-depth feature. Eigen-depth feature is based on depth voxel covariance descriptor and locally rotation invariant. Then we combine depth shape descriptor with skeleton-based feature to form complete depth representation of body shape and physical information. The pipeline of constructing a descriptor for our depth-based person re-id method is illustrated in Figure 2. Our method takes the following steps: (1) segmentation and computation of point cloud and normals of torso and head; (2) extracting depth voxel covariance descriptor and locally rotation invariant Eigen-depth feature; (3) enriching body depth shape descriptor by additionally combining skeleton-based feature. In the second step, the Eigen-depth feature is more suitable due to its stability against local rotation of body when the viewpoint of a person varies obviously, while the depth voxel covariance descriptor will be more effective because of rich information it contains when the viewpoint change of a person is slight. In the third step, the skeleton-based feature can be complementary to the depth shape descriptor extracted from step (2), so more robust matching can be achieved.

In addition, in real-world applications, most of the deployed cameras in existing surveillance systems cannot capture depth information, so how to make depth-based method work in existing system is also a challenge, while existing depth-based and RGB-D-based methods assume depth information is available. Towards overcoming this limitation, we learn the relation between depth features and RGB-based appearance features by a kernelized implicit feature transfer scheme. For this purpose, an auxiliary RGB-D dataset is employed to learn the nonlinear transformation between RGB-based appearance feature and depth feature. When depth device is not ready/available, the depth feature is estimated from RGB image and used to augment the RGB-based appearance feature. The experiment results show that this makes extra improvement on the re-identification performance for top-ranked matching.

We tested our methods on three publicly available datasets, PAVIS [1], BIWI RGBD-ID [2] and IAS-Lab RGBD-ID [3]. The results show the effectiveness of our depth-based approach

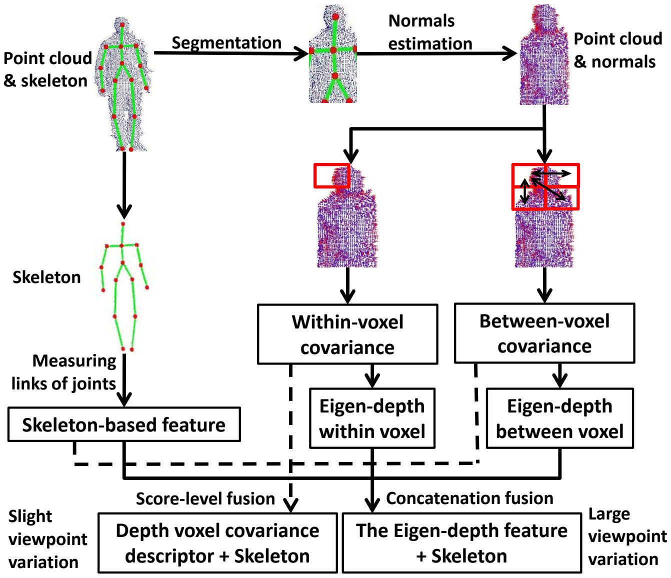


Fig. 2. Pipeline of our depth-based re-identification feature extraction. In the last step, depth shape descriptor and skeleton-based feature are combined. When the viewpoint variation of pedestrians is slight, depth voxel covariance descriptor (DVCov) is exploited as depth shape descriptor; when the viewpoint variation is large, Eigen-depth (ED) is exploited.

for overcoming change of clothes and extreme illumination condition. When clothes are completely different between gallery and probe, RGB appearance-based methods fail while our depth-based method is effective. Our approach outperforms other existing depth-based re-identification methods including skeleton-based methods, PCM, combination of them [3] and recurrent attention model [4]. Compared to other favorable rotation invariant depth shape descriptors, our descriptor also outperforms RIFT2M [5] and Fehr's descriptor [6].

In summary, the contributions of our work are: (1) proposing depth voxel covariance descriptor and Eigen-depth feature for depth-based re-identification and proving the local rotation invariance of Eigen-depth feature in theory; (2) forming a depth re-id recognition framework by unifying depth shape descriptor and skeleton-based feature for a complete representation; (3) proposing a kernelized implicit feature transfer scheme to estimate the Eigen-depth feature from RGB images implicitly when depth device is not available.

## II. Related Work

In this section, we present an overview of related image-based re-id technologies in three aspects: (1) RGB appearance-based re-id, (2) depth-based re-id, and (3) RGB-D re-id. Currently, most person re-identification approaches are based on 2D RGB-based appearance features.

### A. RGB Appearance-based Person Re-identification

Most existing works rely on RGB-based appearance features. Among them, color is most frequently used and it is often encoded into histograms [7]–[12]. Besides, texture-based features are also employed, including HOG-like signature [13], Gabor feature [11], [14], graph model [15], differential filters [11], [14] and Haar-like representations [16]. Many other

[1] Local rotation invariance means that the feature of a body part is invariant when viewpoint change of pedestrian will not make that body part become invisible due to self-occlusion.

hand-crafted features such as covariance descriptor [17], Fisher vector [18], spatial co-occurrence representation [19], custom pictorial structure [20] and SARC3D [21] were also developed for achieving more reliable representations. Recently, feature learning methods have been more focused on, such as salience learning [22], mirror representation [23], salient color names [24], reference descriptor [25], context-based feature [26], deep features [27]–[32], dictionary learning [33]–[35] and attribute learning [36]–[38]. However, in the situations of clothing change or extreme illumination, these RGB-based appearance features tend to fail.

Besides feature representation, a large amount of metric/subspace models [11], [12], [14], [39]–[56], have been developed to achieve more reliable matching, such as LMNN [39], RankSVM [14], RDC [41], PCCA [42], KISSME [40], LFDA [43], CVDCA [50], CRAFT [51], MLAPG [52], TDL [55] and DNS [56]. Some other methods have also been proposed for this purpose, e.g., re-ranking [57], [58] and correspondence structure [59]. Unsupervised learning models [60], [61] have also been developed for person re-identification. However, they cannot solve the illumination and clothing change problems. Compared to RGB-based appearance features, depth information is a solution to this problem, because it is independent of color and maintains more invariant for a longer period of time.

### *B. Depth-based Person Re-identification*

So far, only a few depth-based re-identification methods based on depth image, point cloud and anthropometric measurement [1]–[4], [62]–[64] have been developed. To some extent, depth-based methods can solve the problems of changing clothes and extreme illumination. Barbosa et al. exploited skeleton-based feature [1] based on anthropometric measurement of distances between joints and geodesic distances on body surface. Munaro et al. built a point cloud model for each person as gallery by fusing a set of point clouds from different views and then applied Point Cloud Matching (PCM) to compute the distance between samples [2]. In [3], [62], Munaro et al. combined PCM and skeleton-based feature modified based on Barbosa et al.'s work [1]. These methods needed to align the point clouds, and no depth shape descriptor was applied for describing body shape. Haque et al. proposed a recurrent attention model [4] for depth-video-based person identification, in which 3D RAM model was for still 3D point clouds and 4D RAM model was for 3D point cloud sequences. However, among the above depth-based frameworks, PCM and Haque's method were not suitable for solving person re-identification problem under the setting when there is no overlap on people between training and testing.

Compared to existing depth-based re-identification frameworks, the main difference of our work is that we propose depth voxel covariance descriptor and Eigen-depth feature to describe body shape. Eigen-depth feature is a covariance-based feature, and it is locally rotation invariant and does not require alignment of point clouds. The Eigen-depth feature can be viewed as a depth shape descriptor and thus can remove the ambiguity of using only anthropometric measurement of skeletons in the previous depth modeling for re-identification. Compared to direct utilization of point cloud in PCM [2], it deals with noises of non-rigid human body better.

We also discuss some related depth shape descriptors, including the covariance descriptor in [65], RIFT2M [5] and Fehr's covariance descriptor [6], which were not applied for person re-identification. Compared to the covariance descriptor in [65], Eigen-depth feature is locally rotation invariant. Compared to rotation invariant descriptors RIFT2M [5] and Fehr's covariance descriptor [6], Eigen-depth feature is densely extracted rather than using interest points, so that it contains richer information of body shape. Moreover, its rotation invariance is achieved in eigen-analysis level, so alignment of point cloud is not needed and more compact representation can be obtained by eigen-analysis.

### *C. RGB-D Person Re-identification*

Since RGB and depth information can be obtained simultaneously when using Kinect, some re-identification methods have been developed to combine depth information and RGB appearance cues in order to extract more discriminative feature representation. Pala et al. [66] improved accuracy of clothing appearance descriptors by fusing them with anthropometric measures extracted from depth data. Mogelmose et al. [67] presented a tri-modal method to combine RGB, depth and thermal features. Mogelmose et al. [68] combined color histogram and height feature extracted from depth information. John et al. [69] combined RGB-Height histogram and gait feature of depth information. Satta et al. [70] exploited skeleton to segment human body and extracted color feature. In [71], each color pixel was assigned to the nearest bone in the skeleton, and color histograms were computed for each region. In [72], the proposed feature bodyprint exploited the mean RGB values of regions in different heights. In [73], the descriptor was based on a 3D cylindrical grid that unified color variations together with angle and height. Takac et al. [74] exploited color histograms of upper body and lower body separately. Xu et al. [75] proposed a distance metric using RGB-D data to improve RGB-based person re-identification.

As reported in these works, the combination of RGB and depth is effective. They all assume that depth information is available along with RGB images. In our work, we propose to learn the relation between RGB and depth by a kernelized implicit feature transfer scheme, which enables estimation of depth features from RGB features so as to improve the re-identification performance even though the deployed cameras are not ready for capturing depth information.

A preliminary version of this work appeared in [76]. In this work, apart from providing more in-depth discussion on the proposed Eigen-depth feature and the depth-based person re-identification framework, a kernelized implicit feature transfer scheme is proposed to learn the relation between depth features and RGB features so as to estimate depth features in RGB images when depth sensor is not ready. In addition, more extensive experiments have been conducted.

TABLE I
TERMS AND DEFINITIONS FOR SECTION III.

| symbol | definition | symbol | definition |
|---|---|---|---|
| $\mathbf{f}_{p,i}$ | feature vector of the $i^{th}$ point in voxel $p$ | $\boldsymbol{\mu}_p$ | mean feature vector of voxel $p$ |
| $\mathbf{C}_{Wp}$ | within-voxel covariance of voxel $p$ | $\mathbf{C}_{Bp,q}$ | between-voxel covariance of voxel $p$ and $q$ |
| $\mathbf{C}_1, \mathbf{C}_2$ | any two covariance matrices | $\mathbf{C}_2^N$ | rotation normalized covariance matrix from $\mathbf{C}_2$ to $\mathbf{C}_1$ |
| $\mathbf{U}_p$ | eigenvector matrix of covariance matrix $\mathbf{C}_p$ | $\lambda_{p,q}$ | the $q^{th}$ largest eigenvalue of covariance matrix $\mathbf{C}_p$ |

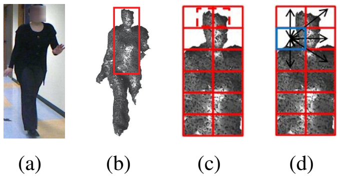


Fig. 3. Illustration of body self-occlusion and feature extraction region. (a) is a sample RGB image, (b) is the corresponding point cloud, (c) is the within-voxel feature extraction region and (d) is the between-voxel feature extraction region.

## III. DEPTH VOXEL COVARIANCE AND EIGEN-DEPTH FEATURE

This section will present the extraction of depth voxel covariance descriptor and locally rotation invariant Eigen-depth feature. Our descriptors are extracted from point cloud, a set of points on object surface expressed by 3D coordinate $(x, y, z)$ in real world converted from depth image. We first tabulate the notations defined in this section in Table I.

### A. Basic Feature Extraction

We first extract basic features of point cloud. We assume another kind of biometric cue, skeleton joints of pedestrian body, is also available along with depth images (e.g., when using Kinect). We intend to extract features on the body parts whose surfaces are more invariant and reliable. As shown in Figure 3 (a) and (b), due to pose difference, sometimes a part of limb surface is not observed under self-occlusion, so the surface shapes of arms and legs contain more noises rather than valuable information. Therefore, we divide each point cloud of the whole body by two shoulder joints and two hip joints and only the points of head and torso are used for feature extraction, while the four limbs are not.

For each point in the point cloud, a normal vector [77] is computed as basic feature. The direction of normal vector describes the shape of a small neighbour region of that point. For a point $\mathbf{x}$, $k$ nearest neighbourhoods of $\mathbf{x}$ are found, and then the direction on which data is least scattered is computed by PCA [78] as the unit normal vector direction $(n_x, n_y, n_z)$. For each point $(x, y, z)$ (unit: mm), a feature vector $F(x, y, z)$ is composed of the coordinate and the unit normal vector

$$F(x, y, z) = [x, y, z, n_x, n_y, n_z]^T. \qquad (1)$$

### B. Depth Voxel Covariance Descriptor

To depict the variation of local feature vectors and alleviate noises, we exploit two types of covariance matrices, namely within-voxel covariance and between-voxel covariance.

*1) Within-voxel Covariance:* We divide a point cloud into rectangular voxels (e.g., 6 × 2 voxels in our case) with 50% overlap, and an example is shown in Figure 3 (c). In each voxel, within-voxel covariance matrix is computed to describe the shape. For a voxel $R_1$, let $\{\mathbf{f}_{1,i}\}_{i=1}^{m}$ be the 6-dimensional feature vectors inside $R_1$. Within-voxel covariance matrix $\mathbf{C}_{W1}$ is then defined as follows:

$$\mathbf{C}_{W1} = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{f}_{1,i} - \boldsymbol{\mu}_1)(\mathbf{f}_{1,i} - \boldsymbol{\mu}_1)^T, \qquad (2)$$

where $\boldsymbol{\mu}_1$ is the mean of the feature vectors of $R_1$.

*2) Between-voxel Covariance:* While within-voxel covariance describes shape in a voxel, the differences of shapes between voxels also contain discriminative information. Similar to standard covariance, we define a novel between-voxel covariance to represent the relation between different voxels.

As shown in Figure 3 (d), the point cloud is divided into $6 \times 2$ voxels without overlap. Between-voxel covariance matrices are computed for each pair of 8-adjacent voxels. For two adjacent voxels $R_1$ and $R_2$, let $\{\mathbf{f}_{1,i}\}_{i=1}^{m}$ and $\{\mathbf{f}_{2,j}\}_{j=1}^{n}$ be the 6-dimensional feature vectors inside $R_1$ and $R_2$, respectively. We define the between-voxel covariance matrix $\mathbf{C}_{B1,2}$ as follows:

$$\mathbf{C}_{B1,2} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{f}_{1,i} - \mathbf{f}_{2,j})(\mathbf{f}_{1,i} - \mathbf{f}_{2,j})^T. \qquad (3)$$

For a depth image, the unification of the within-voxel and between-voxel covariance matrices of all voxels is called the *depth voxel covariance descriptor* (DVCov).

### C. Local Rotation Invariance of Eigenvalues

Assume $p$, $q$ are voxels of a person in depth camera $K$. Let $\{\mathbf{f}_{p,i}^K\}_{i=1}^{N}$ denote the feature vectors, $\mathbf{C}_{Wp}^K$ denote the within-voxel covariance matrix of voxel $p$ and $\mathbf{C}_{Bp,q}^K$ denote between-voxel covariance matrix between voxels $p$ and $q$. We assume only viewpoint rotation and location change take place between two different camera views $A$ and $B$ and the rotation is local so that the body part within voxels $p$ and $q$ will not become invisible due to self-occlusions. To express the transformation from camera view $A$ to camera view $B$, let $\mathbf{R}_{AB1}$ denote the rotation transformation matrix of point coordinate $(x, y, z)$, $\mathbf{R}_{AB2}$ denote the rotation transformation matrix of unit normal vector $(n_x, n_y, n_z)$, and $\mathbf{f}_{AB} = [x_s, y_s, z_s, 0, 0, 0]^T$ denote the shift of pedestrian location. Then the transformations of feature vectors from camera view $A$ to camera view $B$ are

$$\mathbf{f}_{1,i}^B = \mathbf{R}_{AB}(\mathbf{f}_{1,i}^A + \mathbf{f}_{AB}), \qquad (4)$$

$$\mathbf{f}_{2,i}^B = \mathbf{R}_{AB}(\mathbf{f}_{2,i}^A + \mathbf{f}_{AB}), \qquad (5)$$

where $\mathbf{R}_{AB} = \begin{pmatrix} \mathbf{R}_{AB1} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_{AB2} \end{pmatrix}$.

By substituting equations (4), (5) into (2), (3), we have

$$\mathbf{C}_{W1}^{B} = \mathbf{R}_{AB}\mathbf{C}_{W1}^{A}\mathbf{R}_{AB}^{T}, \mathbf{C}_{B1,2}^{B} = \mathbf{R}_{AB}\mathbf{C}_{B1,2}^{A}\mathbf{R}_{AB}^{T}. \tag{6}$$

Since $\mathbf{R}_{AB1}$ and $\mathbf{R}_{AB2}$ satisfy $\mathbf{R}_{AB1}\mathbf{R}_{AB1}^{T} = \mathbf{I}$ and $\mathbf{R}_{AB2}\mathbf{R}_{AB2}^{T} = \mathbf{I}$, we have $\mathbf{R}_{AB}\mathbf{R}_{AB}^{T} = \mathbf{I}$, so that $\mathbf{R}_{AB}$ is orthogonal transformation. Hence, the eigenvalues of within-voxel covariance matrices $\mathbf{C}_{W1}^{B}$ and $\mathbf{C}_{W1}^{A}$ are the same, and the eigenvalues of between-voxel covariance matrices $\mathbf{C}_{B1,2}^{B}$ and $\mathbf{C}_{B1,2}^{A}$ are the same as well. That means the eigenvalues of within-voxel covariance and between-voxel covariance are invariant to rotation and location change.

### *D. Eigen-depth Feature and Analysis*

In this section, we provide more in-depth analysis about the role of those eigenvalues. Let $\mathbf{C}_1, \mathbf{C}_2 \in Sym^{+}(6, \mathbb{R})$ denote two covariance matrices. The eigen-decomposition of $\mathbf{C}_1$ and $\mathbf{C}_2$ are $\mathbf{C}_1 = \mathbf{U}_1\mathrm{diag}(\lambda_{1,1}, \lambda_{1,2}, ..., \lambda_{1,6})\mathbf{U}_1^{T}$ and $\mathbf{C}_2 = \mathbf{U}_2\mathrm{diag}(\lambda_{2,1}, \lambda_{2,2}, ..., \lambda_{2,6})\mathbf{U}_2^{T}$, respectively. Here $\lambda_{1,1} \geq \lambda_{1,2} \geq ... \geq \lambda_{1,6}$ are eigenvalues of $\mathbf{C}_1$, $\lambda_{2,1} \geq \lambda_{2,2} ... \geq \lambda_{2,6}$ are eigenvalues of $\mathbf{C}_2$, and $\mathbf{U}_1$ and $\mathbf{U}_2$ are orthogonal matrices whose columns are the corresponding eigenvectors.

We note that rotation of point clouds and normal vectors can be normalized by matching the principal axes of $\mathbf{C}_2$ and $\mathbf{C}_1$ according to the descending order of eigenvalues. That is, one can find an orthogonal transformation matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{U}_2 = \mathbf{U}_1$, where $\mathbf{Q}$ is the rotation transformation we want to estimate. Hence, we construct a normalized covariance matrix $\mathbf{C}_2^{N} = \mathbf{U}_1\mathrm{diag}(\lambda_{2,1}, \lambda_{2,2}, ..., \lambda_{2,6})\mathbf{U}_1^{T}$, where $\lambda_{2,1}, \lambda_{2,2}, ..., \lambda_{2,6}$ are eigenvalues of $\mathbf{C}_2$ and $\mathbf{U}_1$ contains eigenvectors of $\mathbf{C}_1$. We call $\mathbf{C}_2^{N}$ the rotation normalized covariance matrix from $\mathbf{C}_2$ to $\mathbf{C}_1$.

Now we present how to use the above eigenvalues to construct feature vectors. Let $\mathbf{x}_1 = [\ln\lambda_{1,1}\ \ln\lambda_{1,2}\ ...\ \ln\lambda_{1,6}]^{T}$ and $\mathbf{x}_2 = [\ln\lambda_{2,1}\ \ln\lambda_{2,2}\ ...\ \ln\lambda_{2,6}]^{T}$. Interestingly, we can have the following theorem.

***Theorem** 1:* Computing the Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is equivalent to computing the geodesic distance between covariance matrix $\mathbf{C}_1$ and the rotation normalized covariance matrix $\mathbf{C}_2^{N}$ on the Riemannian manifold.

**Proof.** The Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$\|\mathbf{x}_2 - \mathbf{x}_1\|_2 = \sqrt{\sum_{i=1}^{6}(\ln\lambda_{2,i} - \ln\lambda_{1,i})^2} = \sqrt{\sum_{i=1}^{6}\ln^2\frac{\lambda_{2,i}}{\lambda_{1,i}}}. \tag{7}$$

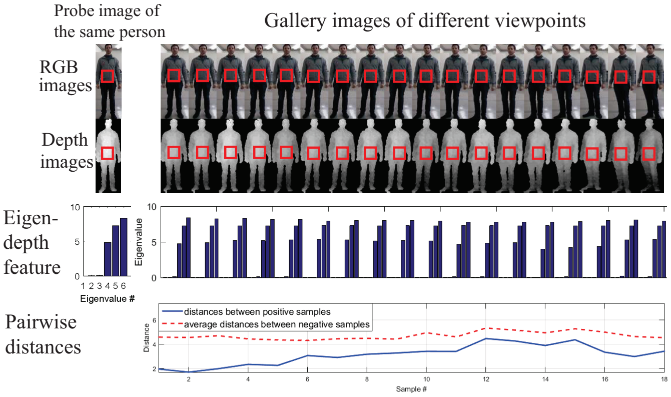


Fig. 4. Visualization of the logarithms of eigenvalues of a fixed voxel and comparison of distance of positive pair (i.e., samples from the same class) and average distance of negative pair (i.e., samples from different classes). The first row shows RGB and depth images of the same person. The second row shows within-voxel Eigen-depth feature of the fixed voxel indicated by red bounding boxes. The third row shows the comparison between the distance of positive pair and the distance of negative pair.

The geodesic distance between $\mathbf{C}_1$ and $\mathbf{C}_2^{N}$ on Riemannian manifold [79] can be calculated as follows:

$$dist(\mathbf{C}_1, \mathbf{C}_2^{N}) = \sqrt{\sum_{k=1}^{6}\ln^2\lambda_k(\mathbf{C}_1, \mathbf{C}_2^{N})}, \tag{8}$$

where $\lambda_k(\mathbf{C}_1, \mathbf{C}_2^{N})_{k=1,...,6}$ are the generalized eigenvalues of $\mathbf{C}_1$ and $\mathbf{C}_2^{N}$, computed by $\lambda_k\mathbf{C}_1\mathbf{z}_k - \mathbf{C}_2^{N}\mathbf{z}_k = 0$, i.e., eigenvalues of $\mathbf{C}_1^{-1}\mathbf{C}_2^{N}$.

$$\begin{aligned}\mathbf{C}_1^{-1}\mathbf{C}_2^{N} &= (\mathbf{U}_1\mathrm{diag}(\lambda_{1,i})\mathbf{U}_1^{T})^{-1}(\mathbf{U}_1\mathrm{diag}(\lambda_{2,i})\mathbf{U}_1^{T}) \\ &= \mathbf{U}_1\mathrm{diag}(\frac{\lambda_{2,i}}{\lambda_{1,i}})\mathbf{U}_1^{T}.\end{aligned} \tag{9}$$

Hence, the $i^{th}$ generalized eigenvalue of $\mathbf{C}_1$ and $\mathbf{C}_2^{N}$ is

$$\lambda_i(\mathbf{C}_1, \mathbf{C}_2^{N}) = \frac{\lambda_{2,i}}{\lambda_{1,i}}. \tag{10}$$

By substituting Equation (10) into (7) and (8), we have

$$\|\mathbf{x}_2 - \mathbf{x}_1\|_2 = dist(\mathbf{C}_1, \mathbf{C}_2^{N}). \tag{11}$$

It can be seen that the geodesic distance on the Riemannian manifold $dist(\mathbf{C}_1, \mathbf{C}_2^{N})$ is equivalent to the Euclidean distance between feature vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.

**Eigen-depth Feature**. The above theorem tells if there exists only local rotation variation, the logarithm eigenvalue vector can convert the distance between covariance matrices on Riemannian manifold to the Euclidean distance between two feature vectors. In our work, we define the Eigen-depth feature (ED) of a covariance matrix $\mathbf{C}_p$ as

$$\mathbf{x}_p = [\ln\lambda_{p,1}\ \ln\lambda_{p,2}\ ...\ \ln\lambda_{p,6}]^{T}, \tag{12}$$

where $\mathbf{C}_p$ is either a within-voxel covariance or a between-voxel covariance. Using eigenvalues makes the feature more compact than using depth voxel covariance descriptor.

To give a direct perception of Eigen-depth features, i.e., the logarithms of eigenvalues, we show some sample images, the Eigen-depth features and distances between positive and negative pairs in Figure 4. For demonstration, we selected one sample as probe image from BIWI RGBD-ID dataset and 18 samples of the same person captured from different views as gallery images. For a fixed voxel indicated in the red bounding box, we extracted its within-voxel Eigen-depth feature and obtained a 6-dimensional feature vector consisting of the logarithms of eigenvalues for each sample. The logarithms of eigenvalues are shown in the second row in Figure 4 by the

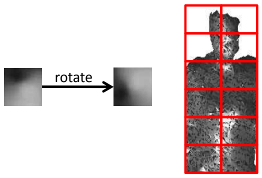


Fig. 5. The left illustrates the confusion caused by rotation invariance, and the right shows how voxels are divided.

histograms. We can find that the histograms of eigenvalues look very similar over the rotation change. Since there are still extra variations but not just local rotation variation in practice, we further make comparison between the distance of positive pair (i.e., samples from the same class) and the distance of negative pair (i.e., samples from different classes) on the third row of Figure 4. We first computed the Euclidean distance between the probe image and each gallery image given above as the distance of positive pair (plotted as blue curve), and computed the average distance between each gallery image and all samples from other classes in this dataset as the distance of negative pair (plotted as red dashed curve). We can observe that the distance between samples of the same class across multiple view angles is less sensitive to rotation in practice. Moreover, the distance of positive pair is smaller than the average distance of negative pair. So the Eigen-depth feature is a useful shape descriptor for recognition.

**Remark.** In existing literatures about covariance descriptor such as [17], geodesic distance on Riemannian manifold is used for measuring the similarities between covariance matrices. However, directly using geodesic distance is not invariant to rotation. Given two rotation transformation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, the eigenvalues of $\mathbf{C}_1^{-1}\mathbf{C}_2$ and $\mathbf{R}_1\mathbf{C}_1^{-1}\mathbf{R}_1^T\mathbf{R}_2\mathbf{C}_2\mathbf{R}_2^T$ are always different. Moreover, covariance matrix does not lie on Euclidean space, so most common machine learning methods are not proper to be applied directly.

In practice, although Eigen-depth feature is proved to be locally invariant to rotation, some problems come along with this property. As shown in Figure 5 on the left, given depth images (in which grayscale denotes depth) of two different voxels of body surface, they can be transformed to each other by rotation. Obviously, their shapes are clearly different but the Eigen-depth features of within-voxel covariance matrices are the same, and such a situation could make confusion in the matching stage, which is also a problem for other rotation invariant depth shape descriptors. This kind of confusion could take place if the voxel size is too small and the voxel contains only a small region of body surface. To alleviate this problem, as illustrated in Figure 5 on the right, we divide the point cloud into 6 × 2 voxels to extract feature for more robust representation. So the voxels are large enough to contain a big area of body surface, making it less possible to cause confusion after rotation.

## IV. Depth-based Re-identification Framework

In the previous section, we have extracted depth voxel covariance descriptors (DVCov), and constructed Eigen-depth feature (ED) for describing body shape. Besides using body shape, incorporating more physical information would have extra benefit on the identification of a person. As indicated in the previous section, the four limbs are not suitable for extracting invariant shape representation, but the lengths of limbs contain physical information which is also a biometric cue for distinguishing people. Hence, to obtain a complete feature representation of pedestrian, we additionally employ the skeleton-based feature (SKL) as complementary physical information, and then build a depth-based re-identification framework by combining the proposed depth shape descriptors and the skeleton-based feature together.

The whole framework is illustrated in Figure 2. For the feature representation of skeleton, we apply the skeleton-based feature in [2]. This skeleton-based feature is a feature vector that contains 13 distance values and ratios computed from the positions of skeleton joints provided by a skeleton tracker. The elements of the feature vector includes: (a) head height, (b) neck height, (c) neck to left shoulder distance, (d) neck to right shoulder distance, (e) torso to right shoulder distance, (f) right arm length, (g) left arm length, (h) right upper leg length, (i) left upper leg length, (j) torso length, (k) right hip to left hip distance, (l) ratio between torso length and right upper leg length (i.e., j/h) and (m) ratio between torso length and left upper leg length (i.e., j/i) (the unit of distances is cm).

After extracting skeleton-based feature, in the stage of feature fusion, we combine our proposed depth shape descriptors and the skeleton-based feature together to form complete representation of human body. In this work, we offer two fusion models below.

- **DVCov+SKL:** When the viewpoint variation of a person across camera views is not large in some special cases (e.g., security check or walking in narrow passage), the influence of rotation can be secondary. In such cases, we select our depth voxel covariance descriptor as depth shape descriptor, as it contains richer information about texture and is more effective for describing the shape. We measure the similarity of two subjects by computing the combined distance $d = d_{DVCov} + d_{SKL}$, where $d_{DVCov}$ denotes the sum of geodesic distances between the corresponding within-voxel covariance matrices and between-voxel covariance matrices, and $d_{SKL}$ denotes the Euclidean distance between skeleton-based features.
- **ED+SKL:** When the viewpoint variation of a pedestrian across different camera views is large, we select Eigen-depth feature as depth shape descriptor since it is locally rotation invariant. Let $\mathbf{x}_W$ and $\mathbf{x}_B$ denote the concatenated Eigen-depth feature vectors of all within-voxel covariance matrices and all between-voxel covariance matrices of a person. Let $\mathbf{x}_{SKL}$ denote the skeleton-based feature. We fuse these three features by concatenating them to obtain a combined feature $\mathbf{x}_C = [\mathbf{x}_W^T \ \mathbf{x}_B^T \ \mathbf{x}_{SKL}^T]^T$. Then we apply LDA [78] to $\mathbf{x}_C$ for feature selection. We first reduce feature dimension to 100 by principal component analysis (PCA) and then extract $c - 1$ discriminant vectors by LDA, where $c$ is the number of classes. After dimension reduction, the projected features are matched by using Euclidean distance.

TABLE II
TERMS AND DEFINITIONS FOR SECTION V.

| symbol | definition | symbol | definition |
|---|---|---|---|
| $\mathbf{f}_i^v$, $\mathbf{f}_i^d$ | visual/depth feature of the $i^{th}$ sample | $y_i$ | label of the $i^{th}$ sample |
| $\phi_v(\cdot)$, $\phi_d(\cdot)$ | nonlinear mapping for visual/depth feature | $\mathbf{W}_v$, $\mathbf{W}_d$ | projection matrix for nonlinear visual/depth feature |
| $\mathbf{S}_{bv}$, $\mathbf{S}_{bd}$ | between-class scatter matrix for visual/depth feature | $\mathbf{S}_{wv}$, $\mathbf{S}_{wd}$ | within-class scatter matrix for visual/depth feature |
| $\mathrm{K}_v(\cdot,\cdot)$ , $\mathrm{K}_d(\cdot,\cdot)$ | kernel function for visual/depth features | $\mathbf{A}_v$, $\mathbf{A}_d$ | combination coefficient matrix for visual/depth features |

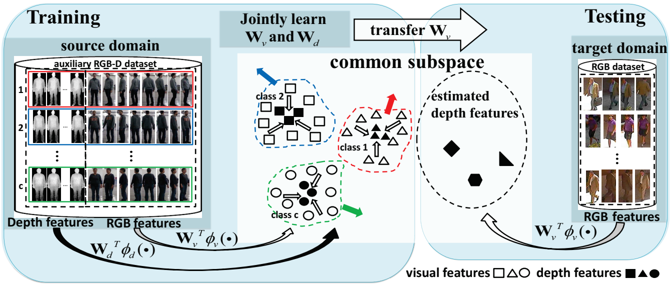


Fig. 6. Overview of kernelized implicit feature transfer scheme. Visual features and depth features are mapped by projections to a common feature subspace.

## V. DEPTH FEATURE TRANSFER

We have proposed a depth-based person re-identification framework in previous sections. However, in most existing surveillance systems, a large amount of cameras do not support capturing depth information, so only RGB images are available. In this section, we exploit a transfer technique to implicitly estimate depth features for RGB person images when depth device is not ready. We tabulate the notations defined in this section in Table II.

### A. Kernelized Implicit Feature Transfer Scheme

Depth features can describe body shape of a person, while some visual features (e.g., HOG [13] and LBP [80]) extracted from RGB images can also describe body shape coarsely to some extent. Therefore, we aim to learn the relation between depth features and these kinds of visual features, so as to estimate the depth features from RGB images when depth device is not ready.

For this purpose, we assume an auxiliary RGB-D dataset is given. This RGB-D dataset is regarded as source domain, and the RGB images from which we want to estimate depth features are regarded as target domain. We propose a kernelized implicit feature transfer scheme to transfer the depth feature from source domain to target domain. The overview of the feature transfer procedure is shown in Figure 6.

In details, suppose there exists an auxiliary RGB-D dataset that consists of RGB-D images for each person. Let the source domain samples be denoted by $\{(\mathbf{f}_i^v, \mathbf{f}_i^d, y_i)\}_{i=1}^{N_s}$, where $N_s$ is the total number of samples, $\mathbf{f}_i^v$ denotes the visual feature of the $i^{th}$ sample, $\mathbf{f}_i^d$ denotes the depth feature of the $i^{th}$ sample, and $y_i \in \{1, 2, ..., C\}$ denotes the label ($C$ is the total number of classes/identities). Depth feature and visual feature are heterogeneous features, and we assume that they can be mapped onto a common latent subspace if they are transformed onto high dimensional nonlinear space implicitly by some kernel functions. Let us denote the nonlinear visual feature as $\phi_v(\mathbf{f}^v) \in \mathbb{R}^{m_v}$ and the nonlinear depth feature as $\phi_d(\mathbf{f}^d) \in \mathbb{R}^{m_d}$, where the dimensions $m_v$ and $m_d$ are unknown. Then we project the visual features and depth features onto a common latent subspace by projection matrices $\mathbf{W}_v \in \mathbb{R}^{m_v \times m}$ and $\mathbf{W}_d \in \mathbb{R}^{m_d \times m}$, respectively, where $m$ is the dimension of the common latent subspace. In the common latent subspace, we aim to make the projected visual feature $\mathbf{W}_v^T\phi_v(\mathbf{f}^v)$ close to the corresponding depth feature $\mathbf{W}_d^T\phi_d(\mathbf{f}^d)$. For this purpose, we minimize the distance between the means of RGB-based visual features and the depth features of each person in the common latent subspace by minimizing

$$\Omega_{vd} = \frac{1}{C}\sum_{c=1}^{C}\left\|\frac{1}{N_c}\sum_{y_i=c}\mathbf{W}_v^T\phi_v(\mathbf{f}_i^v) - \frac{1}{N_c}\sum_{y_i=c}\mathbf{W}_d^T\phi_d(\mathbf{f}_i^d)\right\|_2^2. \tag{13}$$

In addition, we wish that the above transformation between depth and RGB features is learned in a discriminative way. In order to make both visual features and depth features discriminative in the common latent subspace, it is expected to minimize the within-class variance and maximize the between-class variance of both visual features and depth features at the same time. The between-class scatter matrices and within-class scatter matrices are defined as follows:

$$\mathbf{S}_{b*} = \sum_{i,j=1}^{N_s}\mathbf{A}_{i,j}^b(\phi_*(\mathbf{f}_i^*) - \phi_*(\mathbf{f}_j^*))(\phi_*(\mathbf{f}_i^*) - \phi_*(\mathbf{f}_j^*))^T, \tag{14}$$

$$\mathbf{S}_{w*} = \sum_{i,j=1}^{N_s}\mathbf{A}_{i,j}^w(\phi_*(\mathbf{f}_i^*) - \phi_*(\mathbf{f}_j^*))(\phi_*(\mathbf{f}_i^*) - \phi_*(\mathbf{f}_j^*))^T, \tag{15}$$

where $* \in \{v, d\}$, $v$ denotes visual feature and $d$ denotes depth feature, and

$$\mathbf{A}_{i,j}^b = \begin{cases}\frac{1}{N_s} - \frac{1}{N_c} \text{ if } y_i = y_j = c,\\ \frac{1}{N_s} \text{ if } y_i \neq y_j,\end{cases} \tag{16}$$

$$\mathbf{A}_{i,j}^w = \begin{cases}\frac{1}{N_c} \text{ if } y_i = y_j = c,\\ 0 \text{ if } y_i \neq y_j,\end{cases} \tag{17}$$

and $N_c$ is the number of samples of class $c$.

Then we combine the minimization of $\Omega_{vd}$ with the discriminant feature learning that maximizes between-class variance while minimizes within-class variance as follows:

$$\max_{\mathbf{W}_v,\mathbf{W}_d} \frac{\gamma_0\mathrm{tr}(\mathbf{W}_v^T\mathbf{S}_{bv}\mathbf{W}_v) + \gamma_1\mathrm{tr}(\mathbf{W}_d^T\mathbf{S}_{bd}\mathbf{W}_d)}{\beta'\Omega_{vd} + \gamma_0'\mathrm{tr}(\mathbf{W}_v^T\mathbf{S}_{wv}\mathbf{W}_v) + \gamma_1'\mathrm{tr}(\mathbf{W}_d^T\mathbf{S}_{wd}\mathbf{W}_d)}, \tag{18}$$

where $\mathbf{S}_{bv}(\mathbf{S}_{bd})$ and $\mathbf{S}_{wv}(\mathbf{S}_{wd})$ are between-class scatter matrix and within-class scatter matrix of visual features (depth features) respectively, and $\beta'$, $\gamma_0$, $\gamma_0'$, $\gamma_1$ and $\gamma_1'$ are non-negative trade-off parameters. We call the above transfer model the *kernelized implicit feature transfer scheme*. It is unsupervised without using information in target domain.

### *B. Optimization*

We show that the model developed in the last section can be converted to a generalized eigen-decomposition problem. As suggested by the representer theorem [81], the projection matrices can be represented by the combination of training samples, i.e., $\mathbf{W}_v = \mathbf{\Phi}_v \mathbf{A}_v$, $\mathbf{W}_d = \mathbf{\Phi}_d \mathbf{A}_d$, where $\mathbf{\Phi}_v = [\phi_v(\mathbf{f}_1^v), \phi_v(\mathbf{f}_2^v), ..., \phi_v(\mathbf{f}_{N_s}^v)] \in \mathbb{R}^{m_v \times N_s}$ and $\mathbf{\Phi}_d = [\phi_d(\mathbf{f}_1^d), \phi_d(\mathbf{f}_2^d), ..., \phi_d(\mathbf{f}_{N_s}^d)] \in \mathbb{R}^{m_d \times N_s}$ are visual feature matrix and depth feature matrix of training samples respectively and $\mathbf{A}_v \in \mathbb{R}^{N_s \times m}$ and $\mathbf{A}_d \in \mathbb{R}^{N_s \times m}$ are the combination coefficient matrices. For visual feature $\mathbf{f}^v$ and depth feature $\mathbf{f}^d$, we define

$$\widetilde{\mathbf{f}^v} = [\mathrm{K}_v(\mathbf{f}_1^v, \mathbf{f}^v), \mathrm{K}_v(\mathbf{f}_2^v, \mathbf{f}^v), ..., \mathrm{K}_v(\mathbf{f}_{N_s}^v, \mathbf{f}^v)]^T, \tag{19}$$

$$\widetilde{\mathbf{f}^d} = [\mathrm{K}_d(\mathbf{f}_1^d, \mathbf{f}^d), \mathrm{K}_d(\mathbf{f}_2^d, \mathbf{f}^d), ..., \mathrm{K}_d(\mathbf{f}_{N_s}^d, \mathbf{f}^d)]^T, \tag{20}$$

where $\mathrm{K}_v(\cdot, \cdot)$ and $\mathrm{K}_d(\cdot, \cdot)$ are kernel functions for visual feature and depth feature, respectively. The projection of a visual feature $\mathbf{f}^v$ is expressed as $\mathbf{W}_v^T \phi_v(\mathbf{f}^v) = \mathbf{A}_v^T \mathbf{\Phi}_v^T \phi_v(\mathbf{f}^v) = \mathbf{A}_v^T \widetilde{\mathbf{f}^v}$. In the same way, $\mathbf{W}_d^T \phi_d(\mathbf{f}^d) = \mathbf{A}_d^T \widetilde{\mathbf{f}^d}$. To jointly solve $\mathbf{A}_v$ and $\mathbf{A}_d$, we define $\mathbf{A} = [\mathbf{A}_v^T, \mathbf{A}_d^T]^T$, zero-padding transformation matrix $\mathbf{Z}_v = [\mathbf{I}_{N_s}, \mathbf{O}_{N_s \times N_s}]$ for $\widetilde{\mathbf{f}^v}$ and $\mathbf{Z}_d = [\mathbf{O}_{N_s \times N_s}, \mathbf{I}_{N_s}]$ for $\widetilde{\mathbf{f}^d}$. Now the objective function (18) can be reformulated as:

$$\max_{\mathbf{A}} \frac{\gamma_0 \mathrm{tr}(\mathbf{A}^T \mathbf{Z}_v^T \widetilde{\mathbf{S}}_{bv} \mathbf{Z}_v \mathbf{A}) + \gamma_1 \mathrm{tr}(\mathbf{A}^T \mathbf{Z}_d^T \widetilde{\mathbf{S}}_{bd} \mathbf{Z}_d \mathbf{A})}{\beta' \widetilde{\Omega}_{vd} + \gamma_0' \mathrm{tr}(\mathbf{A}^T \mathbf{Z}_v^T \widetilde{\mathbf{S}}_{wv} \mathbf{Z}_v \mathbf{A}) + \gamma_1' \mathrm{tr}(\mathbf{A}^T \mathbf{Z}_d^T \widetilde{\mathbf{S}}_{wd} \mathbf{Z}_d \mathbf{A})}, \tag{21}$$

where $\widetilde{\mathbf{S}}_{bv}$, $\widetilde{\mathbf{S}}_{wv}$, $\widetilde{\mathbf{S}}_{bd}$, $\widetilde{\mathbf{S}}_{wd}$ are scatter matrices defined by $\widetilde{\mathbf{S}}_{b*} = \sum_{i,j=1}^{N_s} \mathbf{A}_{i,j}^b (\widetilde{\mathbf{f}_i^*} - \widetilde{\mathbf{f}_j^*})(\widetilde{\mathbf{f}_i^*} - \widetilde{\mathbf{f}_j^*})^T$, and $\widetilde{\mathbf{S}}_{w*} = \sum_{i,j=1}^{N_s} \mathbf{A}_{i,j}^w (\widetilde{\mathbf{f}_i^*} - \widetilde{\mathbf{f}_j^*})(\widetilde{\mathbf{f}_i^*} - \widetilde{\mathbf{f}_j^*})^T$, $* \in \{v, d\}$.

$$\begin{aligned} \widetilde{\Omega}_{vd} &= \frac{1}{C} \sum_{c=1}^{C} \left\| \frac{1}{N_c} \sum_{y_i = c} \mathbf{A}^T \mathbf{Z}_v^T \widetilde{\mathbf{f}_i^v} - \frac{1}{N_c} \sum_{y_i = c} \mathbf{A}^T \mathbf{Z}_d^T \widetilde{\mathbf{f}_i^d} \right\|_2^2 \\ &= \mathrm{tr}(\mathbf{A}^T \mathbf{B}_{vd} \mathbf{A}), \end{aligned} \tag{22}$$

where $\mathbf{B}_{vd} = \frac{1}{C} \sum_{c=1}^{C} \mathbf{U}_c \mathbf{U}_c^T$, $\mathbf{U}_c = \frac{1}{N_c} \sum_{y_i = c} (\mathbf{Z}_v^T \widetilde{\mathbf{f}_i^v} - \mathbf{Z}_d^T \widetilde{\mathbf{f}_i^d})$.

Let $\mathbf{B}_{bv} = \mathbf{Z}_v^T \widetilde{\mathbf{S}}_{bv} \mathbf{Z}_v$, $\mathbf{B}_{wv} = \mathbf{Z}_v^T \widetilde{\mathbf{S}}_{wv} \mathbf{Z}_v$, $\mathbf{B}_{bd} = \mathbf{Z}_d^T \widetilde{\mathbf{S}}_{bd} \mathbf{Z}_d$, $\mathbf{B}_{wd} = \mathbf{Z}_d^T \widetilde{\mathbf{S}}_{wd} \mathbf{Z}_d$ denote the zero-padding scatter matrices. Finally, the objective function is formulated by:

$$\begin{aligned} \max_{\mathbf{A}} \quad & \mathrm{tr}(\mathbf{A}^T \mathbf{B}_1 \mathbf{A}) \\ s.t. \quad & \mathbf{A}^T \mathbf{B}_2 \mathbf{A} = \mathbf{I}, \end{aligned} \tag{23}$$

where $\mathbf{B}_1 = \gamma_0 \mathbf{B}_{bv} + \gamma_1 \mathbf{B}_{bd}$, $\mathbf{B}_2 = \beta' \mathbf{B}_{vd} + \gamma_0' \mathbf{B}_{wv} + \gamma_1' \mathbf{B}_{wd}$. Hence, a generalized eigen-decomposition problem can be derived below:

$$\mathbf{B}_1 \mathbf{A} = \lambda \mathbf{B}_2 \mathbf{A}. \tag{24}$$

Solving the above is to compute the eigen-decomposition $\mathbf{B}_2^{-1} \mathbf{B}_1 = \mathbf{A} \mathbf{\Lambda} \mathbf{A}^T$, in which $\mathbf{\Lambda}$ is a diagonal matrix with sorted eigenvalues in descending order lying on the diagonal and $\mathbf{A} \in \mathbb{R}^{2N_s \times 2N_s}$ contains the eigenvectors. Since $\mathbf{A} = [\mathbf{A}_v^T, \mathbf{A}_d^T]^T$, we can obtain $\mathbf{A}_v$ by extracting the first $N_s$ rows of $\mathbf{A}$. To specify the dimension of the common latent subspace, we use the first $m$ columns of $\mathbf{A}_v$ to form the projection matrix $\mathbf{A}_v' \in \mathbb{R}^{N_s \times m}$ so that we can project visual feature to the $m$-dimensional common latent subspace.

### *C. Depth Feature Estimation on Target Domain*

After learning the projection to the discriminative common latent subspace, we can implicitly estimate the depth feature of an RGB image in target domain by mapping visual feature to high-dimensional nonlinear space by $\phi_v(\cdot)$ and projecting it to the learned common latent subspace by $\mathbf{W}_v$. Given two new samples $p_1$ and $p_2$ in target domain, let $\mathbf{f}_{p1}^v$, $\mathbf{f}_{p2}^v$ denote the visual features of RGB images. The estimated depth features in the learned discriminative common latent subspace are computed by $\widehat{\mathbf{f}}_{p1}^d = \mathbf{A}_v'^T \widetilde{\mathbf{f}}_{p1}^v$ and $\widehat{\mathbf{f}}_{p2}^d = \mathbf{A}_v'^T \widetilde{\mathbf{f}}_{p2}^v$. Then the distance of depth features between $p_1$ and $p_2$ is computed by

$$dist_D(p_1, p_2) = ||\widehat{\mathbf{f}}_{p1}^d - \widehat{\mathbf{f}}_{p2}^d||_2. \tag{25}$$

## VI. EXPERIMENTS

Our depth-based person re-identification framework was evaluated on three RGB-D person re-identification datasets PAVIS [1], BIWI RGBD-ID [2] and IAS-Lab RGBD-ID [3], which were captured by Kinect. In Section VI-E, the kernelized implicit feature transfer scheme was evaluated on 3DPeS [21] and CAVIAR4REID [20]. The experiment results were presented in Cumulative Matching Characteristic (CMC) curve [82] and rank-$k$ accuracy. Rank-$k$ accuracy is the cumulative recognition rate of correct matches at rank $k$. The CMC curve represents the cumulative recognition rates at all ranks. The evaluation was repeated 10 times and average results were reported.

**Compared Methods**. By following the general re-id setting, we tested Eigen-depth feature (ED), our depth voxel covariance descriptor (DVCov), skeleton-based feature (SKL) and the combinations of depth shape descriptors and skeleton-based feature (ED+SKL and DVCov+SKL). We conducted comparisons with RGB-based appearance features including LOMO feature [12], ELF18 feature [50], color histograms (RGB, HS and YCbCr space) [11], HOG [13] and LBP [80], rotation invariant depth shape descriptors including RIFT2M [5] and Fehr's covariance descriptor [6], and skeleton-based feature designed for depth-based re-id [2]. All RGB-based appearance features were extracted from images which were resized to 128×48. RIFT2M and Fehr's descriptor were densely extracted using the same voxels as Eigen-depth feature. We used LDA to learn the distance metric for all features, except that the skeleton-based feature was matched by Euclidean distance and our depth voxel covariance descriptor was matched by geodesic distance using Equation (8).

### *A. Evaluation on PAVIS*

We used two groups of dataset images in PAVIS dataset [1] for evaluation here. These two groups are denoted by "Walking1" and "Walking2". Images of "Walking1" and "Walking2" were obtained by recording the same 79 people with a frontal view, walking slowly in an indoor scenario. Among the 79 people, 60 people in "Walking2" dressed different clothes from "Walking1".

The characteristic of this experiment is that some people changed their clothes (by wearing one more red shirt) from "Walking1" to "Walking2" as shown in Figure 7. However, one

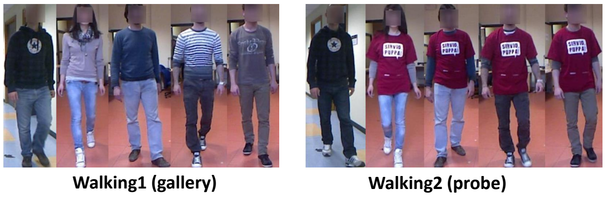


Fig. 7. Examples of images in “Walking1” and “Walking2” in PAVIS. Most persons in “Walking2” dressed different clothes from “Walking1”.

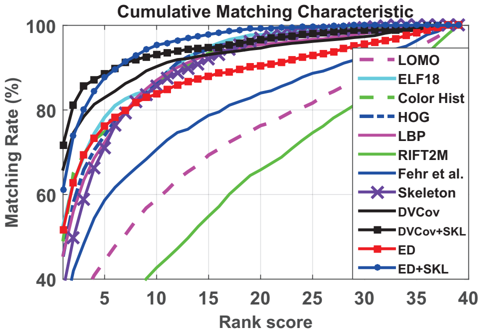


Fig. 8. Performance on PAVIS (multi-shot). Our approach: ED (Eigen-depth feature), ED+SKL, DVCov (our depth voxel covariance descriptor), DVCov+SKL.

could still explore some appearance cues (e.g., trousers and body shape) for matching persons across these two sets. Since the images of frontal bodies were captured from nearly the same view in these two sets, there was little rotation variation of point clouds. In this case, we can apply DVCov+SKL in our framework.

We used images in “Walking1” to form the gallery set and the images in “Walking2” to form the probe. By following the usual train-test policy for person re-identification, we randomly sampled half of the group “Walking1”, i.e., images of 40 persons for training, and the remaining 39 persons were used for testing. Images of these 39 testing persons in “Walking1” were randomly selected as gallery and all images of these 39 persons in “Walking2” were used as probe. In single-shot experiments, one image of each person was randomly selected as gallery. In multi-shot experiments, five images of a person were selected as gallery, and in such a case the distance between each probe image and each gallery class was the minimum distance between each probe image and each gallery image of that class. The performance of the tested methods was reported in Figure 8, Figure 9 (a) and Table III.

The results suggest that both Eigen-depth feature (ED) and our depth voxel covariance descriptor (DVCov) are more effective than RIFT2M and Fehr’s descriptor for describing body shape. Since view angles of persons are nearly the same in “Walking1” and “Walking2”, our depth voxel covariance descriptor is more effective than Eigen-depth feature, because it contains richer information about textures than using only eigenvalues. However, Eigen-depth feature is still more effective than other methods except for our depth voxel covariance descriptor. Using skeleton-based feature alone cannot achieve high performance, but it is complementary information for our depth voxel covariance descriptor and Eigen-depth feature. The combination of our depth voxel covariance descriptor and skeleton-based feature achieves encouraging performance, where the rank-1 accuracy is 67.64% for single-shot recognition and 71.74% for multi-shot recognition. It is clear that the fusion outperformed RGB appearance-based methods and other tested depth-based methods. We note that not all RGB-based appearance features performed badly as we expected, because among the 79 people, 19 people did not change clothes and the other 60 people’s trousers did not change as well from the gallery set to the probe. In conclusion, this test showed the effectiveness of our depth voxel covariance descriptor for shape description.

TABLE III
PAVIS DATASET: RANK-1 AND RANK-5 ACCURACIES (%), INCLUDING RESULTS OF OUR PROPOSED METHODS AND COMPARISONS WITH RGB-BASED APPEARANCE FEATURES AND DEPTH-BASED FEATURES.

| Setting | Single-shot | | Multi-shot | |
|---|---|---|---|---|
| Rank | 1 | 5 | 1 | 5 |
| RGB-based appearance features | | | | |
| LOMO [12] | 12.05 | 35.03 | 19.74 | 44.36 |
| ELF18 [50] | 52.15 | 77.85 | 52.62 | 78.26 |
| Color Hist [11] | 47.90 | 74.97 | 48.92 | 74.82 |
| HOG [13] | 45.03 | 73.49 | 45.33 | 73.95 |
| LBP [80] | 42.92 | 71.33 | 45.64 | 72.36 |
| Depth-based features | | | | |
| RIFT2M [5] | 7.13 | 22.77 | 8.77 | 27.69 |
| Fehr’s [6] | 24.26 | 51.64 | 30.56 | 58.67 |
| Skeleton [2] | 33.13 | 67.85 | 37.33 | 71.13 |
| Proposed | | | | |
| DVCov (depth voxel covariance) | 61.49 | 81.23 | 66.00 | 82.92 |
| DVCov+SKL | **67.64** | **87.33** | **71.74** | **88.46** |
| ED (Eigen-depth feature) | 44.67 | 72.10 | 51.59 | 76.15 |
| ED+SKL | 55.95 | 84.77 | 61.23 | 87.64 |

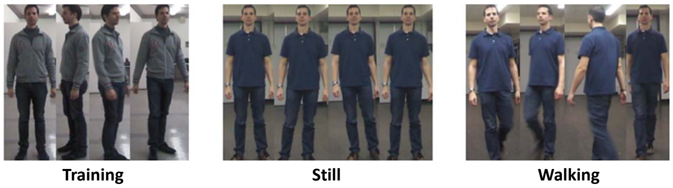
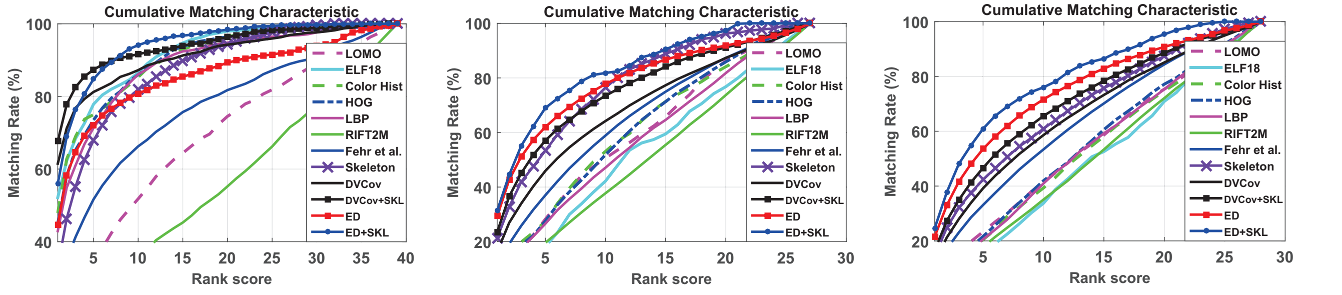


Fig. 10. Examples of images in BIWI RGBD-ID. In “Still”, the persons were captured from frontal view, while in “Training” and “Walking” the persons were captured from multiple views.

### B. Evaluation on BIWI RGBD-ID

The BIWI RGBD-ID dataset [2] contains three groups of sequences “Training”, “Still” and “Walking” captured from 50 different people. For a sequence of each person, there are about 300 frames of depth images and skeletons. Before feature extraction, we converted depth images to point clouds. In “Training”, people performed motions, such as walking and rotating. Only 28 people presented in “Training” were recorded in “Still” and “Walking”, which were collected in a different day and in a different scene, so that most persons were dressed differently. In “Still”, people slightly moved, while in “Walking”, every person walked in different view angles. Examples of images in “Training”, “Still” and “Walking” are shown in Figure 10. Since pedestrians’ viewpoint variation was large here, it was more suitable to use ED+SKL in our framework.



(a) PAVIS (b) BIWI RGBD-ID “Still” (c) BIWI RGBD-ID “Walking”

Fig. 9. Performance on PAVIS and BIWI RGBD-ID (single-shot). Our approach: ED (Eigen-depth feature), ED+SKL, DVCov (our depth voxel covariance descriptor), DVCov+SKL.

TABLE IV
BIWI RGBD-ID DATASET “STILL” AND “WALKING”: RANK-1 AND RANK-5 ACCURACIES (%), INCLUDING RESULTS OF OUR PROPOSED METHODS AND COMPARISONS WITH RGB-BASED APPEARANCE FEATURES AND DEPTH-BASED FEATURES.

| Probe | Still | | | | Walking | | | |
|---|---|---|---|---|---|---|---|---|
| Setting | Single-shot | | Multi-shot | | Single-shot | | Multi-shot | |
| Rank | 1 | 5 | 1 | 5 | 1 | 5 | 1 | 5 |
| RGB-based appearance features | | | | | | | | |
| LOMO [12] | 9.07 | 28.21 | 18.17 | 35.47 | 8.74 | 23.33 | 10.31 | 25.39 |
| ELF18 [50] | 2.79 | 18.18 | 4.11 | 19.13 | 1.32 | 16.03 | 1.50 | 16.77 |
| Color Hist [11] | 7.02 | 25.47 | 10.61 | 31.92 | 5.43 | 19.56 | 5.86 | 21.70 |
| HOG [13] | 8.42 | 25.69 | 12.35 | 30.39 | 6.38 | 21.00 | 6.94 | 23.29 |
| LBP [80] | 7.37 | 26.04 | 10.87 | 33.57 | 4.87 | 20.04 | 5.34 | 23.31 |
| Depth-based features | | | | | | | | |
| RIFT2M [5] | 4.04 | 19.52 | 4.34 | 20.78 | 3.25 | 17.46 | 3.75 | 18.31 |
| Fehr’s [6] | 12.08 | 38.17 | 14.06 | 43.78 | 9.33 | 32.39 | 12.09 | 39.60 |
| Skeleton [2] | 21.34 | 53.32 | 26.55 | 62.73 | 14.52 | 42.36 | 16.94 | 47.18 |
| Proposed | | | | | | | | |
| DVCov | 16.32 | 45.93 | 23.07 | 58.89 | 12.58 | 39.22 | 17.24 | 45.93 |
| DVCov+SKL | 23.49 | 57.06 | 34.37 | 72.77 | 16.59 | 46.67 | 21.40 | 54.12 |
| ED | 28.98 | 61.85 | 36.22 | **73.11** | 20.90 | 51.98 | 28.71 | 63.85 |
| ED+SKL | **30.52** | **67.86** | **39.38** | 72.13 | **24.47** | **60.63** | **29.96** | **65.18** |

For BIWI RGBD-ID dataset, images of the 22 people who only appeared in “Training” were used for training, and images of the remaining 28 people were used for testing. In the testing set, we used images in “Training” as gallery and images in “Still” and “Walking” as probe, so the same person wore different clothes in gallery and probe. We selected the samples for evaluation by face detection as advised in [2]. Since the persons were captured from different view angles, this dataset is suitable to evaluate the effect of the local rotation invariance property of the proposed Eigen-depth feature. The average results of CMC curve and rank-$k$ accuracy over 10 trials were reported in Figure 9 (b), (c) and Table IV.

As shown in Figure 9 (b), (c), RGB-based appearance features completely failed, because most people changed clothes so that color feature was not reliable. Our depth-based methods outperformed all RGB appearance-based methods. On BIWI RGBD-ID, people appeared in different view angles, so the problem became more challenging than the one on PAVIS. On “Walking”, the problem was even more difficult since more frames were captured in multiple viewpoints. In these situations, rotation invariant depth shape descriptor is more suitable, so that our Eigen-depth feature outperformed our depth voxel covariance descriptor. Compared with other rotation invariant depth shape descriptors, Eigen-depth feature outperformed RIFT2M and Fehr’s descriptor. The combination of Eigen-depth feature and skeleton-based feature (ED+SKL) can achieve better performance than using them separately, which is the best on BIWI RGBD-ID. The results showed the local rotation invariance and the effectiveness of body shape description of Eigen-depth feature.

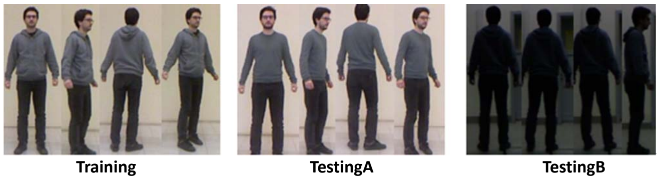


Fig. 11. Examples of images in IAS-Lab RGBD-ID. All samples were captured from multiple views. Compared to “Training”, samples in “TestingA” changed clothes and some samples in “TestingB” were captured in dark environment.

### *C. Evaluation on IAS-Lab RGBD-ID*

There are 11 different people in IAS-Lab RGBD-ID dataset [3]. In this dataset, three groups of sequences “Training”, “TestingA” and “TestingB” were recorded, and each person rotated on himself and walked during the recording. There are about 500 frames of depth images and skeletons for each person. The sequences in “Training” and “TestingA” were acquired when the same person was wearing different clothes. The sequences in “TestingB” were collected in a different room, where each person dressed the same as in “Training”. Some sequences in “TestingB” were recorded in dark environment. Examples of images in “Training”, “TestingA” and “TestingB” are shown in Figure 11.

On this dataset, the evaluation also followed the settings on PAVIS. Half of “Training” sequences were randomly selected to form the training set and the rest were selected to form the gallery in the test. The samples in “TestingA” and “TestingB” corresponding to the gallery persons were selected to form the probe set. By following the settings in [3], all images were used in this experiment. On this dataset, mismatch would be observed when performing the matching between a person image of rear view and his/her image of frontal view, so that it challenges body shape descriptors. The average rank-1 and rank-3 accuracies over 10 trials of evaluation were reported in Table V.

TABLE V
IAS-LAB RGBD-ID DATASET "TESTINGA" AND "TESTINGB": RANK-1 AND RANK-3 ACCURACY (%), INCLUDING RESULTS OF OUR PROPOSED METHODS AND COMPARISONS WITH RGB-BASED APPEARANCE FEATURES AND DEPTH-BASED FEATURES.

| Probe | TestingA | | | | TestingB | | | |
|---|---|---|---|---|---|---|---|---|
| Setting | Single-shot | | Multi-shot | | Single-shot | | Multi-shot | |
| Rank | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 |
| RGB-based appearance features | | | | | | | | |
| LOMO [12] | 26.37 | 65.82 | 25.79 | 66.28 | 30.97 | 75.00 | 30.06 | 79.90 |
| ELF18 [50] | 22.35 | 60.96 | 21.81 | 67.77 | 24.03 | 67.36 | 23.01 | 67.81 |
| Color Hist [11] | 27.69 | 63.71 | 24.42 | 66.48 | 18.45 | 63.33 | 23.89 | 60.93 |
| HOG [13] | 31.00 | 66.48 | 38.89 | 72.67 | 47.21 | 81.16 | 49.62 | 86.79 |
| LBP [80] | 28.71 | 67.97 | 32.81 | 68.22 | 51.38 | 84.28 | 52.88 | 89.81 |
| Depth-based features | | | | | | | | |
| RIFT2M [5] | 19.69 | 60.76 | 20.94 | 60.87 | 19.88 | 59.78 | 19.88 | 60.02 |
| Fehr's [6] | 23.78 | 67.34 | 24.05 | 64.95 | 20.58 | 63.21 | 20.46 | 62.65 |
| Skeleton [2] | 41.36 | 85.29 | 49.83 | **91.49** | 54.18 | 87.07 | 60.25 | **93.58** |
| Proposed | | | | | | | | |
| DVCov | 27.95 | 67.20 | 35.56 | 72.53 | 25.38 | 59.67 | 36.14 | 71.45 |
| DVCov+SKL | 34.10 | 71.00 | 46.57 | 79.23 | 27.74 | 62.28 | 45.91 | 80.42 |
| ED | 32.09 | 75.23 | 31.76 | 75.15 | 35.82 | 73.60 | 39.20 | 79.86 |
| ED+SKL | **48.75** | **90.57** | **52.30** | 90.15 | **58.65** | **94.36** | **63.29** | 91.21 |

On "TestingA", the RGB-based appearance features nearly failed, and Eigen-depth feature and skeleton-based feature outperformed them. On "TestingB", HOG and LBP can still adapt to illumination change to some extent, while color histogram completely failed. Since rotation of samples took place in this dataset, the proposed Eigen-depth feature outperformed our depth voxel covariance descriptor and is more suitable for shape description in such a situation. Eigen-depth feature also outperformed the compared rotation invariant depth shape descriptors RIFT2M and Fehr's descriptor. In most cases, combining Eigen-depth feature with skeleton-based feature worked better than using them separately. Since viewpoint of pose changed from 0° to 360° for each person in the training and testing sets, shape description from front to back for the same person changes notably and thus would cause confusion for matching. Skeleton-based feature is more effective in the cases when there are only 5 persons in testing set, because there are fewer persons of similar somatotype. So skeleton-based feature is better than Eigen-depth feature in this case. In general, the combination of Eigen-depth feature and skeleton-based feature is the most effective. The test showed the effectiveness of our depth-based method when people change clothes and appear in the extreme lighting condition.

### D. Comparison to Depth-based Re-id Frameworks

Existing well-known methods related to depth-based person re-identification include still-image-based recurrent attention model (3D RAM) [4], skeleton-based feature (SKL), Point Cloud Matching (PCM) and the combination of PCM and SKL (PCM+SKL) [3]. 3D RAM, PCM and PCM+SKL are designed under a different setting from the usual one for person re-id; that is they require that the group of persons for training is the same as the one of persons for testing, while there is no overlap on persons between training and testing in the usual re-id setting. To compare our method with the above methods, we tested our Eigen-depth (ED) feature and the combination of Eigen-depth feature and skeleton-based feature (ED+SKL) on PAVIS and IAS-Lab RGBD-ID under the same setting as the compared methods when they were reported in [3], [4]. The experiment results were reported in Table VI. As shown, our method ED and ED+SKL clearly outperformed other existing depth-based frameworks, especially on PAVIS, a much larger dataset with more persons involved.

TABLE VI
PAVIS AND IAS-LAB RGBD-ID*: RESULTS OF COMPARISONS WITH EXISTING DEPTH-BASED RE-ID FRAMEWORKS (%).

| Dataset | Probe | ED | ED+SKL | 3D RAM [4] | PCM [3] | PCM+SKL [3] | SKL [3] |
|---|---|---|---|---|---|---|---|
| PAVIS | Walking2 | 54.4 | **57.0** | 41.3 | - | - | 28.6 |
| IAS-Lab RGBD-ID | TestingA | 44.0 | **49.9** | 48.3 | 28.6 | 25.6 | 22.5 |
| | TestingB | 55.5 | **66.6** | 63.7 | 43.7 | 63.3 | 55.5 |

*The experiments here are under a different setting from the experiments in previous sections. See Sec. VI-D for details.

### E. Depth Feature Transfer Evaluation

The effectiveness of the kernelized implicit feature transfer scheme was evaluated on RGB datasets 3DPeS [21] and CAVIAR4REID [20]. Before showing the experiment results, we first present implementation details of the feature transfer scheme.

**Implementation Details**. In this work, we selected the BIWI RGBD-ID dataset [2] as the auxiliary dataset. In "Training" of BIWI RGBD-ID, there were 50 persons performing actions of rotation and walking. For each of the 50 persons in "Training", 8 RGB images from 8 different views ranged from 0° to 360° were selected as auxiliary RGB images. As for depth information, for each person, 8 point clouds from frontal view were selected for extracting depth features corresponding to those 8 RGB images. Some samples of auxiliary RGB-D dataset are shown in Figure 6.

After constructing the RGB-D auxiliary dataset, we extracted visual features and depth features to establish the connection between two modalities by the kernelized implicit feature transfer scheme. Since depth features describe body shape of pedestrians, the visual features for learning the transformation should also be able to describe body shape to some extent. We used HOG [13] and LBP [80] for describing body silhouette and textures. All RGB images in auxiliary dataset were resized to $128 \times 48$ for extracting HOG and LBP features using $8 \times 8$ cells. We also extracted the same visual features for samples in target domain. As for the point clouds, Eigen-depth feature was extracted to describe body shape.

With the extracted visual features and depth features, we conducted the proposed kernelized implicit feature transfer scheme. We chose the guassian kernel functions for visual feature and depth feature, which are $\mathrm{K}_v(\mathbf{x}, \mathbf{y}) = \exp(-\gamma_v||\mathbf{x}-\mathbf{y}||^2)$ and $\mathrm{K}_d(\mathbf{x}, \mathbf{y}) = \exp(-\gamma_d||\mathbf{x}-\mathbf{y}||^2)$, respectively. Let $dist_{vm}$ and $dist_{dm}$ denote the means of the distances of visual features and depth features between any two samples in the auxiliary dataset, respectively. We set the bandwidth parameters as $\gamma_v = \frac{1}{dist_{vm}^2}$ and $\gamma_d = \frac{1}{dist_{dm}^2}$. As for the parameters setting of the objective function, we empirically set the default parameters as $\beta' = \frac{10}{\mathrm{tr}(\mathbf{B}_{vd})}$, $\gamma_1 = \frac{10}{\mathrm{tr}(\mathbf{B}_{bd})}$, $\gamma_1' = \frac{10}{\mathrm{tr}(\mathbf{B}_{wd})}$, $\gamma_0 = \frac{1}{\mathrm{tr}(\mathbf{B}_{bv})}$, $\gamma_0' = \frac{1}{\mathrm{tr}(\mathbf{B}_{wv})}$ which were normalized by traces. That is to say, the terms related to depth features were assigned much larger weights since we focused on learning the relation between depth feature and visual feature in order to take advantage of the discriminative information in depth features. As for the dimension of the common latent subspace, we set $m = 700$.

**Score-level Feature Fusion**. We estimated depth features on RGB images in order to augment the visual features with complementary information in depth features. Let $dist_{RGB}(p_1, p_2)$ denote the distance between RGB-based appearance features of RGB images between two samples $p_1$ and $p_2$, and $dist_D(p_1, p_2)$ denote the distance between depth features computed according to Equation (25). We fused these two types of distances with a weight $\eta$ as follow:

$$dist_{fusion}(p_1, p_2) = (1-\eta)dist_{RGB}(p_1, p_2)+\eta dist_D(p_1, p_2). \tag{26}$$

In our experiments, each type of distance was normalized by its mean distance between any two samples in training set.

**Experiment Settings**. We evaluated how the transferred Eigen-depth feature (TED) can help to improve the performance when combined with LOMO [12] and ELF18 [50], which were two recently proposed effective RGB-based appearance features in person re-identification. To compute the similarity of RGB-based appearance features, we applied three favorable distance metric learning methods LFDA [43], MLAPG [52] and KISSME [40]. So we had the following different settings, ELF18(LFDA)+TED, LOMO(LFDA)+TED, ELF18(MLAPG) +TED, LOMO(MLAPG)+TED, ELF18(KISSME)+TED, LOMO(KISSME)+TED. For these settings, the corresponding default distance fusion weight $\eta$ was set to 0.3, 0.2, 0.3, 0.15, 0.3, 0.2, respectively. It is reasonable that the distance fusion weight $\eta$ was set to different values when fusing different RGB-based distance metrics with the depth one. Experiments were conducted on 3DPeS and CAVIAR4REID. We followed the experiment settings on PAVIS. For each person in testing set, one image was randomly selected as gallery and the remaining images were used for probing.

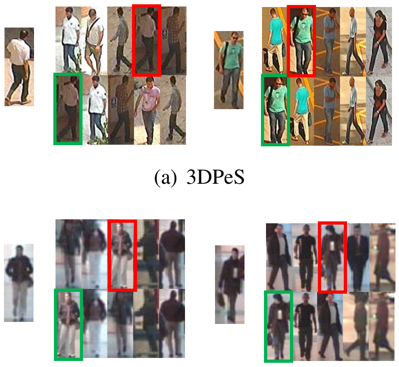
As for baseline methods, CCA [83] and sparse regression [84] were compared. In details, we used CCA to maximize the correlation between RGB feature and depth feature on the auxiliary dataset. As for sparse regression, we made the sparse representation shared between RGB and depth feature dictionaries so as to derive a transferred depth feature. The depth features transferred by CCA and sparse regression are denoted by D-CCA and D-SPA, respectively. We combined the distance of the transferred depth feature with the distance of RGB-based appearance feature for recognition. The average rank-1 to rank-5 accuracies over 10 trials were reported in Table VII.

**Results**. The transferred Eigen-depth feature (TED) can achieve rank-1 accuracy 16.0% on 3DPeS and 27.8% on CAVIAR4REID. For all RGB-based appearance features and distance metrics in our experiments, TED is effective for improving the top-rank matching accuracies. The augmentation of TED can boost rank-1 accuracy of ELF18 using LFDA metric by 4.4% on both 3DPeS and CAVIAR4REID. Although LOMO is a state-of-the-art feature for person re-identification, the transferred depth feature makes consistent improvement especially at the rank-1 matching case. Compared to the baseline methods, the proposed implicit feature transfer scheme clearly outperformed CCA (D-CCA) and sparse regression (D-SPA) when applied for the same purpose. The results indicate that it may not be effective to use CCA and sparse regression to exploit transferred depth feature. Overall, it is evident that the transferred Eigen-depth feature (TED) is complementary to RGB color and texture features, so that it can augment the RGB feature representation and help to get better ranking results.

To analyze the results more specifically, we also compare some matching samples of LOMO and LOMO+TED when using MLAPG metric in Figure 12 to see in what situations

TABLE VII
3DPES AND CAVIAR4REID: EVALUATION OF THE TRANSFERRED DEPTH FEATURE WHEN COMBINED WITH RGB-BASED APPEARANCE FEATURES USING DIFFERENT METRICS (%).

| Dataset | 3DPeS | | | | | CAVIAR4REID | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Rank | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| TED | **16.0** | **21.7** | **26.4** | **29.0** | **32.1** | **27.8** | **35.2** | **39.7** | **43.6** | **46.8** |
| D-CCA | 6.5 | 11.0 | 15.0 | 17.8 | 20.5 | 24.2 | 31.9 | 36.6 | 40.1 | 42.9 |
| D-SPA | 2.7 | 4.5 | 6.4 | 7.5 | 9.0 | 5.7 | 9.2 | 11.8 | 14.7 | 17.9 |
| LFDA metric | | | | | | | | | | |
| ELF18 | 30.3 | 40.5 | 46.4 | 51.5 | 55.3 | 32.6 | 42.9 | 49.6 | 55.2 | 59.0 |
| ELF18+TED | **34.7** | **45.2** | **51.3** | **56.5** | **60.0** | **37.0** | **45.8** | **52.0** | **56.9** | **60.8** |
| ELF18+D-CCA | 30.0 | 40.5 | 47.2 | 52.7 | 56.7 | 35.7 | 44.0 | 48.9 | 53.3 | 56.6 |
| ELF18+D-SPA | 30.3 | 40.5 | 47.1 | 51.5 | 55.6 | 32.2 | 41.9 | 47.6 | 53.0 | 57.5 |
| LOMO | 41.4 | 53.4 | 60.4 | 64.3 | 68.0 | 40.2 | 50.1 | 56.7 | 61.8 | 65.6 |
| LOMO+TED | **43.8** | **54.9** | **61.2** | **65.7** | **68.8** | **42.2** | **51.4** | **56.9** | **62.1** | **65.7** |
| LOMO+D-CCA | 41.2 | 52.8 | 59.6 | 64.0 | 67.2 | 40.9 | 49.3 | 54.9 | 59.5 | 63.3 |
| LOMO+D-SPA | 40.2 | 51.8 | 59.2 | 63.8 | 66.9 | 38.8 | 47.9 | 53.8 | 59.0 | 62.7 |
| MLAPG metric | | | | | | | | | | |
| ELF18 | 35.5 | 47.1 | 54.2 | 59.1 | 62.8 | 34.5 | 46.4 | 54.0 | 60.0 | 65.1 |
| ELF18+TED | **38.6** | **49.7** | **56.6** | **61.8** | **65.1** | **38.5** | **49.3** | **55.7** | **60.8** | **65.7** |
| ELF18+D-CCA | 33.9 | 45.9 | 52.3 | 57.5 | 61.6 | 36.9 | 46.1 | 52.1 | 56.8 | 60.3 |
| ELF18+D-SPA | 34.5 | 46.7 | 53.3 | 59.1 | 62.9 | 34.2 | 45.0 | 52.2 | 58.4 | 62.5 |
| LOMO | 47.1 | 58.5 | 64.5 | 68.5 | 71.7 | 40.6 | 51.8 | 59.4 | 65.2 | 69.4 |
| LOMO+TED | **48.4** | **58.7** | **64.6** | **68.8** | **72.0** | **42.8** | **52.9** | **59.8** | **65.3** | **69.6** |
| LOMO+D-CCA | 43.7 | 55.2 | 62.3 | 66.5 | 69.7 | 41.6 | 50.0 | 56.3 | 60.9 | 64.8 |
| LOMO+D-SPA | 44.3 | 55.8 | 62.3 | 66.5 | 69.3 | 39.0 | 48.8 | 55.3 | 60.9 | 65.2 |
| KISSME metric | | | | | | | | | | |
| ELF18 | 32.4 | 42.8 | 48.9 | 53.5 | 57.0 | 33.3 | 42.6 | 48.7 | 53.5 | 57.7 |
| ELF18+TED | **35.3** | **45.4** | **52.1** | **56.7** | **59.8** | **36.3** | **45.6** | **50.9** | **55.3** | **59.5** |
| ELF18+D-CCA | 32.6 | 42.6 | 49.7 | 54.4 | 57.8 | 35.9 | 43.7 | 48.4 | 52.8 | 56.0 |
| ELF18+D-SPA | 32.4 | 42.6 | 48.5 | 53.5 | 57.0 | 33.1 | 42.1 | 47.8 | 52.4 | 56.7 |
| LOMO | 44.3 | 54.6 | 61.3 | 65.2 | 68.7 | 42.7 | 52.7 | 59.4 | 64.3 | 69.0 |
| LOMO+TED | **45.2** | **55.3** | **61.6** | **65.8** | **69.5** | **43.3** | **53.2** | **59.9** | **64.5** | **69.0** |
| LOMO+D-CCA | 43.3 | 54.3 | 61.2 | 65.3 | 68.9 | 42.4 | 51.8 | 58.1 | 62.9 | 66.9 |
| LOMO+D-SPA | 44.1 | 54.5 | 60.7 | 65.2 | 68.9 | 41.9 | 51.3 | 57.8 | 62.9 | 67.2 |



(a) 3DPeS

(b) CAVIAR4REID

Fig. 12. Top 5 matching gallery images on 3DPeS and CAVIAR4REID. In each group of images, the probe image is on the left. The first and second rows are the matching results of LOMO and LOMO+TED when using MLAPG metric. The bounding boxes show the correct matchings.

the transferred depth features would help to get more robust matching. Among the four groups of images, the two groups on the left show that depth feature can help to find the person with similar body shape, when appearance color changes due to illumination. The other two groups on the right show that, when the appearance color and texture of two pedestrians are very similar, depth feature can help to distinguish the correct matching by body shape.

**Effects of Parameters**. To further analyze the effects of key components in our kernelized implicit feature transfer scheme, we also evaluated two significant parameters in our model, the weight $\beta'$ of the heterogeneous feature mapping term $\Omega_{vd}$ in Equation (18) and distance fusion weight $\eta$ in Equation (26). Let $\beta' = \frac{\beta}{\mathrm{tr}(\mathbf{B}_{vd})}$. We varied $\beta$ from 0 to 1000 and $\eta$ from 0 to 1 in order to see how the performance changed. When varying these two parameters, other parameters were fixed and set to default values. The rank-1 accuracies under different parameter settings were reported in Figure 13. Due to space limitation, we report the results of ELF18(LFDA)+TED and LOMO(MLAPG)+TED on 3DPeS and CAVIAR4REID, while the effects of parameters are similar under other settings.

We first analyse the effect of $\beta$. When $\beta$ is around the default value 10, the best performance is achieved. When the feature mapping term is absent (i.e. $\beta = 0$), the relation between visual features and depth features is not explored. In this case, we can observe that the improvement of TED is minor. The results indicate that the heterogeneous feature mapping term $\Omega_{vd}$ is effective for transferring more effective features by leveraging auxiliary depth information.

Then we analyze the effect of $\eta$. It can be observed that, for ELF18(LFDA)+TED, the performance is improved significantly within range from $\eta = 0.2$ to $\eta = 0.5$; for LOMO(MLAPG)+TED, the best parameter value is around 0.15. The RGB-based appearance features are more powerful in most common cases when people do not change their clothes and there is no severe illumination change, and TED can help removing ambiguities of top-ranked matchings.

### F. Runtime Performance Evaluation

We tested DVCov and Eigen-depth feature on BIWI RGBD-ID and kernelized implicit feature transfer scheme on CAVIAR4REID to compute the computational cost. In Table VIII, the time cost for extracting DVCov, Eigen-depth feature and transferring depth feature is reported.

TABLE VIII
RUNTIME PERFORMANCE OF EXTRACTING DVCOV, EIGEN-DEPTH FEATURE AND TRANSFERRING DEPTH FEATURE.

| Method | Process | Time (s) |
|---|---|---|
| DVCov | Computing normals<br>(one frame) | 0.272 |
| | Extracting DVCov<br>(one frame) | 0.038 |
| | Matching DVCov<br>(one pair of frames) | 0.007 |
| Eigen-depth | Extracting Eigen-depth<br>(one frame) | 0.010 |
| | Training LDA<br>(on training set) | 0.946 |
| | Matching Eigen-depth<br>(one pair of frames) | $3.734 \times 10^{-7}$ |
| Depth feature transfer | Computing kernel<br>(one frame) | 0.006 |
| | Training<br>(on training set) | 6.869 |
| | Matching<br>(one pair of frames) | $1.478 \times 10^{-6}$ |

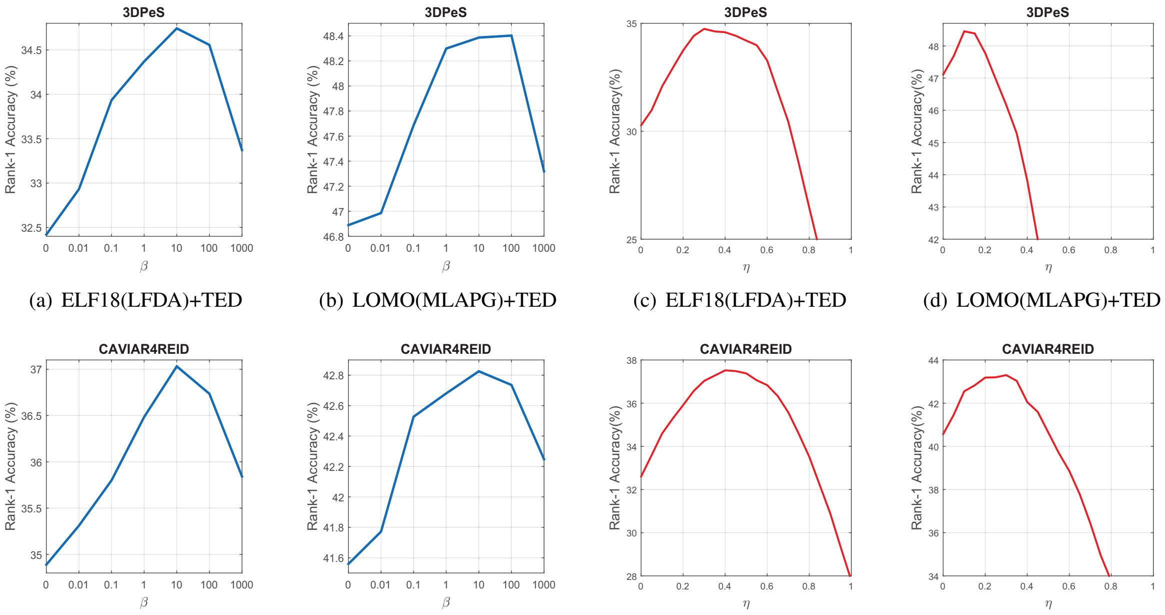
## VII. Conclusion & Discussion

We have developed a depth-based person re-identification model and addressed the bottleneck problem in person re-identification in the cases of clothing change and extreme illumination, which make most existing person re-identification models not workable. In our work, we have proposed two depth shape descriptors: the depth voxel covariance descriptor (within-voxel covariance and between-voxel covariance) and locally rotation invariant Eigen-depth feature. The local rotation invariance property of Eigen-depth feature has been proven in theory. By combining skeleton-based feature which provides complementary information to our proposed depth shape descriptors, a complete depth-based modeling of a person is formed. Extensive experiments on three RGB-D re-identification datasets show that RGB appearance-based methods suffer from clothing change and extreme illumination, while Eigen-depth feature does not and is able to describe shape and deal with rotation better than existing methods.

Since depth information is not always available, we have further proposed a kernelized implicit feature transfer scheme to estimate the Eigen-depth features from RGB images. By augmenting RGB-based appearance feature with the implicitly estimated depth feature, it is helpful for further reducing visual ambiguities for top-ranked matchings and boosting the top rank accuracy in person re-identification.

Finally, we summarize the advantages of our proposed method as follows. Firstly, our framework quantifies depth information, and this can achieve clearly better performance than applying RGB appearance information in the cases of clothing change and extreme illumination change; secondly, compared to existing 3D shape descriptors, the proposed depth-based features can better describe human body shape in depth images due to the extraction of (logarithmic) Eigen-depth feature, which is locally rotation invariant; thirdly, the proposed depth shape descriptor and the skeleton-based feature are complementary to each other, and the combination of them can provide more discriminant information of human body and thus achieve better performance.

For future development, the self-occlusion and mutual-occlusion problems remain largely unsolved in RGB-based person re-identification, and they are indeed also challenging for our proposed depth-based person re-identification model. Especially, when a large part of body is occluded, it is much harder to predict the 3D shape of a body. In such a case, the occlusion becomes a typically difficult problem, and more investigation is needed in the future. A possible way may be to consider the partial person re-identification problem as discussed in [85] for the depth-based approach.



(a) ELF18(LFDA)+TED (b) LOMO(MLAPG)+TED (c) ELF18(LFDA)+TED (d) LOMO(MLAPG)+TED

(e) ELF18(LFDA)+TED (f) LOMO(MLAPG)+TED (g) ELF18(LFDA)+TED (h) LOMO(MLAPG)+TED

Fig. 13. Effects of parameters $\beta$ and $\eta$ (shown by rank-1 accuracy). $\beta$ is the weight of heterogeneous feature mapping term $\Omega_{vd}$ and $\eta$ is the fusion weight of appearance feature and transferred Eigen-depth feature (TED). Two different combinations of appearance features and metrics ELF18(LFDA)+TED and LOMO(MLAPG)+TED on both 3DPeS and CAVIAR4REID are reported here.

## Acknowledgment

The authors would like to thank the reviewers' constructive comments.

## References

[1] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," in *European Conference on Computer Vision (ECCV) Workshop*. Springer, 2012, pp. 433–442.

[2] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. V. Gool., "One-shot person re-identification with a consumer depth camera," in *Person Re-Identification*. Springer, 2014, pp. 161–181.

[3] M. Munaro, A. Basso, A. Fossati, L. V. Gool, and E. Menegatti, "3d reconstruction of freely moving persons for re-identification with a depth sensor," in *The IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 4512–4519.

[4] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] L. J. Skelly and S. Sclaroff, "Improved feature descriptors for 3d surface matching," in *Optics East 2007*. International Society for Optics and Photonics, 2007, pp. 67 620A–67 620A.

[6] D. Fehr, A. Cherian, R. Sivalingam, S. Nickolay, V. Morellas, , and N. Papanikolopoulos, "Compact covariance descriptors in 3d point clouds for object recognition," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, May 2012, pp. 1793–1798.

[7] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *The International Conference on Pattern Recognition (ICPR)*, vol. 3, 2006, pp. 1204–1207.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 2360–2367.

[9] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: what features are important?" in *European Conference on Computer Vision (ECCV) Workshop*. Springer, 2012, pp. 391–401.

[10] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, July 2013.

[11] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision (ECCV)*. Springer, 2008, pp. 262–275.

[12] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2197–2206.

[13] O. Oreifej, R. Mehran, and M. Shah, "Human identity recognition in aerial images," in *The IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 709–716.

[14] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *British Machine Vision Conference (BMVC)*, vol. 2, no. 5, 2010, p. 6.

[15] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*, Oct 2009, pp. 322–329.

[16] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using haar-based and dcd-based signature," in *The IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2010, pp. 1–8.

[17] S. Bak and F. Brémond, "Re-identification by covariance descriptors," in *Person Re-Identification*, 2014, pp. 71–91.

[18] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *European Conference on Computer Vision (ECCV) Workshop*. Springer, 2012, pp. 413–422.

[19] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2007, pp. 1–8.

[20] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *British Machine Vision Conference (BMVC)*, vol. 1, no. 2, 2011, p. 6.

[21] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Proceedings of the joint ACM workshop on Human gesture and behavior understanding*. ACM, 2011, pp. 59–64.

[22] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2013, pp. 2528–2535.

[23] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*. Citeseer, 2015, pp. 3402–3408.

[24] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 536–551.

[25] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person reidentification with reference descriptor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 776–787, April 2016.

[26] 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 Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, Feb 2016.

[27] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 144–151.

[28] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 152–159.

[29] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3908–3916.

[30] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–8.

[31] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[32] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *the European Conference on Computer Vision (ECCV)*, Oct 2016, pp. 791–808.

[33] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 695–704.

[34] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 4516–4524.

[35] L. An, X. Chen, S. Yang, and B. Bhanu, "Sparse representation matching for person re-identification," *Information Sciences*, vol. 355, pp. 74–89, 2016.

[36] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4184–4193.

[37] 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 *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3739–3747.

[38] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *the European Conference on Computer Vision (ECCV)*, Oct 2016, pp. 475–491.

[39] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, 2005, pp. 1473–1480.

[40] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 2288–2295.

[41] W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, March 2013.

[42] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 2666–2672.

[43] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3318–3325.

[44] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3610–3617.

[45] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 1–16.

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

[47] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1629–1642, Aug 2015.

[48] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1846–1855.

[49] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1565–1573.

[50] Y. C. Chen, W. S. Zheng, J. H. Lai, and P. Yuen, "An asymmetric distance model for cross-view feature mapping in person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[51] Y. C. Chen, X. Zhu, W. S. Zheng, and J. H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[52] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3685–3693.

[53] W. S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 591–606, March 2016.

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

[55] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[56] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[57] J. Garcła, N. Martinel, C. Micheloni, and A. Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1305–1313.

[58] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *the European Conference on Computer Vision (ECCV)*, Oct 2016, pp. 405–422.

[59] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3200–3208.

[60] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[61] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised $\ell 1$ graph learning," in *the European Conference on Computer Vision (ECCV)*, Oct 2016, pp. 178–195.

[62] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti, "A feature-based approach to people re-identification using skeleton keypoints," in *The IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 5644–5651.

[63] J. Lorenzo-Navarro, M. Castrillón-Santana, and D. Hernández-Sosa, "An study on re-identification in rgb-d imagery," in *International Workshop on Ambient Assisted Living*, 2012, pp. 200–207.

[64] M. Castrillon-Santana, J. Lorenzo-Navarro, and D. Hernandez-Sosa, "People semantic description and re-identification from point cloud geometry," in *The International Conference on Pattern Recognition (ICPR)*, Aug 2014, pp. 4702–4707.

[65] D. Fehr, W. J. Beksi, D. Zermas, and N. Papanikolopoulos, "Rgb-d object classification using covariance descriptors," in *The IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 5467–5472.

[66] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multimodal person reidentification using rgb-d cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 788–799, April 2016.

[67] A. Mogelmose, C. Bahnsen, T. B. Moeslund, A. Clapes, and S. Escalera, "Tri-modal person re-identification with rgb, depth and thermal features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, June 2013, pp. 301–307.

[68] A. Mogelmose, T. B. Moeslund, and K. Nasrollahi, “Multimodal person re-identification using rgb-d sensors and a transient identification database,” in *The International Workshop on Biometrics and Forensics (IWBF)*, April 2013, pp. 1–4.

[69] V. John, G. Englebienne, and B. Krose, “Person re-identification using height-based gait in colour depth camera,” in *The IEEE International Conference on Image Processing (ICIP)*, Sept 2013, pp. 3345–3349.

[70] R. Satta, F. Pala, G. Fumera, and F. Roli, “Real-time appearance-based person re-identification over multiple kinecttm cameras.” in *VISAPP*, 2013, pp. 407–410.

[71] D. Baltieri, R. Vezzani, and R. Cucchiara, “Learning articulated body models for people re-identification,” in *Proceedings of the ACM international conference on Multimedia*. ACM, 2013, pp. 557–560.

[72] A. Albiol, A. Albiol, J. Oliver, and J. M. Mossi, “Who is who at different cameras: people re-identification using depth cameras,” *IET Computer Vision*, vol. 6, no. 5, pp. 378–387, Sept 2012.

[73] J. Oliver, A. Albiol, and A. Albiol, “3d descriptor for people re-identification,” in *The International Conference on Pattern Recognition (ICPR)*, Nov 2012, pp. 1395–1398.

[74] B. Takac, A. Catala, M. Rauterberg, and W. Chen, “People identification for domestic non-overlapping rgb-d camera networks,” in *The International Multi-Conference on Systems, Signals Devices (SSD)*, Feb 2014, pp. 1–6.

[75] X. Xu, W. Li, and D. Xu, “Distance metric learning using privileged information for face verification and person re-identification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3150–3162, Dec 2015.

[76] A. Wu, W. S. Zheng, and J. H. Lai, “Depth-based person re-identification,” in *The IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 026–030.

[77] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, *Surface reconstruction from unorganized points*. ACM, 1992, vol. 26, no. 2.

[78] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[79] W. Förstner and B. Moonen, “A metric for covariance matrices,” in *Geodesy-The Challenge of the 3rd Millennium*. Springer, 2003, pp. 299–309.

[80] Y. Zhang and S. Li, “Gabor-lbp based region covariance descriptor for person re-identification,” in *The International Conference on Image and Graphics (ICIG)*, Aug 2011, pp. 368–371.

[81] B. Schölkopf, R. Herbrich, and A. J. Smola, “A generalized representer theorem,” in *International Conference on Computational Learning Theory*, 2001, pp. 416–426.

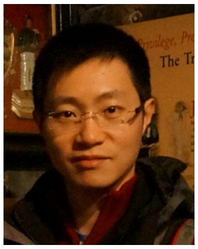
[82] H. Moon and P. J. Phillips, “Computational and performance aspects of pca-based face-recognition algorithms,” *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[83] D. Weenink, “Canonical correlation analysis,” in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, vol. 25. Citeseer, 2003, pp. 81–99.

[84] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Advances in neural information processing systems*, 2006, pp. 801–808.

[85] W. S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, “Partial person re-identification,” in *The IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 4678–4686.

**Ancong Wu** received the bachelor’s degree in intelligence science and technology from Sun Yat-Sen University in 2015. He is pursuing PhD degree with the School of Electronics and Information Technology in Sun Yat-sen University. His research interests are computer vision and machine learning. He is currently focusing on the topic of person re-identification.
Homepage: http://isee.sysu.edu.cn/∼wuancong.



**Wei-Shi Zheng** received the PhD degree in applied mathematics from Sun Yat-sen University in 2008. He is a Professor with Sun Yat-Sen University. He has been a postdoctoral researcher on the EU FP7 SAMURAI Project with Queen Mary University of London. His recent research interests include person re-identification, action/activity recognition, and large-scale machine learning algorithms. He has joined Microsoft Research Asia Young Faculty Visiting Programme. He has outstanding reviewer award in ECCV 2016. He is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of Royal Society-Newton Advanced Fellowship.
Homepage: http://isee.sysu.edu.cn/∼zhwshi.

**Jian-Huang Lai** received the PhD degree in mathematics from Sun Yat-sen University in 1999. He is a Professor of School of Data and Computer Science in Sun Yat-sen university. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications. He has published over 100 scientific papers in international journals and conferences including IEEE TPAMI, IEEE TNN, IEEE TIP, IEEE TSMC-B, PR, ICCV, CVPR, and ICDM.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a document of more experiments and analysis. Contact wuancong@mail2.sysu.edu.cn for further questions about this work.