ISSN 2277-2685
IJESR/April. 2022/ Vol-12/Issue-2/1-30

Video Person Re-Identification using True-Color and Grayscale Images

Dr.G Kishore Kumar, Mrs.I Deepika , Mrs.V Hemasree
Associate Professor[1,3] , Assistant Professor[2]
Department of CSE,
Viswam Engineering College (VISM) Madanapalle-517325 Chittoor District, Andhra Pradesh, India

Abstract

Re-identifying a missing person is a crucial step in many forensics investigations. The vast majority of currently available techniques for re-establishing the identities of missing by the use of several color-accurate cameras. Due to camera failure or specific processing for gray mode, acquired pedestrian footage may sometimes be in grayscale in practice.

Person re-identification from true-color to grayscale pedestrian recordings, which we refer to as color to gray video person re-identification (CGVPR), is required in such scenarios. However, the CGVPR issue is very difficult because to the fact that the color information that is highly crucial to depict a pedestrian is often intensity information and monochromatic in grayscale movies. We present a Semi-coupled Dictionary Pair Learning (SDPL) method based on asymmetric within-video projection to ease the pain points of CGVPR.

SDPL learns a semi-coupled mapping matrix in addition to a true-color and grayscale dictionary for use inside videos at the same time. The within-video projection matrices you've learned can reduce the file size of any video, whether it's in color or black and white. The attributes of full-color and grayscale films may be reconciled with the aid of the learned dictionary pair and the mapping matrix. We create CGVID (color and grayscale video person reidentification dataset), the first dataset of its kind dedicated to pedestrians. Each of the more than fifty thousand frames in our collection was captured in a genuine environment. Extensive assessments show that the gathered CGVID dataset is quite difficult, and it may be utilized for future study of person re-identification. Evidence from experiments demonstrates

Xiao-Yuan Jing (email: jingxy 2000@126.com) and Zhiping Peng (email: pengzp@foxmail.com) are the writers who may be reached through email.

F. Ma may be reached at mafei0603@163.com. He is affiliated with the Computer Schools at Guangdong University of Petrochemical Technology in Maoming, China; Wuhan University in Wuhan, China; and Pingdingshan University in Pingdingshan, China.

X. Zhu works at the Henan Key Laboratory of Big Data Analysis and Processing and the School of Computer and Information Engineering at Henan University in Kaifeng, 475001, China. His name appears among the others as a co-author.

Z. Tang may be reached at tang.zm@mail.njust.edu.cn or via the School of Computer Science and Engineering at Nanjing University of Science and Technology in Nanjing 210094, China.

Z. Peng may be contacted at pengzp@foxmail.com, and he works in the School of Computer at the Guangdong University of Petrochemical Technology in Maoming, China.

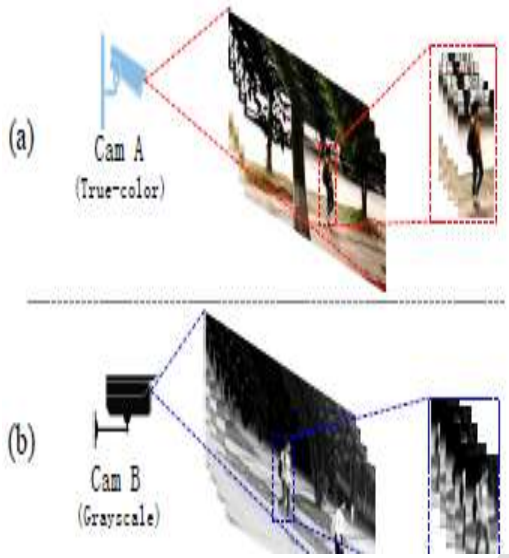


Figure 1: A common video person re-id situation in both true color and grayscale.

(a) Real-world color video shot using a color camera. b) The recorded sequences of grayscale images from Grayscale photography equipment. When comparing the images of the same individual taken with two cameras set to various settings, it is clear that there is a noticeable difference.

As compared to other approaches, ours performs better on the CGVPR challenge.

Re-identification of individuals in grayscale footage; HD video, dictionary study in full color

INTRODUCTION

PEOPLE re-identification (or "person re-id") is gaining attention in video surveillance because it may be used to link images of the same pedestrians captured by multiple cameras.

[1],[2], [3], [4]. Both feature learning-based [5, 6, 7, 8, 9, 10, 12, 13] and distance learning-based [14, 15, 16, 17, 18] approaches may be used for person re-id. Learned robust and discriminative representations from individual samples are the goal of feature learning based approaches. The goal of the distance metric learning based approaches is to discover an efficient metric for the person re-id issue that can unite images captured by several cameras.

In real life, there are situations when the camera only records in grayscale. To save memory, cameras may be switched to grayscale mode (each grayscale pixel is recorded with 8 bits, whereas each true-color pixel takes 24 bits). For real-world use, switching to grayscale mode will wipe out all of that lovely color data.

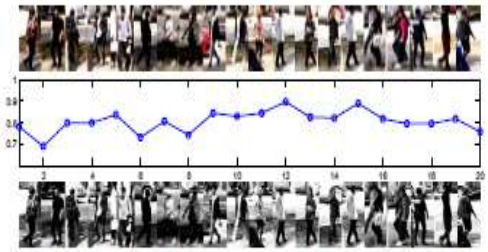


Fig. 2. The cosine similarity of true-color and grayscale images from our new dataset CGVID. *Top row:* True-color images corresponding to 20 persons captured by true-color camera. *Middle row:* the cosine similarity of true-color and grayscale images. *Bottom row:* The grayscale images are generated from the corresponding true-color images.

films of pedestrians, which will make it more challenging to match routinely acquired (true-color) videos with grayscale recordings. A normal human being is shown in Fig. 1. re-id situation between full-color and black-and-white recordings

To examine the impact of the grayscale mode, we first choose 20 photographs representing various people and, using the pixel values, compute the cosine similarity of the color and grayscale versions of each. Figure 2 displays the obtained outcomes. When comparing two photographs, a bigger cosine value indicates a higher degree of resemblance. Since the average cosine similarity between two photos is around 0.8, we may deduce that effective information is being lost in grayscale person sequences. In this article, we propose a new term, "Color to Gray Video Person Reidentification," to describe the process of matching faces in full color and grayscale pedestrian recordings (CGVPR).

I. Intrinsic Drive

Even though CGVPR has many practical uses, it has received little academic attention.

The true-color and grayscale video person re-id issue is not yet successfully solved, despite the fact that many works on person re-id have been provided for the regular circumstances [19, 20], [21]. The fundamental cause is that these techniques ignore the impact grayscale mode has on the visual appearance characteristics and the space-time features.

To combat this, we want to develop a method that mitigates the impact of grayscale mode on subsequent re-identification, an issue that affects both true-color and grayscale images of people. The literature [24], [25], [38], and [27] has inspired us to investigate the connection between the characteristics of color and monochrome films shot by various cameras. Videos come in a wide variety of both color and black and white, however there is 1 These numerous common grayscale situations occur in real life: The probe sets are grayscale while the gallery sets are in full color; ii) both the probe and gallery are gray; iii) one of the probe and gallery is a combination of true-color and grayscale movies; iv) both the probe and gallery are a combination of true-color and grayscale videos. The first scenario is the focus of this study.

True-color and grayscale movies of the same individual have inherent connections. Enhancing the heterogeneous person re-id job

Finding connections between the full-color and grayscale samples.

The heterogeneous picture issue may be effectively tackled by using the semi-coupled approach, which identifies the connection between images/videos with varying properties via the acquisition of a mapping matrix. A semi-coupled dictionary learning approach is presented in [24] for super-resolution and cross-style pictures. To find the hidden connection between the disparate data, this technique trains a dictionary pair for pictures of varying modalities and a semi-coupled mapping matrix. Instead of concentrating on the re-id process of identifying individuals, this piece explores the translation between photographs of varying fashion. Comparing samples taken at HR and LR resolution reveals strikingly distinct properties. The semi-coupled method is used for the HR and LR person re-id issue in the literature [25, 27]. To some extent, the loss of features due to poor resolution may be compensated for by using the semi-coupled mapping matrix. In light of these studies, we want to tackle the CGVPR issue using the semi-coupled method. The challenge of matching truecolor and grayscale films remains unsolved, however, since these efforts are primarily concerned with elucidating the connection between HR and LR pictures. From this study, we want to develop a semi-coupled mapping matrix that relates films with dissimilar qualities to one another. Differences between full-color and grayscale films may be minimized using the learnt mapping matrix.

We are unaware of any publically accessible person re-id databases that include both color and black-and-white footage. This paper's secondary goal is to provide a novel true-color and grayscale video person re-id dataset (CGVID) to the CGVPR research community.

Supporting Evidence, Part B

Following are three bullet points that outline the paper's contributions:

1) This is the first published study to deal with the challenge of re-identifying individuals in films shot in both full color and grayscale.

2) We provide a true-color and grayscale video person reidentification dataset (CGVID) as a baseline for CGVPR. This dataset includes 52,723 frames of 200 people captured by two cameras. The CGVID system captures 26,827 frames in full color for each of 200 subjects, and another 25,896 in grayscale. The generic

CGVID's capacity to detect pedestrians in three different video-based datasets.

Thirdly, we provide a solution to the CGVPR issue via semi-coupled dictionary pair learning (SDPL) method, which learns a mapping matrix and dictionaries for both color and monochrome videos at the same time.

Learned grayscale projection matrices may make grayscale videos more compact, while true-color projection matrices can help smooth out the bumps in between scenes. Both full-color and monochrome videos are accurately represented by the dictionary pair that was learnt. When comparing truecolor and grayscale movies, the semi-coupled mapping matrix might help you out.

Here's how the remainder of the paper is laid out: Section II provides a summary of the literature on the subject of person re-id databases and techniques. In Part III, we provide the specifications of our reference true-color and grayscale pedestrian dataset.

Our methodology is described in further detail in Section IV. In Section V, we break down the optimization process that underpins our methodology. We provide the experimental setup and methodology assessments in Section VI. In Section VII, we sum up our findings and draw conclusions.

2. CONNECTED DOCUMENTS

Here, we provide an overview of the efforts that have already been done on the topic of person re-id datasets and video-based person re-id techniques.

Reidentification Datasets, Type A

To better understand the existing person re-id datasets, we may classify them as either: image-based person datasets, i. These databases include either a single picture or many photographs of the same person taken with various cameras; ii) video-based person datasets. Each individual is represented by many, sequential frame sequences from various cameras in these databases.

databases for visual person re-identification. Many image-based datasets [29, 30], [33] are now in use for the person re-id issue.

The VIPeR [29] dataset was created in 2007 and contains 1264 photos of 632 people taken from a wide variety of angles and lighting situations. To achieve a total width and height of 12848 pixels, each picture has been resized. A total of 971 individuals were recorded in 2012 by two separate cameras located across the CUHK01 [30] campus. Two pictures exist for each individual. In 2013, ten cameras at the Chinese University of Hong Kong (CUHK02 [31]) captured 1816 people and 7264 photos on campus. Two photographs of each individual may be found in each camera.

With an average of 4.8 photos per camera, CUHK03 [32] has 13,164 images of 1,360 pedestrians captured by two separate cameras. The Market1501 [33] dataset is massive, including 32,668 pictures.

out of 1,501 subjects caught on 6 cameras. Six cameras are used in the SYSUMM01 dataset [34], which serves as a standard for cross-modality person re-id.

There are a total of 491 people in the database, shown throughout 287,628 RGB photos and 15,792 IR (Near-infrared) photographs. For the SYSU-MM01 dataset, there are a total of six cameras available (four RGB and two IR). Large discrepancies between the two modalities make this dataset particularly difficult to work with. Using an IR camera in a low-light setting opens up novel possibilities for person re-id in a tense setting.

The static photos used in these re-id databases are stuffed with details about the subjects' physical appearance. They are unable, however, to provide data on mobility.

Databases for the automatic recognition of people in videos. For the purpose of person re-id, various video-based datasets have been released as of late [22], [23], [44], [36]. These datasets often include several images of each individual, making it possible to get valuable spatiotemporal data that would otherwise be impossible to extract from a single, static photograph. There are 148 individual video clips that make up ETHZ [22]. The average number of frames in a sequence is 56, although the number of frames in any one sequence might be anything from 6 to 356. There are a total of 1134 individuals in PRID 2011 [23], and 200 of them have been captured on video by two separate, stationary security cameras. You may choose between an image-based version of PRID 2011 or a video-based version. The durations of frame sequences in the video version dataset vary widely, from 5 to 675 frames, with an average of 117 frames captured by Camera A and from 5 to 179 frames captured by Camera B. All photos are shot using high-quality color cameras, and most sequences of individual frames have uncomplicated backgrounds and foregrounds.

Background of iLIDS-VID [44] is complicated, since it was filmed in an airport's arrivals hall with a CCTV system.

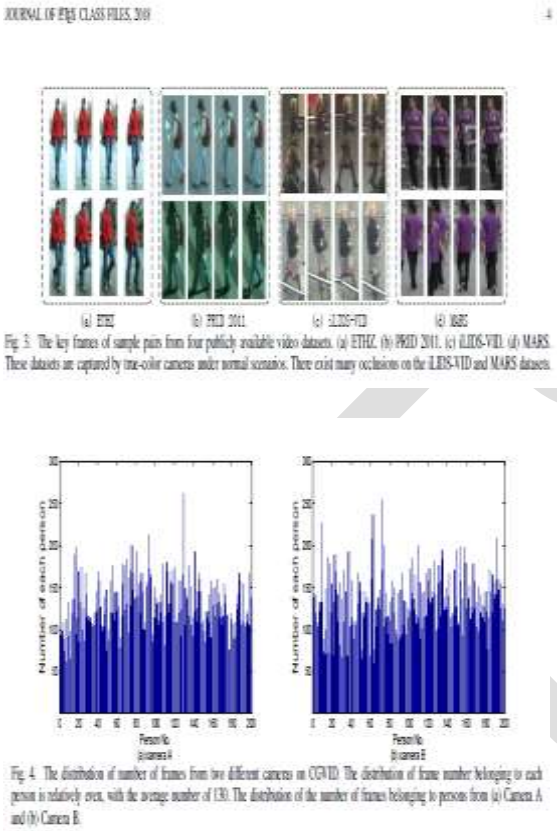
Each picture sequence has an arbitrary length, averaging 73 frames long. The range is 23-192. The iLIDSVID dataset is more difficult than other person reid datasets due to similarities in clothing, lighting, perspective, and crowded backgrounds and occlusions. Space-time features have been extracted from both the PRID 2011 and iLIDS-VID datasets.

Six cameras at Tsinghua University collect the MARS (Motion Analysis and Re-identification Set) [36], which includes 1,261 individuals and over 20,000 tracklets. In addition, MARS includes 500,000 incorrect detection results as distractor bounding boxes.

Different modalities of human datasets have been available in recent years. We show two useful databases for re-identification purposes. In 2014, RGB and depth cameras were used to compile the RGB-D dataset BIWI RGBD-ID [37]. Only 50 unique people are included in the 50 training and 56 test sessions.

Synchronized RGB pictures, depth images, segmentation maps, and skeletal data are all part of the collection.

The RGB-D IAS-Lab RGBD-ID dataset [39] includes 11 training sequences and 22 testing sequences.



individual differences. In the training and testing-A sequences, participants wear unique outfits in each room, while in the testing-B set, participants wear the same outfits in both rooms.

Multi-modal datasets take into account more information, which is useful for learning stable features.

Many different types of person re-id datasets are already available, as was mentioned above. Despite this, there are not many video-based person re-id datasets available. The majority of current image/video-based person re-id datasets are captured by true-color cameras, which typically have three color channels.

B. Person Re-Identification Techniques That Use Multiple Technologies

Several approaches to the cross-modality person re-id problem have emerged in recent years. A multi-model uniform deep learning (MMUDL) [26] method is introduced to address the issue of person reidentification in an RGB-D scenario.

Effective anthropometric features are extracted from depth images using the deep network in MMUDL. This allows MMUDL to make use of the data contained in three-channel depth images. This paper marks the first time a deep neural network has been trained to perform person re-id using depth images. Two models, semi-coupled low-rank discriminant dictionary learning (SLD2L) [25] and discriminative semi-coupled projective dictionary learning (DSPDL) [27], have been proposed to deal with the high resolution (HR) and low resolution (LR) person re-id problem. Dictionary pair and mapping matrices can be learned by SLD2L using HR and LR features.

Images used for training purposes. The semi-coupled mapping can bridge \sthe gap between HR and LR pedestrian images. The semi-coupled mapping method is used in these works for samples with varying characteristics and a concentration on the LR image-based person re-id problem.

A deep zero-padding (DeepZero) method [34] is introduced to train a single-stream network with the goal of automatically evolving domain-specific nodes in the network for the cross-modality problem, which is of particular importance when attempting to solve the RGB-IR person re-id problem. DeepZero provides a versatile and unique cross-modality model alternative. A novel cross-modality generative adversarial network, cmGAN [35], is introduced to take on the cross-modality problem. cmGAN utilizes \sa cutting-edge generative adversarial training based \sdiscriminator to learn discriminative feature representation \sfrom different modalities. cmGAN can map crossmodality \ssamples into one common subspace.

Multi-modality person re-id is a common problem in \sreal-world scene. Even though these techniques can address the cross-modality issue, they are primarily concerned with image-based person re-id tasks.

C. Person Re-Identification Techniques Using Video

A handful of new techniques for re-identifying individuals from video footage have emerged recently. These techniques can be broken down into two groups: those that are based on metric learning, and those that are based on feature learning. Wang et al. [44], [45] \sintroduced FEP (Flow Energy Profiling) to segment the

DVR (Discriminative selection in video ranking) uses a person's walking patterns to mechanically choose the most discriminative video clips from a set.

still image sequences. Using the FEP as a foundation, STFV3D [13] is introduced as a means to extract 3D space-time characteristics from individual walking-cycle pieces. In order to find a solution to the temporal alignment issue, STFV3D takes into account walking cycles and body-action units. For the video-based person re-id issue, we develop SI2DL [21], which is capable of learning both an intra-video projective matrix and an inter-video distance metric from the spacetime feature sets concurrently. By projecting one video onto another, we may reduce the number of features needed to identify a given clip, which speeds up the re-identification process. Using a top-push restriction, TDL [20] is developed to address the problem of inter-class variation in individual films. The top-push restriction may both shorten the distance between well matched samples and increase the distance between incorrectly matched samples.

In order to derive a deep feature vector from a video set's visual appearance and optical flow information, McLaughlin et al. [11] established a convolutional neural network (CNN) and recurrent neural network (RNN) architecture.

This is the first study to use CNN and RNN to try the person re-id job using video. In order to fully use the spatial and temporal attentive information in movies, Zhou et al. [9] introduced an end-to-end deep neural network architecture for video-based person re-id by concurrently learning features and metrics. Xu et al. [48] developed a combined attentive spatial and temporal pooling network (ASTPN) for video-based person re-identification, which is based on CNN and RNN architecture. Spatial pooling allows ASTPN to pick out

interesting parts of each frame, whereas attention temporal pooling helps it pick out the most informative frames over the whole sequence.

While the aforementioned techniques are useful for video-based person re-id, they are not well-suited to the CGVPR issue since they are developed for matching between true-color films without taking the effects of grayscale into account.

AN EXPLANATION OF THE DATASET (III)

In this study, we include a new video person re-id dataset in both true color (three color channels) and grayscale (one color channel) (CGVID). This section provides an overview of the recently amassed CGVID dataset.

First, let's look into A. CGVID

Our CGVID and its collecting methodology are described here.

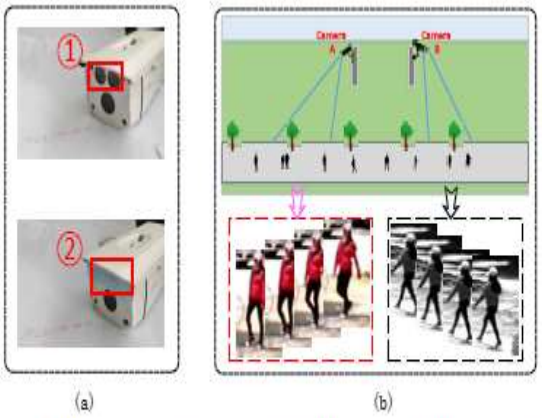


Fig. 5. Illustration of collecting the CGVID dataset. (a) IP camera with monochrome and true-color modes. (b) The illustration of cameras' layout. Two disjoint cameras are monochrome and true-color for collecting CGVID, respectively. We set camera A in true-color mode and camera B in monochrome mode by covering Near-Infrared lamp.

1) The gathering procedure. The data for CGVID was gathered at Wuhan University. The following are the procedures that were followed to compile the CGVID dataset.

Method of obtaining data: we watch recordings of people going down the street from two cameras set up in opposite directions. Fig. 5 (a) shows the Dahua IP camera's (DH-IPCHDW3302B) monochrome camera lens recording grayscale video, whereas Fig. 5 (b) depicts the Dahua IP camera's (DH-IPCHDW3302B) color mode recording true-color video. Cameras A and B both have a resolution of 1280 x 720 pixels. By shielding the IP camera's Near-Infrared (NIR) light source (shown in Fig. 5(a)1, 2), we were able to capture movies in grayscale. A frame at a rate of 30 fps is achieved. The arrangement of the cameras, shown in Fig. 5 (b), allows for many angles of the same persons to be recorded.

As part of the person-extraction process, footage of pedestrians is transformed into frame sequences. We use custom-built software1 to semi-automatically extract bounding boxes of people in each frame, which works well in challenging conditions including changing lighting, busy backgrounds, and occlusion. Finally, the recovered images are resized to uniformity, and human identities are encoded into the picture sequences.

The second option is a look at our data set. The 52,723 photos in CGVID were taken by two separate cameras and do not overlap. There are anything from 58 to 262 frames in each video, with an average of 130 frames per person per camera. Each individual's unique statistical frame number is shown in Fig. 4. From the results of the two cameras, we can see that the amount of frames captured for each individual is distributed quite evenly. There is sufficient space-time detail in only 58 frames of video, which is more than enough to capture two full walking cycles. Many people are hidden by oncoming traffic, trees, or vehicles. All

TABLE I
BASIC STATISTICS OF CGVID AND EXISTING REPRESENTATIVE VIDEO PERSON RE-ID DATASETS.

Dataset	People	Total Frames	Average length	Min/Max length	Occlusion degree	Year	Cameras	Label	Color Model
ETHZ	146	8,580	56	6/556	Few	2007	1	manual	true-color
PRID 2011	200	40,035	73	5/675	Few	2011	2	manual	true-color
iLIDS-VID	300	42,459	100	23/192	Partial	2014	2	manual	true-color
MARS	1261	1,191,003	95	30/490	Partial	2016	6	auto	true-color
CGVID(Ours)	200	52,725	130	58/262	Partial	2018	2	manual	true-color/grayscale

Fig. 6. The key frames of sample pairs from the CGVID dataset. The top row belongs to camera with true color and the bottom row comes from the grayscale camera. The samples in the dotted-line rectangle belong to the same person. Many people are occluded by cars, other pedestrians.

After separating the human from each video, the visual sequence is normalized to 64 by 128 pixels. A new CGVID dataset may be accessed at URL2.

B. Analyzing Known Video Datasets

Table I displays some descriptive metrics for the five aforementioned video person reid datasets as well as our own CGVID. Table I shows that the frame number distribution of several existing video datasets is significantly diverse, and that some pedestrians have fewer than 10 frames in the ETHZ and PRID 2011 datasets. Existing video-based person re-id datasets are recorded under standard conditions, such as true-color mode cameras in well-lit settings. The ETHZ and PRID 2011 datasets have minimal occlusions. As far as video-based person re-id datasets go, MARS is by far the biggest. The MARS dataset has a high number of spurious positive detections. Tracking tasks include those that attempt to identify and remove people from films [40, 41, 42, 43].

The DPM detector and GMMCP tracker [36] automatically create MARS, unlike other current video datasets are created by humans. Our dataset has longer average and minimum lengths compared to other available datasets. More walking cycles [44, 13] and other space-time details may be found in the longer frame sequences.

Some couples from the CGVID dataset are shown as examples in Fig. 6. Here is a quick rundown of our dataset and its features:

There are noticeable discrepancies between sample pairs captured by the two cameras. Color data from Fig. 6's (top row) samples taken with camera A

camera A's true-color mode yields a greater quantity of usable samples than camera B's grayscale mode.

2) There is a veil around our dataset's samples.

Figure 6 shows how challenging it may be for a human to re-identify commonplace items like trees, automobiles, and other pedestrians.

3. Since the films are collected by two notably diverse angles, the perspectives of most people in our sample are distinct. You can see a person's front side in one camera and their rear side in another. Meanwhile, pedestrians may be obscured by an object like a purse or an umbrella.

4) Everyone in our sample population is out and about on the road.

Videos of moving people, as opposed to those of people who are standing or sitting stationary, may be used to more accurately gauge the passage of time. From these image sequences, we can reliably extract whole walking cycles. More than two walking cycles are included in each video. About 20 frames make up a single walking cycle.

C Observational Procedures for Assessment

When assessing the efficacy of person re-id techniques, we use a standard Cumulated Matching Characteristics (CMC) [29] curve. We compare the state-of-the-art person re-id approaches against the standard space-time feature STFV3D and two sample deeply-learned features RNNCNN and PCB to learn more about the robustness and generalizability of our new dataset. We divided the dataset into two equal subsets (100 people each) and did trials.

A. Error Term in Video Reconstruction

Let's pretend that in the training set, samples from camera A are full color and those from camera B are monochrome.

The feature sets for the color and black-and-white training videos are denoted by A = [A1;A2;:::;Ai;:::;AN] and B = [B1;B2;:::Bi;:::;BN]. In this case, N is the total amount of data used for training. The feature set of the ith video, Ai = [ai 1;:::; ai j ;:::; ai ni], is represented by the equation Ai = [ai 1;:::; ai j ;:::; ai ni]. The ai j 2 Rd element stands for the characteristic of the jth walking cycle that applies to the ith individual. For the ith individual, ni is the total number of walking cycles. The ith video's feature set, Bi = [bi 1;:::; bi j ;:::; bi ni], is denoted as Bi 2 Rdni. Each individual's unique walking pattern is encoded in a feature vector, where bi j 2 Rd is the feature vector for the ith person's jth walking pattern. D stands for "video features dimension." The acquisition of dictionaries is a powerful tool for improving sample representation [49, 50]. As a matter of common sense, we can pick up the necessary knowledge to acquire a dictionary pair for video representation. Let's say that the lexicons DC and DG represent the color and monochrome video characteristics captured by cameras A and B, respectively. Consider two coding coefficient matrices, A over DC and B over DG, and label them X and Y. The error term in video reconstruction may thus be defined as follows.

$$E_{rep} = \|W^T A - D_C X\|_F^2 + \|V^T B - D_G Y\|_F^2, \quad (1)$$

where Wand V are the true-color and grayscale samples' asymmetric within-video projection matrices. In actuality, there are often substantial differences,

Individually challenging aspects of finding a matched pair include occlusion by other objects in each person's feature set. W and V are asymmetric projection matrices inside the video that attempt to reduce individual differences in their feature set. Thus, we create the following asymmetrical word for video projection inside the movie:

$$E_{proj} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left\|W^T(a_j^i-\mu^i)\right\|_2^2+\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left\|V^T(b_j^i-\mu^i)\right\|_2^2, \quad (2)$$

Semi-coupled mapping term

Grayscale-to-true-color sample matching is a notoriously difficult issue in several fields.

the useful software for monitoring large areas using cameras. It seems to reason that by illuminating the connections between full-color and black-and-white footage, we may bring down the differences between the two. To solve the problem of heterogeneity, the semi-coupled mapping method has been effectively used in the fields of photo-sketch synthesis and recognition [24], [27]. This motivates us to study the semi-coupled mapping, which allows us to close the gap between the coding coefficients of full-color and monochrome films. Let's pretend that the footage from camera A is in full color and the footage from camera B is monochrome. The characteristics of the actual grayscale video captured by camera B are brought closer to the features of the true color video captured by camera A of the same person thanks to the learnt mapping matrix. To see the whole picture, check out Fig. 7. For this reason, we implement the following mapping function:

$$E_{mapping} = \|X - PY\|_F^2, \quad (3)$$

where P is the mapping matrix between true color and grayscale video characteristics that is semi-coupled. Using the trained mapping P, we find that the drop

The loss of data due to grayscale may be partially remedied.

Discerning Loyalty Term C

Given that we want to use our model for person re-identification across color and black-and-white movies of people, it is essential that the feature data it processes be highly discriminable. In order to bring the same individuals from multiple cameras closer together, while keeping the different people from other cameras apart, we may devise a discriminative fidelity term. To explain what is meant by "discriminative faithfulness," we might say the following:

$$E_{disc} = \frac{1}{|S|}\sum_{<i,j>\in S}\left\|X^i-PY^j\right\|_F^2-\frac{\beta}{|D|}\sum_{<i,j>\in D}\left\|X^i-PY^j\right\|_F^2, \quad (4)$$

where the ith element and the jth element both belong to the same individual (I j >2 S). Separate ith and jth elements are indicated by I j >2 D. S

and D represent collections of duplicate samples and unique samples, respectively. Set size is denoted by the notation jj. is a variable used to fine-tune a system. Using the acquired mapping P, camera B's grayscale feature may approach the true color feature of camera A.

D. Purpose and SDPL

Within-video projection, semi-coupled mapping, and the inaccuracy in video reconstruction are all taken into account.

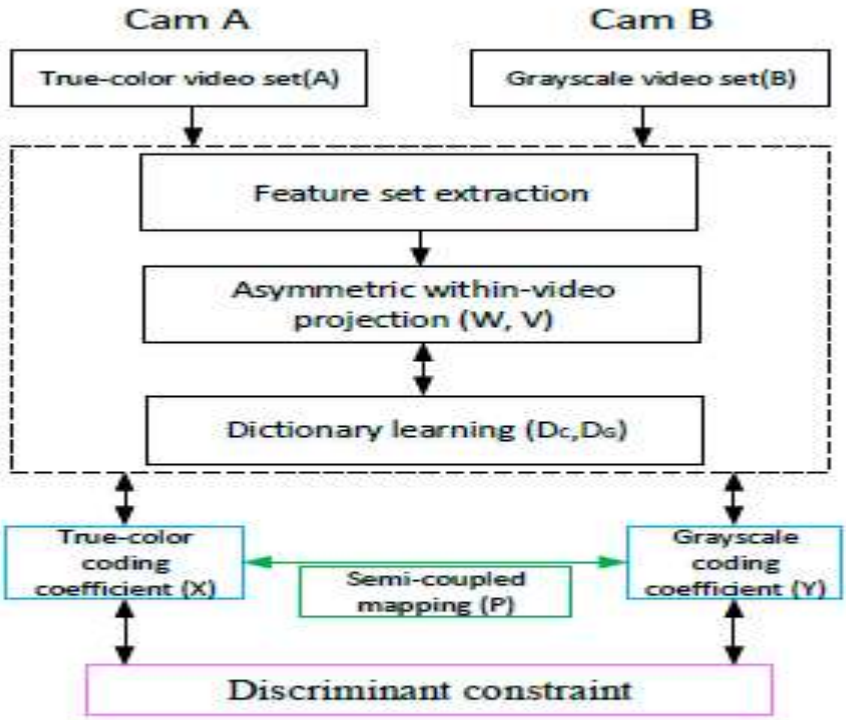


Fig. 7. Illustration of our SDPL approach. First, we extract the space-time features from the training samples, and then learn the within-video feature projection matrices (W, V) respectively. Then, SDPL updates the dictionary pair (D_C, D_G), the corresponding to coefficients (X, Y) and the semi-coupled mapping matrix (P). Finally, we update the coefficients with the discriminant constrains.

design the following objective function

design the following objective function:

$$\min_{\substack{D_C,D_G\\P,W,V}} E_{rep} + \rho_1 E_{mapping} + \rho_2 E_{proj} + \alpha E_{disc} + \lambda E_{reg}$$
$$s.t. \quad \|d_C^i\|_2^2 \leq 1, \|d_G^i\|_2^2 \leq 1, \forall i, \tag{5}$$

where α and λ are balancing factors. ρ_1 and ρ_2 separately control the effects of within-video projection matrices and semi-coupled mapping, and are empirically set to $\frac{1}{N}$. $E_{reg} = \|W\|_F^2 + \|V\|_F^2 + \|P\|_F^2 + \|X\|_F^2 + \|Y\|_F^2$ is the regularization term, which can regularize the coding coefficients, the within-video projection matrices,

min

Part V: Improving SDPL

Here, we'll go out the optimization process in further depth.

Despite the absence of any theoretical assurance

Hence if all other variables are held constant, the objective function (5) is convex with regard to (W; V; P;DC;DG). As a result, we optimize the objective function via an iterative approach.

The objective function of SDPL may be decomposed into four sub-problems, namely, updating coding coefficients, updating dictionary pairs, learning within-video projection matrices, and updating the mapping matrix of representation coefficients, all with the goal of minimizing the energy in (5).

A. Revising the W and V Projection Matrix

The objective function (5) may be expressed as when all other factors except W and V are held constant.

follows:

$$\min_W \|W^T A - D_C X\|_F^2 + \rho_2 \sum_{i=1}^{N} \sum_{j=1}^{n_i} \|W^T (a_j^i - \mu^i)\|_2^2 + \lambda \|W\|_F^2, \quad (6)$$

$$\min_V \|V^T B - D_G Y\|_F^2 + \rho_2 \sum_{i=1}^{N} \sum_{j=1}^{n_i} \|V^T (b_j^i - \mu^i)\|_2^2 + \lambda \|V\|_F^2. \quad (7)$$

By setting the derivative with respect to W, we have the solution of (6) as follow:

$$W = (AA^T + \rho_2 \sum_{i=1}^{N} \sum_{j=1}^{n_i} (a_j^i - \mu^i)(a_j^i - \mu^i)^T + \lambda I)^{-1} A X^T D_C^T, \quad (8)$$

where I is an identity matrix. The solution of V is similar to W. By setting the derivative with respect to V, we have the solution of (7) as follow:

$$V = (BB^T + \rho_2 \sum_{i=1}^{N} \sum_{j=1}^{n_i} (b_j^i - \mu^i)(b_j^i - \mu^i)^T + \lambda I)^{-1} B Y^T D_G^T. \quad (9)$$

B. Updating Coding Coefficients X and Y

By fixing the other variables, the objective function with respect to the coefficient X can be written as:

$$\min_{X^i} \|W^T A - D_C X\|_F^2 + \rho_1 \|X - PY\|_F^2 + \alpha\left(\frac{1}{|\mathcal{S}|} \sum_{<i,j> \in \mathcal{S}} \|X^i - PY^j\|_F^2 - \frac{\beta}{|\mathcal{D}|} \sum_{<i,j> \in \mathcal{D}} \|X^i - PY^j\|_F^2\right) + \lambda \|X\|_F^2. \quad (10)$$

]

The solution of (10) can be obtained by setting the derivative with respect to X^i to zero. The solution is as:

$$X^i=(D_C^T D_C+(\lambda+\rho_1+\frac{\alpha}{n_i}(1-\beta))I)^{-1}$$
$$(D_C^T W^T A^i+\rho_1 P Y^i+\alpha\rho_1 P(Y^i-\frac{\beta}{|N-1|}\sum_{j=1,j\neq i}^{N-1} Y^j)). \quad (11)$$

Similar to the solution of X, the objective function with respect to Y^i can be written as:

$$\min_{Y^i}\|V^T B-D_G Y\|_F^2+\rho_1\|X-PY\|_F^2$$
$$+\alpha(\frac{1}{|\mathcal{S}|}\sum_{<i,j>\in\mathcal{S}}\|X^i-PY^j\|_F^2-\frac{\beta}{|\mathcal{D}|}\sum_{<i,j>\in\mathcal{D}}\|X^i-PY^j\|_F^2)$$
$$+\lambda\|Y\|_F^2. \quad (12)$$

By setting the derivative with respect to Y^i to zero, we can get the solution of Y^i:

$$Y^i=(D_G^T D_G+\frac{\alpha}{n_i}(1-\beta)P^T P+(\rho_1+\lambda)I)^{-1}$$
$$(D_G^T V^T B^i+\alpha\rho_1 P(X^i-\frac{\beta}{|N-1|}\sum_{j=1,j\neq i}^{N-1} X^j)). \quad (13)$$

C. *Updating Dictionaries D_C and D_G*

When updating D_C and D_G, the other variables are fixed. The objective function with respect to D_C and D_G can be separately written as:

$$\min_{D_C}\|W^T A-D_C X\|_F^2,\quad s.t.\left\|d_C^i\right\|_2^2\leq 1,\forall i, \quad (14)$$

$$\min_{D_G}\|V^T B-D_G Y\|_F^2,\quad s.t.\|d_G^i\|_2^2\leq 1,\forall i. \quad (15)$$

We can get the solutions of (14) and (15) by using ADMM algorithm similar to [28].

D. Updating Mapping Matrix P

With the other variables fixed, we can get P as follows:

$$\min_P \rho_1 \|X - PY\|_F^2 + \frac{\alpha}{|\mathcal{S}|} \sum_{<i,j>\in\mathcal{S}} \|X^i - PY^j\|_F^2 - \frac{\alpha\beta}{|\mathcal{D}|} \sum_{<i,j>\in\mathcal{D}} \|X^i - PY^j\|_F^2 + \lambda\|P\|_F^2. \tag{16}$$

By setting the derivative with respect to P to zero, we can get the solution:

$$P = \left(\frac{\alpha}{|\mathcal{S}|} \sum_{<i,j>\in\mathcal{S}} X^i Y^{jT} - \frac{\alpha\beta}{|\mathcal{D}|} \sum_{<i,j>\in\mathcal{D}} X^i Y^{jT} + \rho_1 X^b Y^{iT}\right) \left(\frac{1}{|\mathcal{S}|} \sum_{<i,j>\in\mathcal{S}} Y^b Y^{jT} - \frac{\beta}{|\mathcal{D}|} \sum_{<i,j>\in\mathcal{D}} Y^i Y^{jT} + \lambda I + \rho_1 Y^b Y^{iT}\right)^{-1}. \tag{17}$$

The optimization procedure of our approach is summarized in Algorithm 1.

Algorithm 1 The Optimization Procedure of SDPL

Input: The true-color and grayscale space-time feature sets A and B.

Initialization Initialize D_C, D_G, P, W and V; Parameters α, β, λ, ρ_1 and ρ_2.

For each iteration Until convergence:

1. Fix other variables, update W and V by (6) and (7), respectively;

2. Fix other variables, update X and Y according to (11) and (13), respectively.

3. Fix other variables, update D_C and D_G by (14) and (15), respectively;

4. Fix other variables, update the mapping P by (17).

Output: Dictionary pair D_C and D_G, the semi-coupled mapping matrix P, the within-video projection matrices W and V.

E. Time Complexity

In the training phase of our model, the computational cost is proportional to the size of dictionary pair and the dimension of the within-video projection matrices. The time complexity of initializing the within-video projection matrix W (V) is $O(N^2)$. Updating the coefficient X (Y) takes $O(s^2p+s^3+sp^2+N(s^2+sp))$, where s is the size of dictionary. N is the total number of samples in A or B. p is the dimension of samples. The dictionary size s is smaller than p. In each iteration, the time complexity of updating the true-color dictionary D_C (grayscale dictionary D_G) is $O(p^2N+qNs+s^2N+s^3+qs^2)$. Updating W (V) costs $O(p^2N+p^3+Nsp+p^2s)$ in each iteration.

F. Matching

With the learned dictionary pair D_C and D_G, the mapping P, the within-video projection matrices W and V, we can get robust and effective representations for the test videos. Let F be the features of a grayscale probe video, and C be the features of the true-color gallery videos. We perform the matching as follows:

1) With the learned P, W and V, encoding the representation coefficient f of the probe video over grayscale dictionary D_G by solving (12) as

$$f=(D_G^TD_G+(\lambda+\rho_1)I+\alpha(1-\beta)P^TP)^{-1}D_G^TV^TF. \quad (18)$$

2) Encoding the representation coefficients g of the gallery videos over true-color dictionary D_C by solving (10), whose solution is derived as

$$g=(D_C^TD_C+(\lambda+\rho_1+\alpha(1-\beta))I)^{-1}D_C^TW^TC. \quad (19)$$

3) Re-identifying the probe video in gallery videos: with the obtained representation coefficients, we compute the distance between the probe and gallery sets.

Finally, we sort the distances in ascending order, and the gallery video with the smallest distance is the truly matching for the probe video.

Algorithm 2 Re-identification

1. **Learning within-video projection matrices**
   Learning the within-video projection matrices W and V from the training sets by Algorithm 1, respectively;
2. **Learning dictionary pair**
   Learning the true-color dictionary D_C and the grayscale dictionary D_G from the training sets by Algorithm 1, respectively;
3. **Re-identification**
   Computing the distance between the probe and gallery sets, and then sorting the distance in ascending order, finally the gallery video with the smallest distance is the truly matching for the probe video.

VI. EXPERIMENTS

A. Settings

Evaluation Settings: In experiments, we follow the evaluation protocol in [44]. Specifically, we randomly split the dataset into two subsets with equal persons, one for training and one for testing. We conduct 10 random splits and show the cumulative matching characteristic (CMC) curves.

Feature Extraction: In experiments, we evaluate our approach and dataset CGVID by using two typical categories of features, including video-based feature STFV3D [13] and deeply-learned feature PCB [46].

Each pedestrian video has many walking cycles that may be used to better match the space-time characteristics.

Spatial and temporal characteristics of walking cycles are determined.

by the creators of [13]'s STFV3D. Based on the attention model and resNet50 [47], the usual deep-learned feature PCB2 has improved matching rates on various publicly accessible person reid datasets. We take PCB characteristics from both the training set and the test set, and retrain them using the competing approaches. At this point, we put the characteristics of the testing set through a review.

Adjusting the Settings: Five values,,,, 1, and 2, make up the variables in our model. For testing purposes, these values are often manipulated to = 0:04, = 0:06, and = 0:2 on CGVID. The values for and are determined experimentally to be 1 N, where N is the total number of training samples. Parameters for our SDPL are determined with the use of 5-fold cross validation on the training data.

A Comparison of Approaches: We have compared SDPL to a number of other methods in order to gauge its efficacy: I several video-based and dictionary-based person re-id methods, such as STFV3D [13], TDL [20],

KISSME [51], XQDA [52], SI2DL [21], JDML[53] (codes provided by authors); ii) several typical deep learning methods, such as RNNCNN [11], ASTPN [48], We introduce DeepZero, a cross-modality image-based person reid system. The foundation of PCB is the attention model, which is in turn built on resNet50.

The Siamese network architecture is the foundation of both RNNCNN and ASTPN.

Analyses and Results from Experiments

We provide a reference full-color and grayscale movie to hasten the study of the CGVPR topic.

dataset CGVID, and offer forth the idea of using SDPL to deal with the CGVPR issue. From CGVID, we derive two types of features: the effective spacetime feature STFV3D, and the usual deep features like RNNCNN and PCB. Experiments show that RNNCNN provides little benefit over approaches based on either metric learning or dictionary learning. We use STFV3D and PCB in our studies because of this.

Table II and Figure 8 display the data from the experiments. Our freshly gathered video dataset, CGVID, shows two things: I the findings are rather consistent across various person re-id approaches; and 2)) our SDPL obtains superior matching rates than the competing methods with different categories of characteristics. When compared to the best competitive approach, JDML, SDPL's Rank-1 matching rate improves by 3.2%(=23.3%-20.1%) with the space-time feature STFV3D and by 2.0%(=49.9%-47.1%) with the deeply-learned feature PCB.

(Gallery set is true color, while the probe is grayscale) 2.6%(=22.1%-18.5%) with STFV3D and 2.3%(=48.2%-45.9%) with PCB (Gallery set is grayscale, while the probe is color)

are the actual colors. Our method offers three major benefits: The semi-coupled mapping approach is used by SDPL to mitigate the impact of grayscale video.

ii) Our model's asymmetric within-video projection matrices may condense the individual's collection of features. In a complicated setting, true-color and grayscale characteristics benefit from the discriminant dictionary pair.

Most current approaches find it challenging to bridge the gap between grayscale and full color movies due to the former's lower information density compared to the latter. Conventional features (STFV3D) have rank-1 matching rates below 23%, while deep features (PCB) have rates below 50%. Therefore, I person re-id between the true-color and grayscale films is much more difficult than that under normal situation; ii) our new dataset CGVID comprises of more than 50K frames and is ideal for deep-learning approaches.

3. C. Deeper Dissection

The Outcomes with Regular Video Datasets 1) Tables III and IV show the outcomes of our method applied to various available video-based person datasets. Both the iLIDS-VID and PRID 2011 datasets' respective outcomes are documented in their respective primary literatures. The literature [13] reports the outcomes of STFV3D and KISSME. The outcomes of both RNNCNN and ASTPN are documented in the literature [48].

The findings of RNNCNN and ASTPN for the MARS dataset can be seen in [48]. We discuss the results of studies conducted using PCB on three video-based datasets.

It is clear that the matching rates achieved by all approaches are greater in the typical scenario than they are in the CGVPR scenario. On three different video-person datasets, PCB network design yields superior results. When compared to cross-modality situations, normal scenarios have more effective information available. When compared to the results of competing approaches on three native datasets, our approach SDPL is the closest. Two benefits result: I The space-time data may be aligned by using the several walking cycles included in each pedestrian video. When minimizing individual differences in appearance, the video projection concepts W and

V are very useful. ii) Dictionary learning is a powerful categorization approach that provides accurate representations of data.

Two, the Semi-coupled Mapping Term's Impact: We devise the following validation experiment to measure the efficacy of semi-coupled mapping. In this case, we only get rid of the semi-coupled mapping term in

TABLE II
TOP R RANKED MATCHING RATE WITH STANDARD DEVIATION OF AVERAGE VALUES (%) ON THE CGVID DATASET. BEST RESULTS ARE IN BOLDFACE FONT.

Feature	Method	Gallery: True color, Probe: Gray scale					Gallery: Gray scale, Probe: True color				
		r=1	r=5	r=10	r=20	mAP	r=1	r=5	r=10	r=20	mAP
STFV3D	STFV3D	11.8±3.1	32.1±3.2	53.8±2.6	69.8±2.8	6.3±2.7	10.3±3.1	32.2±3.9	52.9±2.9	68.3±3.4	5.4±2.3
	KISSME	15.9±2.9	43.4±3.4	62.6±1.0	75.2±2.9	16.6±2.3	14.3±3.6	43.5±3.4	62.1±3.7	73.7±3.5	15.9±3.1
	XQDA	16.3±2.1	48.6±3.3	64.2±3.1	79.1±3.0	17.3±1.9	14.7±2.9	48.7±2.3	64.3±3.4	77.4±4.1	16.1±2.6
	TDL	18.1±3.3	47.5±2.3	65.6±1.1	83.5±0.5	18.7±2.9	16.7±3.0	47.4±2.9	65.6±3.1	80.1±2.9	18.8±2.3
	SI^2DL	19.9±2.7	52.0±3.4	70.2±1.9	84.3±2.3	19.4±2.6	18.1±2.7	51.9±2.9	70.2±3.3	80.4±4.1	18.9±2.6
	JDML	20.1±3.1	49.7±2.9	69.4±2.1	83.5±1.9	18.3±2.6	18.5±3.4	49.7±3.0	69.5±2.9	81.9±3.4	18.1±3.1
	SDPL	**23.3±2.1**	**53.3±2.8**	**74.6±2.0**	**87.3±1.8**	**19.6±1.8**	**22.1±3.1**	**53.4±2.8**	**73.7±3.0**	**86.0±2.6**	**19.1±2.7**
CNN	RNNCNN	33.0±1.3	76.5±0.9	84.6±1.3	93.1±1.0	23.6±1.1	29.1±1.4	74.4±1.0	84.6±0.9	91.0±0.8	22.5±1.3
	ASTPN	34.7±1.1	77.4±0.8	87.9±1.1	95.2±0.9	24.8±1.0	30.8±1.1	77.3±0.8	87.8±0.8	93.3±0.9	23.9±1.1
	DeepZero	26.7±1.4	72.6±0.9	82.9±1.4	90.2±1.0	19.7±1.1	25.1±1.6	72.5±1.1	83.0±1.1	88.5±1.1	18.9±1.4
PCB	PCB	43.4±1.6	80.3±0.9	91.3±1.1	95.5±0.9	41.1±1.5	41.3±1.3	80.2±1.4	91.2±1.1	94.9±0.9	37.2±1.1
	KISSME	46.8±1.1	82.7±1.0	91.5±0.9	95.9±0.6	43.2±1.0	45.4±1.4	80.4±1.3	89.2±1.2	93.7±1.1	42.6±1.3
	XQDA	46.1±1.3	80.4±1.1	90.8±1.3	95.5±0.8	39.7±1.1	45.2±0.9	79.4±0.9	89.1±0.9	93.6±1.2	38.9±0.8
	TDL	47.8±1.1	77.9±0.9	90.9±0.8	97.1±0.7	43.1±0.9	45.7±1.3	76.9±1.4	89.8±0.8	95.1±1.1	41.6±1.1
	SI^2DL	47.6±1.4	81.6±0.8	92.0±0.9	96.9±1.0	42.9±1.3	45.7±1.4	80.6±0.9	91.0±1.1	95.2±0.9	**42.8±1.3**
	JDML	47.9±1.0	83.2±1.1	92.7±1.1	98.3±0.9	39.7±0.9	45.9±1.3	82.2±1.1	91.7±1.0	96.3±1.0	38.9±0.9
	SDPL	**49.9±1.1**	**85.2±1.0**	**95.4±0.9**	**98.6±0.9**	**43.9±1.0**	**48.2±1.1**	**84.5±1.0**	**94.7±1.1**	**96.9±0.9**	**42.8±0.8**

TABLE III
TOP R RANKED MATCHING RATE (%) ON TWO NATIVE VIDEO-BASED PERSON DATASETS, INCLUDING PRID 2011 AND iLIDS-VID. THE STANDARD DEVIATIONS OF AVERAGE VALUES OF OUR APPROACH ARE PROVIDED.

Method	PRID 2011				iLIDS-VID			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
STFV3D [13]	42.1	71.9	84.4	91.6	37.0	64.3	77.0	86.9
KISSME [13]	64.1	87.3	89.9	92.0	44.3	71.7	83.7	91.7
CNN+XQDA [36]	77.3	93.5	-	99.3	53.0	81.4	-	95.1
TDL [20]	56.7	80.0	87.6	98.3	56.3	87.6	95.6	98.3
SI^2DL [21]	76.7	95.6	96.7	97.3	48.7	81.1	89.2	97.3
RNNCNN [11]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
ASTPN [48]	77.0	95.0	99.0	99.0	62.0	86.0	94.0	98.0
PCB [46]	83.1	97.3	99.4	99.6	58.3	83.9	91.3	97.3
STFV3D+SDPL	76.7±2.6	95.4±1.9	96.5±2.1	97.2±2.0	49.8±3.3	81.3±3.1	89.1±4.1	97.6±1.5
PCB+SDPL	83.3±1.3	97.1±1.1	99.3±0.2	99.5±0.1	57.7±2.6	83.6±1.4	91.9±1.9	97.1±1.3

TABLE IV
TOP R RANKED MATCHING RATE (%) ON THE NATIVE VIDEO-BASED PERSON DATASET MARS. THE STANDARD DEVIATIONS OF AVERAGE VALUES OF OUR APPROACH ARE PROVIDED.

Method	MARS			
	r=1	r=5	r=10	r=20
CNN+KISSME [36]	65.0	81.1	-	88.9
LOMO+XQDA [36]	30.7	46.6	-	60.9
CNN+XQDA [36]	65.3	82.0	-	89.0
SI^2DL [21]	72.8	90.4	92.3	96.2
RNNCNN [48]	40.0	64.0	70.0	77.0
ASTPN [48]	44.0	70.0	74.0	81.0
PCB [46]	57.8	75.3	81.4	86.4
PCB+SDPL	63.8±3.1	82.3±2.7	85.8±3.1	89.8±2.6

SDPL, with the reworked version being dubbed SDPL-P. We test CGVID with spacetime features STFV3D and PCB via deep learning.

Table V details the findings. The limited palette of 256 shades in a grayscale video's lone color channel limits the amount of information it can convey. Because of its greater information density, true-color video (which has three channels and 256,255,256 levels) is more popular.

illustration of physical things. Without the semi-coupled mapping matrix, one can observe that the matching rates drop dramatically.

In a sense, mapping may make up for the information loss inherent to grayscale mode.

TABLE V
TOP *r* RANKED MATCHING RATE WITH STANDARD DEVIATION (%) OF SDPL AND SDPL-P ON CGVID.

Feature	Method	CGVID			
		r=1	r=5	r=10	r=20
STFV3D	SDPL-P	20.2±2.3	50.7±2.4	71.9±2.3	84.4±1.7
	SDPL	23.3±2.1	53.3±2.8	74.6±2.0	87.3±1.8
PCB	SDPL-P	48.1±1.3	83.6±1.1	93.8±1.3	97.9±1.3
	SDPL	49.9±1.1	85.2±1.0	95.4±0.9	98.6±0.9

Thirdly, the effect of the projection term inside the video itself is that W and V may reduce the size of the original video by making it grayscale or full color. As an attempt at measuring withinvideo's efficacy,

projection matrices, we perform our method using the semi-coupled mapping matrix and the discriminant dictionary learning, excluding the withinvideo projection terms from SDPL in the process. Here are the specifics of the experiments: Our goal is what we refer to as the

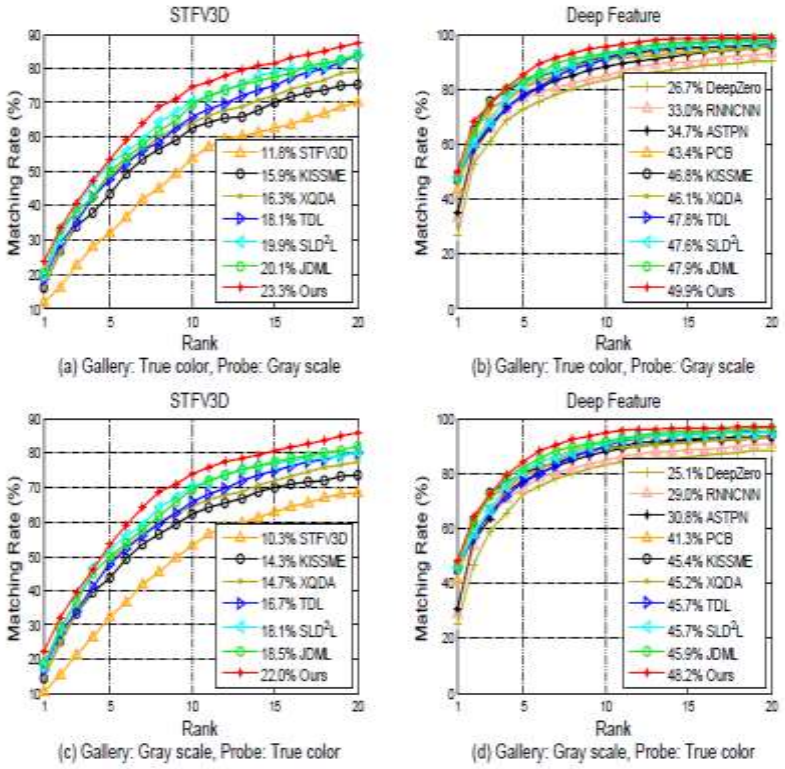


Fig. 8. Results on new dataset CGVID. The rank-1 matching rate of each method is provided in the legend. (a) Results with STFV3D (Gallery: True color, Probe: Gray scale). (b) Results with deeply-learned features (Gallery: True color, Probe: Gray scale). (c) Results with STFV3D (Gallery: Gray scale, Probe: True color). (d) Results with deeply-learned features (Gallery: Gray scale, Probe: True color).

function out of WV as SDPL-WV, i.e., take two projection matrices from inside the video out of the objective function; ii) We take one of the two terms from within the video out of the objective function.

that either the SDPL-W or SDPL-V refers to. We test CGVID with the space-time feature STFV3D and the deep-learning feature PCB. Results are shown in Table VI. We can observe that when we remove the within-video projection words, the matching rates drop dramatically. As a result, the projection terms inside the video may help lower the within-video variance and contribute significantly to the overall aim.

TABLE VI
TOP *r* RANKED MATCHING RATE WITH STANDARD DEVIATION OF AVERAGE VALUES (%) OF SDPL AND SDPL WITHOUT WITHIN-VIDEO PROJECTION MATRICES ON CGVID.

Feature	Method	CGVID			
		r=1	r=5	r=10	r=20
STFV3D	SDPL-WV	19.6±2.3	48.0±1.9	66.5±2.6	79.1±2.1
	SDPL-V	21.2±3.1	53.0±2.6	69.9±2.1	85.2±2.3
	SDPL-W	20.9±2.0	51.7±3.0	72.1±3.1	85.3±2.6
	SDPL	23.3±2.1	53.3±2.8	74.6±2.0	87.3±1.8
PCB	SDPL-WV	46.3±1.3	81.1±1.6	91.8±1.6	96.5±1.1
	SDPL-V	47.1±1.4	83.2±1.1	92.7±1.3	97.3±1.2
	SDPL-W	47.6±1.0	81.5±1.3	92.3±1.4	97.1±1.1
	SDPL	49.9±1.1	85.2±1.0	95.4±0.9	98.6±0.9

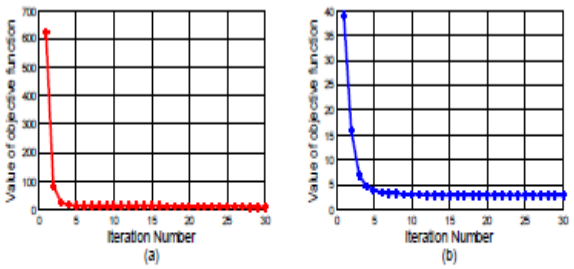


Fig. 9. Convergence curve on CGVID. (a) The convergence of our approach SDPL with STFV3D; (b) The convergence of our approach SDPL with PCB.

4) Convergence Analysis: In this work, we alternate between updating the dictionary pair, the semi-coupled mapping, and the within-video representation in order to maximize the objective function (5).

matrix projections in a recursive fashion. Each every step's sub-problem may be written as a convex one. The primary variable of interest in this experiment is the evolution of the energy value of our goal function as a function of iteration.

Figure 9 demonstrates that our approach will converge on CGVID in less than 20 rounds. When doing CGVID's training phase computations in Matlab 9.0 on a machine with an Intel i7 3.00GHz and 32G RAM, the process takes around 170 seconds. In Fig. 9, the first step has a high objective function value (a). A possible explanation for this is

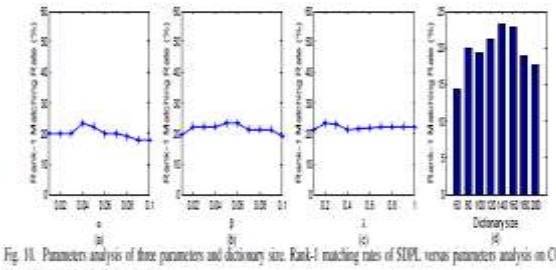


Fig. 10. Parameters analysis of three parameters and dictionary size. Rank-1 matching rates of SDPL versus parameters analysis on CGVID. (a) α, (b) β, (c) λ, (d) dictionary size.

might significant variations in STFV3D characteristics of the same individual occur when seen from different angles. In Fig. 9, PCB results in a low objective function value (b). PCB is

a learnt trait with little inter-camera variation for the same individual using the PCB feature.

Finally, we perform a running cost experiment using our CGVID dataset for training and testing set, each of which consists of 100 human pairs, to assess the training and testing efficiency of the compared competing approaches. Every procedure is run on a computer outfitted with an Intel i7-5960X processor running at 3.00GHz, 32GB of RAM, and a Geforce GTX1080Ti graphics processing unit. Traditional approaches like KISSME, XQDA, TDL, JDML, SI2DL, and SDPL are carried out using the PCB function. Principal component analysis takes PCB features, whose initial dimensionality is 12288, and reduces them to 200 for use in traditional techniques (PCA). Training and testing times for 10 random CGVID trials are shown in Table VII.

The computational cost of distance-learning methods like KISSME, SI2DL, and XQDA scales linearly with the size of the feature dimension, and these methods need O(n) time (N2). The number of feature dimensions is N. As a matter of fact, it runs in under 1 second. Distance learning is the basis of TDL, and in order to compute an optimal positive semidefinite matrix using stochastic gradient descent, it requires several rounds. In terms of iterations, TDL has a complexity of O(kN2). It takes around 5.3 seconds to train, but less than 1 second to put through its paces in a test environment. Dictionary learning, upon which JDML is built, requires periodic revisions to the coding coefficients. Complexity in terms of time is O. (N3). Both RNNCNN and ASTPN use a network structure that is developed by deep learning. Convolution, recurrent, and back propagation take up the majority of their processing time, and they have a computational cost of O(M2K2CinCout). The feature map generated by the convolutional kernel has a size of M. The size of the convolutional kernel, K, determines its effectiveness. The convolutional kernel's channel number is denoted by Cin. This layer's convolutional kernel has a number associated with it, denoted by Cout.

While RNNCNN and ASTPN each take around 100 seconds to extract features, examining the probe set's 128 dimensions only takes about 0.003 seconds. Tests show that our method takes less than 0.2 seconds to correctly re-identify a pedestrian.

TABLE VII
COMPARISON OF TRAINING AND TESTING TIME (SECONDS) ON THE CGVID DATASET.

Method	Ave Time (Training)	Ave Time (Testing)
KISSME	0.02	0.01
TDL	5.3	0.009
SI^2DL	9.4	0.13
XQDA	0.06	0.001
JDML	16.6	0.26
PCB	266*400	4.3
RNNCNN	10*500	0.003
ASTPN	13*500	0.003
SDPL	170	0.16

Here in Section 6, we analyze three of the most important elements of our methodology. To get the greatest results from our trials, we empirically tune these variables.

rate of similarity The discriminant term's impact is tempered by the parameter. This parameter compensates for the bias introduced by sample pairs with a negative value. The significance of regular phrases is modulated by a parameter. In order to assess one parameter, all the others must remain constant. We plot the rank-1 matching rates of our method against various values of in Figs. 10 (a)-(c). Maximum rank-1 matching occurs at. is stable between [0.04,0.06] and [0.1,0.3] and is stable between [0.04,0.06] and [0.1,0.3]. The optimum performance of our method is reached when, for example, 0.04, 0.06, and 0.2 are used.

A significant aspect in the goal function is the dictionary size, which is the sum of the atomic numbers in DC and DG. We test out varying dictionary sizes, from 60 to 200, in order to see what happens. Both dictionaries of a pair (DC and DG) have the same size.

Rank-1 matching rate vs dictionary size is seen in Fig. 10 (d). It has been shown that the optimal performance is attained with a dictionary size of 140.

We investigate the failed case pairs using two types of features, namely the space-time feature STFV3D and the deeplylearned feature PCB, to better gauge the efficacy of our new dataset CGVID and our method SDPL. Figure 11 displays the compatibility findings for positions 1 through 5. One channel is used in the grayscale

movies, which will lead to less useful data being lost. If two adjacent pixels have different values for each of the RGB channels,

In grayscale mode, the pixel values may be equivalent.

For instance, in true-color mode, the pixel values of pure red and pure blue are distinct even if they may be quite close. to a close, grayscale Fig. 11 (a). Figure 11 (a) and (b) incorrectly match true-color mode with grayscale mode, even though there are noticeable variances between the two (c). However, there is still a lot of space for improvement in the CGVPR situation, since the rank-1 matching rates are only near to 50% with deeply-learned feature PCB.

Rankings Across Multiple Datasets (9) Taking cues from [36], we analyze how well the CGVID dataset generalizes.

Here, we implement experiments using several datasets.

We choose the MARS, PRID 2011, and iLIDSVID datasets, which are three of the most popular human reid video resources.

We use the whole CGVID as a training set and run tests with two different sets of input data: I both cameras capture in full color, and ii) the tested video sequences are converted from camera B to grayscale while those from camera A remain in full color. We do an investigation into how well our new dataset generalizes to other situations. We directly extract features from three video datasets using the deep model we obtained by first training the PCB model on the CGVID dataset. Next, we pick two common distances for matching, the Euclidean distance and the Mahalanobis distance, and calculate the matching rates using the characteristics of the video datasets. Table VIII details the findings. As can be shown across all three datasets, KISSME allows the CGVID-learned model to provide superior re-id performance compared to Euclidean. The CGVID dataset has enough picture sequences to train deep-learning techniques adequately. For further study of the person re-id issue, CGVID's large size (more than 50K frames) will be an asset.

TABLE VIII
TOP R RANKED MATCHING RATE (%) ON TESTING DATASETS WITH THE CGVID DATASET FOR CROSS PERFORMANCE. TRAINED ON CGVID AND DIRECTLY TESTED ON MARS, PRID 2011 AND iLIDS-VID. * DENOTES THE TESTED DATASETS WITH BOTH TRUE-COLOR MODALITIES.

Method	Tested On	r=1	r=5	r=10	r=20
Euclidean	MARS	22.5	36.4	47.7	60.8
	PRID 2011	26.6	58.0	75.8	86.2
	iLIDS-VID	10.2	28.2	41.4	60.4
KISSME	MARS	27.6	46.0	56.2	68.0
	PRID 2011	33.2	64.8	80.0	90.6
	iLIDS-VID	15.0	36.2	48.0	66.0
Euclidean*	MARS	41.6	59.4	66.8	74.0
	PRID 2011	22.4	52.8	69.8	85.4
	iLIDS-VID	16.4	30.0	44.0	62.8
KISSME*	MARS	44.5	65.2	73.6	80.4
	PRID 2011	34.6	64.2	77.2	89.0
	iLIDS-VID	26.8	46.2	63.0	78.0

BOTTOM LINE

In this study, we have given a standard video pedestrian re-id dataset in both true color and grayscale.

CGVID in a real-world context. The 52,723 frames in our sample represent 200 unique individuals, making it the largest video re-id dataset to date.

Each sequence also has a frame count more than 58, which is optimal for space-time analyses since it represents at least two full walking cycles. We present a semi-coupled dictionary pair learning (SDPL) method based on asymmetric within-video projection to solve the CGVPR issue. The learnt withinvideo projection matrices allow for a level of uniformity to be applied to each true-color or grayscale video. To assist subsequent matching between true-color and grayscale recordings, a semi-coupled mapping matrix is learnt to reveal the connection between characteristics of each video type.

We test with two common feature types, the standard space-time feature STFV3D and the deep feature PCB. The suggested method has been shown to be successful in experiments for the CGVPR problem. Re-identifying an individual in CGVID is difficult because to the diversity of the movies captured by the many cameras. CGVID is favored for additional testing of new algorithms because to its large size (over 50,000 frames). In addition, the detection quality of each individual is 100% since each bounding box is tagged by hand. Our long-term goal is to enhance the efficiency of existing matching algorithms for the CGVPR problem.

ACKNOWLEDGMENT

The authors appreciate the helpful feedback and ideas from the editors and the blind reviewers. The National Key Research and Development Program of China (Grant No. 2017YFB0202001), the National Nature Science Foundation of China (Grant Nos. 61672208, U1504611, 41571417), the Natural Science Foundation Key Project for Innovation Group in Hubei Province (Grant No. 2018CFA024), and the Science and Technique Development Program of China (Grant No. 2018CFA024) all provided funding for this study.

REFERENCES

[1] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi and S. Z. Li, "Salient

Color Names for Person Re-identification," in Proc. Eur. Conf.

Comput. Vis. (ECCV), 2014, pp. 536-551.

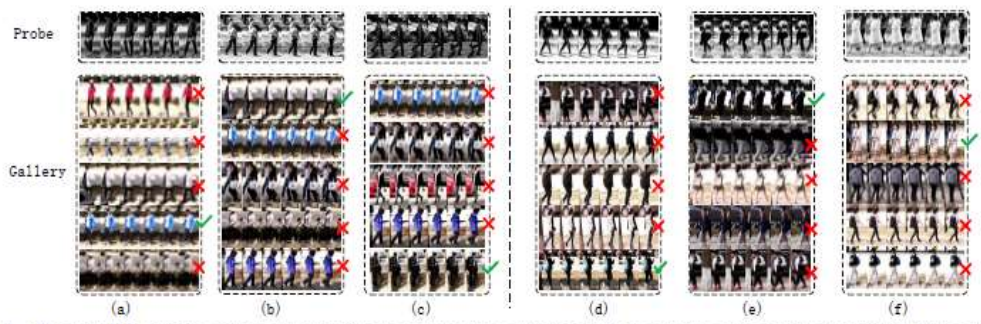


Fig. 11. The rank 1-5 matching results corresponding to persons from grayscale camera and true-color camera respectively. The red mark × means the wrongly matched samples from gallery set. The green mark ✓ denotes the same sample from true-color camera. The top row is the sample sequences from the probe with grayscale mode. The remained rows are the matching results from gallery set with true-color. (a)-(c) The matching results with feature PCB. (d)-(f) The matching results with feature STFV3D.

[2] W. Zhang, B. Ma, K. Liu and Rui Huang, "Video-Based Pedestrian Re-Identification by Adaptive Spatio-Temporal Appearance Model," IEEE Trans. Image Processing, vol. 26, no.4, pp. 2042-2054, 2017.

[3] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen and R. Hu, "Person Reidentification via Ranking Aggregation of Similarity Pulling and Dissimilarity Pushing," IEEE Trans. Multimedia, vol. 18, no. 12, pp. 2553-2566, 2016.

[4] X. Zhu, X.-Y. Jing, L. Yang, X.G. You, D. Chen, G. Gao and Y. Wang, "Semi-Supervised Cross-View Projection-Based Dictionary Learning for Video-Based Person Re-Identification," IEEE Trans. Circuits Syst. Video Techn., vol. 28, no. 10, pp. 2599-2611, 2018.

[5] S. Iodice and A. Petrosino, "Salient feature based graph matching for person re-identification," Pattern Recognition, vol. 48, no.4, pp. 1074-1085, 2015.

[6] E. Ahmed, M. J. Jones and T. K. Marks, "An improved deep learning architecture for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3908-3916.

[7] S. Yi, Z.Y. He, Y.M. Cheung and W.S. Chen, "Unified Sparse Subspace Learning via Self-Contained Regression," IEEE Trans. Circuits Syst. Video Techn., vol. 28, no. 10, pp. 2537-2550, 2018.

[8] D. Li, X. Chen, Z. Zhang and K. Huang, "Learning Deep Context-Aware Features over Body and Latent Parts for Person Reidentification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 7398-7407.

[9] Z. Zhou, Y. Huang, W. Wang, L. Wang and T. Tan, "See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 6776-6785.

[10] S. Zhou, J. Wang, J. Wang, Y. Gong and N. Zheng, "Point to Set Similarity Based Deep Feature Learning for Person Re-Identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp.5028-5037.

[11] N. McLaughlin, J. M. del Rinc´on and P. C. Miller, "Recurrent Convolutional Network for Video-Based Person Re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1325-1334.

[12] S. Yu, X.G. You, W. Ou, X. Jiang, K. Zhao, Z. Zhu, Y. Mou and X. Zhao, "STFT-like time frequency representations of nonstationary signal with arbitrary sampling schemes," Neurocomputing,204, pp. 211-221, 2016.

[13] Kan Liu, Bingpeng Ma, Wei Zhang, Rui Huang, "A Spatio-Temporal Appearance Representation for Video-Based Pedestrian Re-Identification," in Proc. IEEE Conf. ICCV, 2015, pp. 3810-3818.

[14] J. Chen, Z. Zhang and Y. Wang, "Relevance Metric Learning for Person Re-Identification by Exploiting Listwise Similarities," IEEE Trans. Image Processing, vol. 24, no. 12, pp. 4741-4755, 2015.

[15] N. Mel, Christian Micheloni, Gian Luca Foresti, "Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning," IEEE Trans. Image Processing, vol. 24, no. 12, pp. 5645-5658, 2015.

[16] S. Bak and P. Carr, "One-Shot Metric Learning for Person Reidentification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2017, pp. 1571-1580.

[17] D. Tao, L. Jin, Y. Wang and X. Li, "Person Reidentification by Minimum Classification Error-Based KISS Metric Learning," IEEE Trans. Cybernetics, vol. 45, no. 2, pp. 242-252, 2015.

[18] X. Yang, M. Wang and D. Tao, "Person Re-Identification With Metric Learning Using Privileged Information, " IEEE Trans. Image Processing, vol. 27, no. 2, pp. 791-805, 2018.

[19] Z.-X. Feng, J. Lai and X. Xie, "Learning View-Specific Deep Networks for Person Re-Identification," IEEE Trans. Image Processing, vol. 27, no. 7, pp. 3472-3483, 2018.

[20] J. You, A. Wu, X. Li and W.-S. Zheng, "Top-Push Video-Based Person Re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1345-1353.

[21] X. Zhu, X.-Y. Jing, X.G. You, X. Zhang and T. Zhang, "Video-Based Person Re-Identification by Simultaneously Learning Intra-Video and Inter-Video Distance Metrics," IEEE Trans. Image Processing, 27(11), pp. 5683-5695, 2018.

[22] A.s Ess, B. Leibe and L. J. Van Gool, "Depth and Appearancefor Mobile Scene Analysis," in Proc. IEEE Conf. Comput. Vis. (ICCV), 2007, pp. 1-8.

[23] H. Martin, B. Csaba and Bischof, "Person re-identification by descriptive and discriminative classification," Scandinavian Conference on Image Analysis, pp. 91-102, 2011.

[24] S. Wang, L. Zhang, Y. Liang and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp. 2216-2223.

[25] X. -Y. Jing, X. Zhu, F. Wu, R. Hu, X.G. You, Y. Wang and J.- Y. Yang, "Super-Resolution Person Re-Identification With Semi-Coupled Low-Rank Discriminant Dictionary Learning," IEEE Trans. Image Proces., vol. 26, no. 3, pp. 1363-1378, 2017.

[26] L. Ren, J. Lu, J. Feng and J. Zhou, "Multi-modal uniform deep learning for RGB-D person re-identification," Pattern Recognition, 72, pp. 446-457, 2017.

[27] K. Li, Z. Ding, S. Li and Y. Fu, "Discriminative Semi-Coupled Projective Dictionary Learning for Low-Resolution Person Re-Identification," in Proc. AAAI, 2018, pp. 2331-2338.

[28] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in Proc. Conf. on Neural Inform. Proces. Sys. (NIPS), 2014, pp. 793–801.

[29] D. Gray, Hai Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2008, pp. 262-275.

[30] W. Li, R. Zhao and X. Wang, "Human Reidentification with Transferred Metric Learning," in Proc. ACCV, 2012, pp. 31-44.

[31] W. Li and X. Wang, "Locally Aligned Feature Transforms across Views," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2013, pp. 3594-3601.

[32] W. Li, R. Zhao, T. Xiao and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2014, pp. 152-159.

[33] L. Zheng, L. Shen, L. Tian, S.n Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," in Proc. IEEEConf. Comput. Vis. (ICCV), 2015, pp. 1116-1124.

[34] A. Wu, W.-S. Zheng, H. Yu, S. Gong and J. Lai, "RGB-Infrared

Cross-Modality Person Re-identification," in Proc. IEEE Inter. Conf. on Comput. Vis. (ICCV), 2017, pp. 5390-5399.

[35] P. Dai, R. Ji, H. Wang, Q. Wu and Y. Huang, "Cross-Modality Person Re-Identification with Generative Adversarial Training," in Proc. Inter. Joint Conf. on Artif. Intell. (IJCAI), 2018, pp. 677-683.

[36] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang and Q.Tian, "MARS: A Video Benchmark for Large-Scale Person Re-Identification," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016,pp. 868-884.

[37] M. Munaro, A. Fossati, A. Basso, E. Menegatti and L. Gool,"One-shot person re-identification with a consumer depth camera,"In Proc. ACVPR, pp. 161–181. Springer, London, 2014.

[38] Z. Li, X.-Y. Jing and X. Zhu, "On the multiple sources and privacy preservation issues for heterogeneous defect prediction,",TSE, 2017.

[39] M. Munaro, A. Basso, A. Fossati, L. Van Gool, E. Menegatti,"3D reconstruction of freely moving persons for re-identification with a depth sensor,"in Proc. ICRA, pp. 4512-4519, May 2014.

[40] F. Jahan, M. K. Islam and J.-H. Baek, "Person Detection, Reidentification and Tracking Using Spatio-Color-based Model for Non-Overlapping Multi-Camera Surveillance Systems," Smart CR,vol.2, no.1, pp. 42-59, 2012.

[41] F. Fleuret, H. Shitrit and P. Fua, "Re-identification for Improved People Tracking," Person Re-Identification, 2014, pp. 309-330.

[42] Z. He, S. Yi, Y.-M. Cheung, X.G. You and Y. Tang, "Robust Object Tracking via Key Patch Sparse Representation". IEEE Trans. Cybernetics, vol. 47, no. 2, pp. 354-364, 2017.

[43] Z. Chen, X. You, Boxuan Zhong, Jun Li, Dacheng Tao, "Dynamically Modulated Mask Sparse Tracking," IEEE Trans. Cybernetics,vol. 47, no. 11, pp. 3706-3718, 2017.

[44] T. Wang, S. Gong, X. Zhu and S. Wang, "Person Re-identification by Video Ranking," in Proc. Eur. Conf. Comput. Vis. (ECCV),2014, pp. 688-703.

[45] T.Wang, S. Gong, X. Zhu and S.Wang, "Person Re-Identification by Discriminative Selection in Video Ranking," IEEE Trans.Pattern Anal. Mach. Intell. vol. 38, no.12, pp. 2501-2514,2016.

[46] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling," in Proc. Eur.Conf. Comput. Vis. (ECCV), 2018.

[47] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.

[48] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang and P. Zhou,"Jointly Attentive Spatial-Temporal Pooling Networks for Video-Based Person Re-identification," in Proc. IEEE Conf. Comput. Vis.(ICCV), 2017, pp. 4743-4752.

[49] M. Yang, L. Zhang, X. Feng and D. Zhang, "Fisher Discrimination Dictionary Learning for sparse representation," in Proc. IEEE Conf. Comput. Vis. (ICCV), 2011, pp. 543-550.

[50] S. Li, M. Shao and Y. Fu, "Cross-View Projective Dictionary Learning for Person Re-Identification," in Proc. Inter. Joint Conf. Artif. Intell. (IJCAI), 2015, pp. 2155-2161.

[51] M. K¨ostinger, M. Hirzer and H. Bischof, "Large scale metric learning from equivalence constraints," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp. 2288-2295.

[52] S. Liao, Y. Hu, X. Zhu and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015,pp. 2197-2206.

[53] Q. Zhou, S. Zheng, H. Ling, H. Su and S. Wu, "Joint dictionary and metric learning for person re-identification," Pattern Recognition,72, pp. 196-206, 2017.

Fei Ma received the M.S. degree in computer software and theory from Yunnan Normal University in 2006. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.

Xiao-Yuan Jing received the Ph.D. degree of Pattern Recognition and Intelligent System in the Nanjing University of Science and Technology, 1998. Now he is a Professor with the School of Computer, Wuhan University, and the School of Computer, Guangdong University of Petrochemical Technology, China. His research interests include pattern recognition, image processing, computer vision and machine learning.

Zhiping Peng received his Ph.D. in Computer Application Technology from South China University of Technology in 2007. He is currently the Vice President and Professor of the School of Computer Science at the Guangdong University of Petrochemical Technology, China. His research interests include artificial intelligence application, cloud computing and information processing.

Xiaoke Zhu received the Ph.D. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China in 2017. Now he is an associate professor with the School of Computer, Henan University, China. His research interests include computer vision and machine learning.

Zhenmin Tang received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China. He is a Professor at the Nanjing University of Science and Technology. He is also the leader of several key programs of the National Nature Science Foundation of China. His major research interests include intelligent system, pattern recognition and machine learning.