Celebrities-ReID: A Benchmark for Clothes Variation in Long-Term Person Re-Identification

Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong
School of Electrical and Data Engineering
University of Technology, Sydney
Sydney, Australia
Yan.Huang-3@student.uts.edu.au, Qiang.Wu@uts.edu.au, JingSong.Xu@uts.edu.au, Yi.Zhong@student.uts.edu.au

***Abstract*—This paper considers person re-identification (re-ID) in the case of long-time gap (*i.e.*, long-term re-ID) that concentrates on the challenge of clothes variation of each person. We introduce a new dataset, named Celebrities-reID to handle that challenge. Compared with current datasets, the proposed Celebrities-reID dataset is featured in two aspects. First, it contains 590 persons with 10,842 images, and each person does not wear the same clothing twice, making it the largest clothes variation person re-ID dataset to date. Second, a comprehensive evaluation using state of the arts is carried out to verify the feasibility and new challenge exposed by this dataset. In addition, we propose a benchmark approach to the dataset where a two-step fine-tuning strategy on human body parts is introduced to tackle the challenge of clothes variation. In experiments, we evaluate the feasibility and quality of the proposed Celebrities-reID dataset. The experimental results demonstrate that the proposed benchmark approach is not only able to best tackle clothes variation shown in our dataset but also achieves competitive performance on a widely used person re-ID dataset Market1501, which further proves the reliability of the proposed benchmark approach.**

***Index Terms*—person re-identification, clothes variation, body parts, deep learning**

I. INTRODUCTION

Person re-identification (re-ID) aims to associate a target (person) who appeared under multiple camera views at distinct time points. It is a challenging computer vision task and underpins significant applications such as person retrieval, cross camera object tracking, target reacquisition, *etc* [1]. According to the gap of time, person re-ID can be categorized into two types: the short-term and the long-term scenarios. A majority of current researches belong to the former which assumes a person will reappear under non-overlapping camera views within a brief period (*e.g.*, less than 30 minutes) without the changes of clothes [2]–[8], [31]. On the other hand, persons in the long-term scenario inherently undergo a great possibility to change their clothes or carry different objects after a long-time gap (*e.g.*, several days or even months) when they reappear under the same view or large spatial separation between views [1]. The certain application of short-term person re-ID is person tracking and retrieval within time and space limitations. To break these limitations, long-term person re-ID is indispensable in many real-world applications such as forensic search and identity verification for the security surveillance.

In the previous short-term person re-ID studies, the main challenge normally comes from 'general' variations, *e.g.*, illuminations, viewpoints, poses, occlusions, weather in outdoor scenes, and background clutters. To this end, several datasets have been proposed to tackle such variations [2], [7], [9]–[12], [32]. However, dressing differently, which is one of the most important variation in the long-term person re-ID scenario, has not been observed amongst the above-mentioned datasets. Apparently, the same person wearing different clothes involves more indistinguishable and ambiguous factors (*e.g.*, causing large inter-class indistinctness and intra-class discreteness) which result in a significant challenge in person re-ID. An ideal re-ID system should embed the capacity to tackle such challenge in realistic [1]. However, this problem is still not well addressed in current researches. Therefore, we will introduce a dataset with a proper benchmark method proposed.

Some researches attempt to propose datasets that do not rely on the cue of clothes. To the best of our knowledge, these datasets can be categorized into three types, *i.e.*, the RGB-D datasets [13]–[15], the appearance of impaired datasets [16], and the general video-based dataset [17]. Apart from person re-ID, the variation of clothes can be also observed in gait recognition dataset [18], [33]. However, unfortunately, all of them are not suitable to the long-term re-ID study due to the limited experimental conditions (*e.g.*, under controlled environments such as indoor scenes [18] or need special cameras such as *Kinect* to capture RGB-D images [13]–[15]) and the scale (the detailed statistics of each dataset are listed in Section II).

Apart from datasets, current person re-ID approaches mainly focus on dealing with challenges in the short-term scenario such as variations of viewpoint, pose, illumination, *etc*. These methods heavily rely on the appearance cue such as color and texture of clothes since a person will not change chothes. These approaches can be categorized into two types: traditional learning methods and deep learning methods. The former utilizes appearance-based descriptors such as color, texture, or their combinations to represent a person for metric learning [4], [6], [19], [20]. The latter targets to optimize the performance of a deep neural network by training from the scratch [2], [5] or fine-tuning off-the-shelf pre-trained models [3], [7], [8], [21], [22]. Impressive progress has been achieved in current state-of-the-art methods by using deep

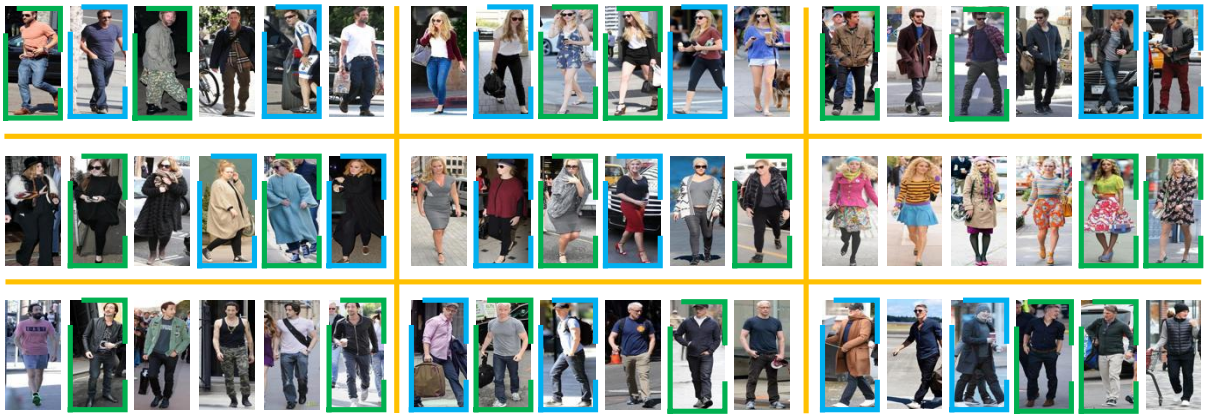

Fig. 1. Nine persons with six randomly selected images per person are shown from the proposed Celebrities-reID dataset. Each person shows different clothes. The blue and green boxes respectively indicate instances of viewpoint and pose variations within a person. The disturbance of background clutters and illuminations also can be observed amongst the nine persons. The glasses and profiles make the face cue of each person unreliable.

learning approaches. However, this work will reveal that such methods are not adequate to tackle the challenge of long-term person re-ID where a person will change clothes.

From the above discussion, in this paper, we concentrate on the study of clothes variation for the case of long-term person re-ID. To this end:

- **Firstly**, we propose a new person re-ID dataset that contains 590 IDs with 10,842 images. The same person does not wear the same clothing twice to make it more suitable for the clothes variation person re-ID study. Also, some other disturbances, *e.g.*, variations of pose, viewpoint, background clutter, and illumination are also embodied in our dataset to make it even more challenging (see Fig. 1). On the other hand, capturing hundreds of persons wearing different clothes after a long-time gap is difficult. To collect a certain number of IDs, we use images of celebrities to build our dataset, and name it as 'Celebrities-reID'. As we use images of celebrities, there is a bias towards real-life the person re-ID scenario (*e.g.*, using pedestrian images captured by surveillance cameras). Therefore, we choose the street snap-shots of celebrities, which is closer to the real-life scenarios. To the best of our knowledge, Celebrities-reID is the largest dataset with a person wearing different clothes.
- **Second**, to verify the feasibility of the new dataset, we evaluate it using several state-of-the-art methods to testify their performance under the new challenge. It well demonstrates that Celebrities-reID is extremely challenging to the existing methods. We introduce a two-step fine-tuning strategy on human body parts (2SF-BPart) for the clothes variation study. In the first fine-tuning step, an overall optimization to a specific backbone neural network (*e.g.*, the pre-trained resnet [23]) is executed on full images. Then, we separately fine-tune the optimized network on each human body part which is divided by a structure-guided partition approach. The motivation to use the part-based approach on our dataset is that: we consider different body regions play distinct roles in matching a person wearing different clothes (*e.g.*, the detailed cues of the body are more critical than appearance, on the contrary, appearance is still important in matching the heads of persons). Also, compared with current methods that directly co-train all body parts or train them separately from the scratch[1], the proposed 2SF-BPart approach demonstrates its superior in eliminating useless information (color or texture of clothes), and keeping effective information (details of the body and appearance of the head) to the proposed Celebrities-reID dataset concurrently.

The contributions of this paper can be summarized in two-folds. **First**, we propose a new dataset Celebrities-reID which is particularly for the clothes variation study in the long-term person re-ID scenario. A new data acquisition pipeline is introduced by using the street-snap shots of celebrities in the uncooperative environment to depress the bias between our data and the real-life scenario. Also, we comprehensive evaluate existing state of the arts to verify the feasibility and new challenge exposed by the proposed dataset. **Second**, a new 2SF-BPart benchmark approach is introduced which simultaneously achieves the best performance on the proposed Celebrities-reID dataset and also well demonstrates its generality to the existing person re-ID dataset.

This paper is organized as follows. We first review some related works in Section II. In Section III, we introduce the pipeline of our data acquisition. In Section IV, we present the benchmark method to the proposed Celebrities-reID dataset. The experiments are shown in Section V. The conclusion is in Section VI.

II. Related Work

Person re-ID dataset. Most previous person re-ID datasets are based on the short-term scenario, such as VIPeR [9], CUHK03 [2], Market-1501 [10], DukeMTMC-reID [7], [24],

[1]In the following section, we use scratch to represent the pre-trained backbone network (*e.g.*, the resnet) on ImageNet, instead of random initialization.

Airport [11], MSMT17 [12], *etc*. These datasets are mainly proposed to tackle the variations of illuminations, poses, viewpoints, background clutters, occlusions, and weather in outdoor scenes, based on an invariant factor: appearance, specifically **clothes**. Since the above-mentioned disturbances are already existed in many existing person re-ID datasets, thereby, in this paper, we concentrate on a new challenge, *i.e.*, clothes variation of long-term person re-ID. In addition, some other disturbances such as illuminations, poses, viewpoints, background clutters are also involved in our dataset to make it even more challenging.

To tackle the change of clothes, the depth cue is utilized to extract 3D soft-biometric feature which is insensitive to appearance variations in RGB-D person re-ID datasets, such as PAVIS [13], BIWI [14], IAS-Lab [14], and DPI-T [15] where a person may wear different clothes. Table I lists the comparison between the proposed Celebrities-reID dataset and other clothes variation datasets. Amongst them, three appearance impaired datasets [16] also can be found. Two of them are people wearing similar dark clothes (without changing clothes). The remaining one (*i.e.*, TSD [16]) just contains 9 IDs with 27 sequences of video with clothes variation. In addition, a general video-based long-term person re-ID dataset 'Motion-ReID' is introduced in [17]. This dataset includes 30 IDs, and each of them contains two different suits of clothes. Compared with the above-mentioned clothes variation datasets, our Celebrities-reID dataset contains the largest number of ID. Up to 590 subjects with around 20 different clothes per ID makes it especially for the clothes variation study in the long-term person re-ID scenario.

Re-ID approach. Current person re-ID methods can be divided into two categories. One is traditional leaning methods, and the other is deep learning methods. Overall, deep learning achieves superior performance. Amongst them, an IDE model that casts the training of person re-ID to an image classification task is presented in [8]. Beyond that, a two-stream architecture, which simultaneously considers the classification and verification functions, is also proposed in [8]. Both utilize the resnet-50 [23] as the backbone network. Moreover, the SVDNet is introduced to decorrelate the learned weight vectors for improving the discriminative learning of the deep CNN network [25]. To utilize the body part cue, the Spindle net [21] is designed to extract human body parts using fourteen body joints, then all the features from different body parts are fused to represent a person. Also, a part aligned representation is given [22] to embed an attention mechanism in training, which allows the model to decide which part to focus on by itself. A refined part model that lays emphasis on the content consistency within each part is proposed in [26]. These approaches are recently published state of the arts which achieve competitive results on various short-term person re-ID datasets. To verify the feasibility of the proposed Celebrities-reID dataset, all of them will be used in our experiments.

In addition to deep learning methods, some traditional learning methods also demonstrate their effectiveness in recent years. However, since most person re-ID datasets are short-term, they mainly rely on appearance-based descriptors. Amongst them, the Local Maximal Occurrence (LOMO) [19] and the Gaussian Of Gaussian (GOG) [20] descriptors demonstrate superior performance. The LOMO [19] is a high dimension descriptor that includes the color and Scale Invariant Local Ternary Pattern (SILTP) histograms. Bins in the same horizontal stripe undergo max pooling and a three-scale pyramid model are built to consider the multi-scale information. Motivated by LOMO, GOG [20] is presented that models each region of a person image by multiple Gaussian distributions to simultaneously describe color and texture information. Whereas, unlike LOMO, GOG further adds covariance information in each hierarchy to fuse color and texture features into a single descriptor. Both use the Cross-view Quadratic Discriminant Analysis (XQDA) [19] to learn a metric model for measuring the similarity between two samples in testing. In this paper, we attempt to utilize the two state-of-the-art descriptors and XQDA metric learning on the proposed Celebrities-reID dataset.

Apart from the above-mentioned approaches, we propose a two-step fine-tuning strategy on human body part as a benchmark on our dataset. The experimental results demonstrate that the proposed method is able to best tackle the clothes variation challenge in the long-term re-ID. Fine-tuning has been adopted in person re-ID [21] in order to adapt the various of different data domains. On the contrary, our method is designed within only one data domain, but lays emphasis on the discrepancy between different body regions when a person wearing different clothes.

III. Dataset Description

The data acquisition process to the newly proposed Celebrities-reID dataset consists of the following five steps:

1) **Determine the List of Celebrities.** We use the street snap-shots of celebrities. We consider the quantity of street snap-shots being proportional to the popularity of celebrities. Therefore, social media networks such as twittercounter[2] are leveraged to determine the name list. We initially determine 2600 celebrities, after data filtering (step 3), 590 of them are retained in the final version.
2) **Crawl Data.** After determining the list, we crawl their street snap-shots images on Google Image with keywords: name + street + snap-shot (*e.g.,* Justin Bieber street snap-shot). For each celebrity, the top 100 returned results are accepted as candidates. Thus, there are 2600×100 images left for manually filtering.
3) **Data Filtering.** Data filtering is a critical step that needs to select images of each celebrity manually. First, the labeler needs to verify the identity of images for each celebrity. Any wrong returned result is discarded. Second, person normally shows standing pose with the whole body in typical re-ID datasets, therefore other poses are not considered. Third, if one celebrity wears

[2]https://twittercounter.com/pages/100

TABLE I
COMPARISON BETWEEN CELEBRITIES-REID AND OTHER CLOTHES VARIATION PERSON RE-ID DATASETS.

Dataset	#Subjects	#Images	change clothing
Image-based dataset (in images)			
Celebrities-reID	**590**	**10,842**	yes, and each person does not wear the same clothes twice.
RGB-D based dataset (video in sequences)			
PAVIS [13]	79	316	some dress differently
BIWI [14]	50	50	most dress differently
IAS-Lab [14]	11	33	some dress differently
DPI-T [15]	12	300	5 different sets of clothes on average for each subject.
Appearance impaired dataset (video in sequences)			
iLIDS-VID BK [16]	91	-	no, but all people wear dark clothes
PRID2011 BK [16]	35	-	no, but all people wear dark clothes
TSD [16]	9	81	27/81 with people wear different clothes, 54/81 are distractors.
Video-based long-term person re-ID dataset (video in sequences)			
Motion-reID [17]	30	240	two suits of clothes for each person

TABLE II
THE SPLIT OF OUR DATASET FOR TRAINING, VALIDATION, AND TESTING.

split	training (490 identities)		testing (100 identities)	
subsets	training	validation	query	gallery
subjects	490	490	100	100
images	8531	490	887	934

the same clothing in more than one images, only one of them is retained. Finally, celebrities with less than 9 images left are excluded from our dataset.

4) **Image Preprocessing.** After data filtering, the raw dataset with 590 identities and 10,842 images is obtained. In this step, each image needs to be processed to produce a standard person image. Three processes are included. First, a bounding box of a person is manually cropped from the original image. After that, paddings are extracted from the original image along the surrounding of the cropped bounding box. These paddings are used to pad the bounding box to a fixed length-width ratio (3:1). Compared with directly resizing the cropped image, using paddings can prevent the changes of body proportion. Finally, all the cropped images are resized to 480×180 according to the pre-defined length-width ratio. Finally, each image is down-sampled to 128×64.
5) **Data Split.** We randomly split our dataset into two sets: training and testing. To ensure complete separation between them, all the images of the same celebrity fall in the same set. That is, it ensures the subjects being tested do not register in the training set. During training, the first image of each subject in the training set is selected to constitute a validation set. We also divide the testing set into two subsets named query and gallery. Each of them covers around 50% images over all the subjects in the testing set. Specifically, given a celebrity in the testing set, half images are used for query and the others are used for gallery. Table II provides more details about the data split.

IV. BENCHMARK FOR CELEBRITIES-REID

We attempt to use several state-of-the-art methods to verify the feasibility of our newly proposed dataset. However, these approaches cannot yield satisfactory results against the challenge of clothes variation. For this reason, we propose a benchmark for our Celebrities-reID dataset that performs promisingly against other approaches.

The new benchmark utilizes a two-step fine-tuning strategy based on human body parts. It is well motivated by considering different body parts can play distinct roles in matching a person wearing different clothes. Specifically, since the appearance cue (*e.g.*, color and texture) is no more reliable in distinguishing the ID of a person who wears different clothes, the detailed cues (*e.g.*, bared parts, the body shape and proportion) become a deterministic factor in matching the similarity between persons. On the contrary, the head still lays emphasis on the appearance (*e.g.*, hair or unclear face, *etc*.) which is beneficial to the re-ID task.

We use the two-stream IDE (2S-IDE) neural network [8] as our backbone since it concurrently considers the identification and verification signals in a co-training architecture. In the first fine-tuning stage, the training set (full image) of Celebrities-reID is utilized to fine-tune the resnet-50 embedded in the two-stream IDE model. The last fully-connected layer is changed to have K neurons to predict K classes. In the second fine-tuning stage, five body parts are respectively utilized to fine-tune the two-stream model optimized from the first stage. The architecture of the 2S-IDE network is shown in Fig. 2.

To exactly locate different body parts, we use an external cue upon the latest progress of human pose estimation [27]. Since drastic appearance changes incurred by clothes variation, we use a coarse-grained instead of a fine-grained partition to retain more effective detailed cues on each part rather than the general appearance cues such as color and texture. Fig. 3 illustrates the partition result. After pose estimation [27], 18 keypoints of a body are obtained, the full image is divided into three parts according to keypoints of the neck and waist.

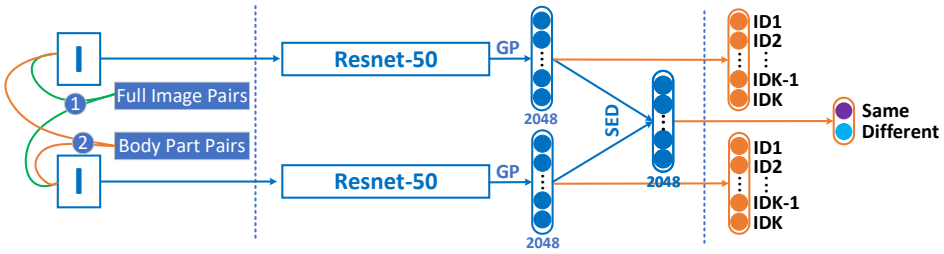

Fig. 2. The architecture of the 2S-IDE network. Image pairs are fed into two paralleled resnet-50 (without weights sharing). After global average pooling (GP), a squared Euclidean distance (SED) is used to calculate the difference between two inputs. Two fully connected layers (K neurons) with cross-entropy loss are used after GPs to classify the ID of each input person (identification signal). A two-class cross-entropy loss is connected after the SED layer to determine whether the two input images belong to the same person or not (verification signal).

Fig. 3. Body parts partition assistances from pose estimation. Given a full image, the keypoints of neck and waist are utilized to divide the body into three parts ①, ② and ③. The last two parts are constituted by ① + ② and ② + ③, respectively.

Their combinations are used to constitute other two parts (see Fig. 3).

Naturally, an alternative of the two-step fine-tuning strategy is to directly train each body part from the scratch. However, the result is inferior compared with our approach. This is because parameters learned by the full image of the proposed Celebrities-reID can provide a better guide towards each body part compared with directly optimizing them from other domain (*e.g.*, ImageNet-trained resnet). Fig. 4 illustrates the effectiveness of learning between scratch $\rightarrow$ body part and full image $\rightarrow$ body part. Also, we can observe that the activation map of our 2SF-BPart method can effectively eliminate more useless information and concentrate on other detailed cue (*e.g.*, bared arms or hands) under drastic appearance changes.

In testing, we feed forward each part to their corresponding fine-tuning network and obtain a 2,048-dim descriptor from the GP layer (see Fig. 2) respectively. Thus, the similarity between two images (*e.g.*, x and y) is calculated by:

$$\begin{aligned} Sim(x,y) \quad &= D(x_{head}, y_{head}) + D(x_{head\_up}, y_{head\_up}) + \\ &\sigma * (D(x_{up}, y_{up}) + D(x_{low}, y_{low}) + \\ &D(x_{up\_low}, y_{up\_low})), \end{aligned} \tag{1}$$

where $D(\cdot)$ is the squared Euclidean distance. x_{head}, x_{up}, x_{low}, x_{head_up}, and x_{up_low} respectively represent the part of head, upperbody, lowerbody, head + upperbody, and upperbody + lowerbody of an image. σ represents the weighs of upperbody, lowerbody, and their combination. We empirically set the $\sigma = 0,5$ because of the drastic appearance changes caused by clothes variation of the three parts.

V. Experiments

A. Evaluation Summary

We will evaluate the quality of our proposed Celebrities-reID dataset and the new benchmark approach in four aspects. First, since we pursue person re-ID task, we expect the face does not play an important role. Thus, we will verify faces are not functionable in the proposed Celebrities-reID dataset. Second, state-of-the-art approaches will be harnessed to testify the challenge of clothes variation in our new dataset. Third, the effectiveness of the proposed benchmark approach will be verified on our dataset, and its generality will be testified on currently publicly available dataset concurrently. Finally, a comparison between the new benchmark method and other state of the arts will be made.

To fairly compare our results with the state of the arts, we use the same split over all the methods. Since the proposed Celebrities-reID is large, we fix the training and testing split instead of randomly repeating it for 10 or 20 times as in [2]. In training, 490 subjects (including the validation set) are used to train a model for measuring the similarity between two images in query and gallery, respectively. In testing, given a query image, all gallery images are used to calculate similarity scores. Then, a ranking process is performed to rank all gallery images according to their similarity to the query image. Sine each query image has multiple true matching in the gallery, both of the Cumulative Matching Characteristic (CMC) curve [9] and mean Average Precision (mAP) [10] are used to evaluate the performance. The CMC curve is utilized to show the probability that all gallery images appear in different ranking positions according to their similarity to query, which focuses on a retrieval precision. While, the mAP provides the overall performance when multiple true matching exists.

B. Experimental Setup

1) Person re-ID dataset: Two datasets are involved in our experiment. The first is the proposed Celebrities-reID dataset that used for challenging the task of clothes variation. The second is Market1501 [10] which is one of the most commonly used short-term (without clothes variation) person re-ID dataset that contains 12,936 training images and 19,732 testing images. The number of identities is 751 and 750 in

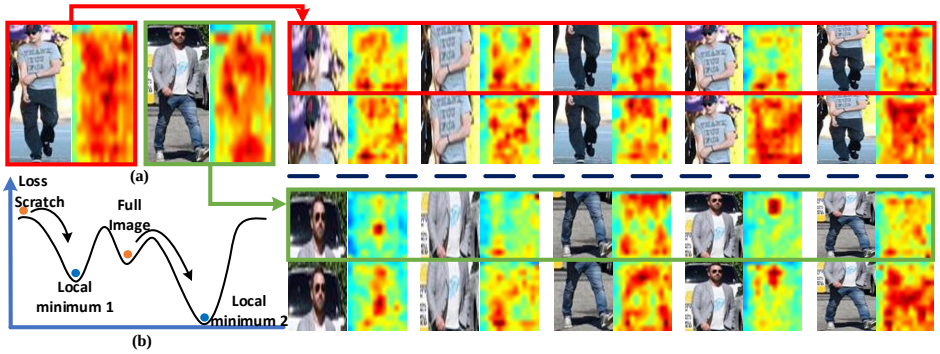

Fig. 4. (a) shows activation maps of two different full person images using the network optimized by the first step fine-tuning. The right figure provides five corresponding body parts of the two persons with parts activation maps obtained through the second step fine-tuning (the first and third rows) and learning from the scratch (the second and fourth rows), respectively. (b) Gives an intuitive understanding of the merit through our 2SF-BPart approach. A superior result (local minimum 2) can be observed by fine-tuning from the optimized network (learned by full images) to each body part.

training and testing respectively. Only the single query setting is utilized.

2) 2SF-BPart for evaluation: The Matconvnet [28] package is used to implement our benchmark. We adopt the 2S-IDE network [8] as our backbone because of its compelling performance on Celebrities-reID (see Table VI). In our first fine-tuning step, the full image is resized to 256×256 before being randomly cropped into 224×224. The ImageNet-trained resnet-50 is utilized to fine-tune the full images. We modify the fully-connected layer of resnet-50 to have 490 and 751 neurons for the proposed Celebrities-reID dataset and Market1501 respectively. The learning rate is set to 0.1 and decay to 0.01 after 40 epochs, we stop training after 60 epochs in the first fine-tuning step. The batchsize is set to 32. In the second fine-tuning step, we use the same batchsize and input size. Five body parts (see Fig. 3) are separately used to fine-tune the 2S-IDE model which optimized on the full image. In this stage, we set the learning rate to 0.1 and decay to 0.01 after 20 epochs. We stop training after 40 epochs. In training, the dropout is used after the GP layer and the SED layer (see Fig. 2) to reduce over-fitting with rate of 0.75 and 0.9 after the layers respectively. In addition, we follow the original design (see [8]) which sets the initial ratio between positive (belong to the same person) and negative (belong to different persons) pairs of inputs as 1:1, and it will be multiplied by a factor of 1.01 every epoch until it reaches 1:4.

Fig. 5. Face covered version of the proposed Celebrities-reID. We use black masks to explicitly hide the face which is detected by a state-of-the-art face detector S^3FD [29].

Methods	rank-1	mAP
Celebrities-reID		
2S-IDE	17.36%	9.85%
2S-IDE face covered	16.46%	9.09%
2S-IDE head	16.33%	9.45%
Market1501		
2S-IDE head	28.98%	13.21%

TABLE III
THE COMPARISON BETWEEN THE 2S-IDE MODEL DIRECTLY TRAINED ON THE FULL IMAGE (2S-IDE), ITS FACE BEING SPECIFICALLY COVERED (2S-IDE FACE COVERED), AND ONLY THE HEADS OF CELEBRITIES-REID AND MARKET1501 RESPECTIVELY. RANK-1 ACCURACY AND MAP ARE LISTED.

C. Effectiveness and Quality of the Proposed Dataset

1) Impact of face: The face cue is commonly unreliable in the realistic scenario of person re-ID. Therefore, we expect the impact of faces is intrinsically weak on our proposed dataset. To this end, we respectively train the 2S-IDE approach (the backbone of our benchmark) on the proposed Celebrities-reID dataset with full image and its face being specifically covered (see Fig. 5). It can be observed in Table III that the result of the 2S-IDE is on par with the 2S-IDE face covered in rank-1 accuracy (17.36% *vs.* 16.46%) and mAP (9.85% *vs.* 9.09%). Apparently, the face cue is not vital to the re-ID performance of Celebrities-reID. We further verify the effect of the head (including the hair and face) by directly training the head part only (see Fig. 3) using the 2S-IDE model after resizing the input size as the same as the full image. As a comparison, we also verify the effect of heads on Market1501 by using the same part partition approach. Table III shows the performance by only using the head for person re-ID on both datasets (2S-IDE head). It can be observed that the head cue of our proposed Celebrities-reID dataset is even more indistinguishable than Market1501, which deteriorates the recognition performance. This result indicates the proposed Celebrities-reID dataset is inherently challenging either on the

TABLE IV
COMPARISON BETWEEN OUR TWO-STEP FINE-TUNING STRATEGY AND DIRECTLY TRAINING EACH BODY PART FROM SCRATCH.

Methods	Training from the scratch		Two-step fine-tuning	
	rank-1	mAP	rank-1	mAP
2S-IDE baseline	-	-	17.36%	9.85%
2S-IDE head	15.12%	9.22%	**16.33%**	**9.45%**
2S-IDE up	4.51%	3.33%	**6.76%**	**3.62%**
2S-IDE low	5.30%	4.03%	**6.76%**	**4.16%**
2S-IDE head+up	**16.70%**	9.35%	16.01%	**10.00%**
2S-IDE up+low	7.78%	4.75%	**9.02%**	**4.92%**
Combination	23.69%	12.95%	**26.76%**	**14.01%**

TABLE V
COMPARISON OF USING DIFFERENT COMBINATIONS OF BODY PARTS AND WEIGHTS. THE SUBSCRIPT OF D IS CORRESPONDING TO DIFFERENT BODY PARTS IN FIG. 3

Methods	rank-1	mAP
$D_1 + 0.5 * D_2 + 0.5 * D_3$	21.76%	12.25%
$D_1 + 0.5 * D_4 + 0.5 * D_5$	22.21%	12.85%
$D_1 + D_2 + D_3 + D_4 + D_5$	23.45%	12.97%
$D_1 + 0.5 * D_2 + 0.5 * D_3 + D_4 + 0.5 * D_5$	**26.76%**	**14.01%**

clothes variation or the unreliable biometric trait, *i.e.*, face and head to person re-ID.

2) Challenges of the proposed Celebrities-reID Dataset: We attempt to harness several state-of-the-art approaches (**with publicly available source code which are suitable to our experiment**) to testify the challenge of the proposed Celebrities-reID dataset. Table VI lists ten methods we used on our dataset. Amongst them, LOMO+XQDA [19] and GOG$_{fusion}$+XQDA [20] are the traditional learning methods, and others are deep learning methods. The performance of the two traditional learning methods is inferior (*e.g.*, the best rank-1 accuracy is only 5.47%) to other deep learning methods since they heavily rely on the appearance information (*i.e.*, texture and color). The 2S-IDE achieves the best rank-1 accuracy (17.36%) and mAP (9.85%). It can be observed that there is a large margin between the performance of Celebrities-reID and Market1501 (*e.g.*, the PCB [26] achieves 92.30% rank-1 accuracy on Market1501 but only obtains 16.69% on Celebrities-reID). This result verifies that our new dataset is extremely challenging to the state of the arts because of the drastic appearance changes caused by clothes variation.

3) Quality of the new benchmark: To testify the quality of the newly proposed benchmark, we not only verify its performance on our proposed Celebrities-reID dataset (with clothes variation) but also testify its generality to the widely used person re-ID dataset Market1501 (without clothes variation). Table VI shows that our 2SF-BPart approach achieves the best performance on Celebrities-reID with 26.76% rank-1 accuracy and 14.01% mAP. This result concurrently outperforms the second best 2S-IDE method by a large margin in rank-1 accuracy (26.76% *vs*. 17.36%) and mAP (14.01% *vs*. 9.85%). In addition, our 2SF-BPart method also achieves a competitive result on Market1501 with 91.21% rank-1 accuracy and 77.17% mAP, which demonstrates its generality to different person re-ID scenarios (with or without clothes variation).

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART METHODS.

Methods		Celebrities-reID		Market-1501	
		rank-1	mAP	rank-1	mAP
Traditional Learning Methods					
LOMO	CVPR15 [19]	4.85%	3.40%	-	-
GOG	CVPR16 [20]	5.47%	4.95%	-	-
Deep Learning Methods					
SVDNet	CVPR17 [25]	13.16%	8.82%	82.30%	62.10%
Part-aligned	CVPR17 [22]	8.46%	6.05%	81.00%	63.40%
Spindle Net	CVPR17 [21]	7.49%	-	76.90%	-
1S-IDE	TOMM17 [8]	12.06%	7.44%	73.69%	51.48%
2S-IDE	TOMM17 [8]	17.36%	9.85%	83.31%	65.49%
PSE	CVPR18 [30]	15.19%	8.73%	87.70%	69.00%
HACNN	CVPR18 [34]	16.2%	11.5%	91.20%	75.70%
PCB	ECCV18 [26]	16.69%	9.03%	**92.30%**	**77.40%**
Our 2SF-BPart		**26.76%**	**14.01%**	91.21%	77.17%

We also conduct an ablation study by directly training each part from the scratch. In this experiment, we use the same 2S-IDE model as our backbone network; the only difference is that we skip the process of training full images. Table IV lists the results. By using our two-step fine-tuning strategy, superior results can be achieved on the five body parts, except the rank-1 accuracy of the head+upperbody which is marginally lower than directly training from scratch (16.01% *vs*. 16.70%). Using the Eq. 1, we integrate all the parts to produce the final result, compared with training from scratch, our 2SF-BPart approach achieves better rank-1 accuracy (26.76% *vs*. 23.69%) and mAP (14.01% *vs*. 12.95%) simultaneously.

In addition, we change the number and the weight of different body parts to compare the performance using different combinations. Table V lists the results. We can observe that all variants are experimentally validated as inferior to the combination by using the five parts with 0.5 weights on upperbody, upperbdoy + lowerbody, and lowerbody. Other adaptive combinations may achieve better performance compared with our empirically setting (*i.e.*, Eq. 1).

D. Performance of the New Benchmark Approach

Table VI lists the overall experiment results on two datasets. The proposed 2SF-BPart achieves state of the art on our Celebrities-reID dataset compared with other nine approaches by using their available source code. Our method outperforms all existing methods with a large improvement of $+9.40\%$ (rank-1) and $+4.16\%$ (mAP) compared with the 2nd best method 2S-IDE. In the meantime, we achieve 91.21% rank-1 accuracy and 77.17% mAP on the widely used Market1501 dataset. This result is competitive with other state-of-the-art methods. It can be observed that the PCB [26] obtains the best performance on Market1501 with 92.30% in rank-1 accuracy and 77.40% in mAP, which is slightly higher than our 2SF-BPart approach. However, the PCB is jointly trained on all body parts, which does not pay attention to the discrepancy between each part when a person does change clothes in the long-term person re-ID scenario. Although it performs slightly

better than our benchmark method on Market1501, our method significantly outperforms it on our Celebrities-reID dataset (26.76% *vs*. 16.69% in rank-1 accuracy and 14.01% *vs*. 9.03% in mAP). Noting that the Spindle Net [21] is firstly trained on eleven existing short-term person re-ID datasets (all of them without clothes variation) and then fine-tuned on each target dataset. Actually, because of the large bias between Celebrities-reID and the other short-term datasets, Spindle Net only achieves 1.6% rank-1 accuracy when Celebrities-reID is combined with all the other short-term datasets in training. Therefore, we try to only use Celebrities-reID to train the Spindle Net, which achieves a better result shown in Table VI (7.49% in rank-1 accuracy). In addition, the performance of the 2S-IDE (83.31% rank-1 accuracy) is higher under our experimental setting (see Section V-B2) compared with the result reported in [8] (79.51% rank-1 accuracy) on Market1501.

VI. Conclusions

In this paper, we introduce a new dataset which is particularly used for the clothes variation study in the long-term person re-ID scenario. The new dataset will enable research possibilities in multiple directions, *e.g.*, large scale and long-time gap person re-ID where a person undergoes a great possibility to change clothes. To tackle the challenge of clothes variation, a benchmark approach is proposed by utilizing a two-step fine-tuning strategy on human body parts. Extensive evaluations are conducted on the proposed Celebrities-reID dataset, and a commonly used short-term dataset Market1501 where person does not change their clothes. Superior experiment results are achieved to demonstrate the effectiveness and the generality of the proposed benchmark approach.

Acknowledgment

We thank Qian Liu, Xingyuan Ding, Chunxi Gui, and Shengdong Li for working on the new dataset. This research is supported by an Australian Government Research Training Program Scholarship.

References

[1] Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person re-identification. Springer (2014)

[2] Li, W., Zhao, R., Xiao, T. and Wang, X., 2014. Deepreid: Deep filter pairing neural network for person re-identification. In CVPR, pp. 152-159.

[3] Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z. and Zhang, J., 2019. Multi-Pseudo Regularized Label for Generated Data in Person Re-Identification. In IEEE TIP, 28(3), pp.1391-1403.

[4] Huang, Y., Sheng, H. and Xiong, Z., 2016. Person re-identification based on hierarchical bipartite graph matching. In ICIP, pp. 4255-4259.

[5] Huang, Y., Sheng, H., Zheng, Y. and Xiong, Z., 2017. DeepDiff: learning deep difference features on human body parts for person re-identification. In Neurocomputing, 241, pp.191-203.

[6] Huang, Y., Sheng, H., Liu, Y., Zheng, Y. and Xiong, Z., 2015. Person re-identification by unsupervised color spatial pyramid matching. In KSEM, pp. 799-810.

[7] Zheng, Z., Zheng, L. and Yang, Y., 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In ICCV, pp. 3774-3782.

[8] Zheng, Z., Zheng, L. and Yang, Y., 2018. A discriminatively learned cnn embedding for person reidentification. In ACM TOMM, 14(1), p.13.

[9] Gray, D., Brennan, S. and Tao, H., 2007. Evaluating appearance models for recognition, reacquisition, and tracking. International Workshop on Performance Evaluation for Tracking and Surveillance, Vol. 3, No. 5, pp. 1-7.

[10] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q., 2015. Scalable person re-identification: A benchmark. In ICCV, pp. 1116-1124.

[11] Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O. and Radke, R.J., A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. In IEEE TPAMI, (1), pp.1-1.

[12] Wei, L., Zhang, S., Gao, W. and Tian, Q., 2018. Person transfer gan to bridge domain gap for person re-identification. In CVPR, pp. 79-88.

[13] Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L. and Murino, V., 2012. Re-identification with rgb-d sensors. In ECCV, pp. 433-442.

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

[15] Haque, A., Alahi, A. and Fei-Fei, L., 2016. Recurrent attention models for depth-based person identification. In CVPR, pp. 1229-1238.

[16] Gou, M., Zhang, X., Rates-Borras, A., Asghari-Esfeden, S., Camps, O.I., Sznaier, M., 2016, Person re-identification in appearance impaired scenarios. In BMVC.

[17] Zhang, P., Wu, Q., Xu, J. and Zhang, J., 2018, March. Long-Term Person Re-identification Using True Motion from Videos, 2018, In WACV, pp. 494-502.

[18] Yu, S., Tan, D. and Tan, T., 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In ICPR, Vol. 4, pp. 441-444.

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

[20] Matsukawa, T., Okabe, T., Suzuki, E. and Sato, Y., 2016. Hierarchical gaussian descriptor for person re-identification. In CVPR, pp. 1363-1372.

[21] Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X. and Tang, X., 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In CVPR, pp. 1077-1085.

[22] Zhao, L., Li, X., Zhuang, Y. and Wang, J., 2017. Deeply-Learned Part-Aligned Representations for Person Re-identification. In ICCV, pp. 3239-3248.

[23] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In CVPR, pp. 770-778.

[24] Ristani, E., Solera, F., Zou, R., Cucchiara, R. and Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, pp. 17-35.

[25] Sun, Y., Zheng, L., Deng, W. and Wang, S., 2017. SVDNet for Pedestrian Retrieval. In ICCV, pp. 3820-3828.

[26] Sun, Y., Zheng, L., Yang, Y., Tian, Q. and Wang, S., Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline), 2018, In ECCV.

[27] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., Realtime multi-person 2d pose estimation using part affinity fields, 2017, In CVPR.

[28] Vedaldi, A., Lenc, K.: Matconvnet: Convolutional neural networks for matlab, 2015, In ACMMM, pp. 689-692.

[29] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X. and Li, S.Z., 2017. S^3FD: Single Shot Scale-Invariant Face Detector. In ICCV, pp. 192-201.

[30] Sarfraz, M.S., Schumann, A., Eberle, A. and Stiefelhagen, R., 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In CVPR, Vol. 7, pp. 8.

[31] Zhang, W., Ma, B., Liu, K. and Huang, R., 2017. Video-based pedestrian re-identification by adaptive spatio-temporal appearance model. In IEEE TIP, Vol. 26(4), pp.2042-2054.

[32] Zhang, W., Li, Y., Lu, W., Xu, X., Liu, Z. and Ji, X., 2018. Learning intra-video difference for person re-identification. In IEEE TCSVT.

[33] Ben, X., Gong, C., Zhang, P., Jia, X., Wu, Q. and Meng, W., 2019. Coupled Patch Alignment for Matching Cross-view Gaits. In IEEE TIP.

[34] Li, W., Zhu, X. and Gong, S., 2018, February. Harmonious attention network for person re-identification. In CVPR, Vol. 1, pp. 2.