Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification

Yan Huang, Jingsong Xu, Qiang Wu, *Senior Member, IEEE*, Yi Zhong, Peng Zhang, Zhaoxiang Zhang, *Senior Member, IEEE*

***Abstract*—Current person re-identification (re-ID) works mainly focus on the short-term scenario where a person is less likely to change clothes. However, in the long-term re-ID scenario, a person has a great chance to change clothes. A sophisticated re-ID system should take such changes into account. To facilitate the study of long-term re-ID, this paper introduces a large-scale re-ID dataset called "Celeb-reID" to the community. Unlike previous datasets, the same person can change clothes in the proposed Celeb-reID dataset. Images of Celeb-reID are acquired from the Internet using street snap-shots of celebrities. There is a total of 1,052 IDs with 34,186 images making Celeb-reID being the largest long-term re-ID dataset so far. To tackle the challenge of cloth changes, we propose to use vector-neuron (VN) capsules instead of the traditional scalar neurons (SN) to design our network. Compared with SN, one extra-dimensional information in VN can perceive cloth changes of the same person. We introduce a well-designed ReIDCaps network and integrate capsules to deal with the person re-ID task. Soft Embedding Attention (SEA) and Feature Sparse Representation (FSR) mechanisms are adopted in our network for performance boosting.**

Experiments are conducted on the proposed long-term re-ID dataset and two common short-term re-ID datasets. Comprehensive analyses are given to demonstrate the challenge exposed in our datasets. Experimental results show that our ReIDCaps can outperform existing state-of-the-art methods by a large margin in the long-term scenario. *The new dataset and code will be released to facilitate future researches.*

***Index Terms*—person re-identification, Long-term scenario, cloth change, vector-neuron capsules.**

I. Introduction

Person re-identification (re-ID) is a typical computer vision problem which aims to associate person images captured by different camera views. It is an important research topic and has a wide range of applications such as cross-camera object tracking, target re-acquisition, *etc* [1]. In past years, several datasets have been proposed and contribute significantly to the community, *e.g.*, VIPeR [2], CUHK03 [3], Market1501 [4], DukeMTMC-reID [5], *etc*. However, most of them are based on the assumption that each person appears under one camera will re-appear under another one in a short period (*e.g.*, less than 30 minutes). Under this assumption, a person is less likely to change clothes. We define this kind of scenario as "short-term" re-ID. On the contrary, if a person appears again after a long-time gap (*e.g.*, more than one day), the chance of changing clothes or carrying different objects will become large. This type of scenario can be regarded as "long-term" re-ID. This paper will investigate the feasibility of the existing datasets for long-term person re-ID in terms of the definition above. The long-term re-ID is a more challenging case commonly seen in large-scale video security surveillance.

According to the existing research outcomes and potential applications in practice, a dataset is suitable for the long-term person re-ID researches should satisfy the following requirements: 1) a large number of person IDs, 2) highly diverse environment/backgrounds, 3) multiple camera views, 4) various shooting conditions (*e.g.*, illumination and resolution), and 5) highly dynamic appearance of each person. By investigating the existing datasets regardless of indoor or outdoor datasets, there are mainly two methods for collecting datasets: 1) collecting data in a free setup environment with non-collaborative people [2], [3], [4], [5]; 2) collecting data in a constrained environment with collaborative people (*e.g.*, actors) [6], [7], [8], [9]. Either method has its advantages but also clear problems when the aforementioned requirements are to be satisfied.

The first method can best align with the situations of a real surveillance environment. However, in practice, the data annotation is extremely difficult even just for a few thousand people ID. In order to allocate the same ID to the same person, the annotation staff has to rely on the people face information (if that is available) to give a typical ID for the same person when they appear under different cameras. If the face information is not available due to poor image quality or poor camera view (*e.g.*, back view), the samples have to be discarded or marked based on experiences. For the case of short-term person re-ID, such difficulty is still manageable, where the annotation staff may use cloth information to match people across cameras. However, in the case of long-term person re-ID, the assumption above does not exist. Thus, the annotation for long-term person re-ID purpose is impossible. That is the main reason that current datasets such as VIPeR [2], Market1501 [4], and CUHK03 [3] are not suitable for long-term person re-ID for the requirement of the highly dynamic appearance of each annotated person although they are collected in a free control (at least less constrained) environment.

The second method can best simulate many extremely difficult cases, such as very unusual camera views and cloth changes. For example, a person re-ID dataset is proposed in [9] that demonstrates the scenarios of person re-ID in the case of

Yan Huang, Jingsong Xu, Qiang Wu, and Peng Zhang are with the Global Big Data Technologies Centre (GBDTC), School of Electrical and Data Engineering, University of Technology Sydney, Australia.

Yi Zhong is with with the School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China. (Email: yi.zhong@bit.edu.cn)

Zhaoxiang Zhang is with the Research Center for Brain-Inspired Intelligence, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

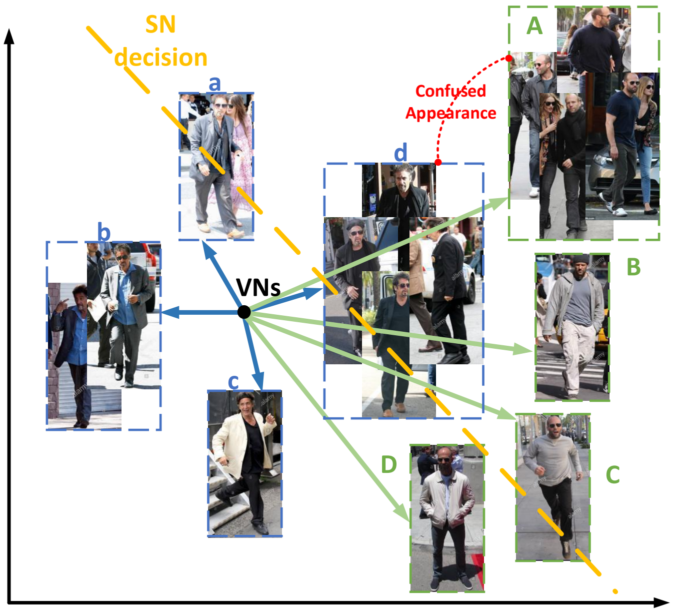


Fig. 1. SN *vs.* VN capsule in person re-ID. The (a)-(d) and (A)-(D) are images belonging to two different IDs in our Celeb-reID dataset. The (d) and (A) include two persons with similar dark clothes. The VN capsules use the length of the vector to represent different IDs, while its orientations are used to perceive different types of clothes. With two-dimensional perception capability, different IDs can be distinguished easier by using the length of capsules. Instead, typical SN cannot make a decision between confused appearance (*e.g.*, some images in (d) are regraded as the ID in the green bounding box). Best viewed in color.

clothes change. It tries to recognize different people through gait information across view angles. However, the main issue of the datasets conducted by the second method is the scale, including the number of people ID, diversity of environment, and the number of camera view. This is mainly because building-up such dataset in a mimic environment relies on a certain number of actors who follow the predefined way to complete the shooting. In practice, it is not manageable to have a few thousand actors. Moreover, shooting usually is completed in some specific places based on predefined design. So it is hard to mimic real surveillance situations.

Based on the experiences of existing datasets, the key challenging problems to build a dataset, particularly, suitable for long-term person re-ID are: 1) sufficient large number of people IDs, 2) dynamic shooting with true environments, 3) various clothes on each person. In this paper, a new dataset called "Celeb-reID" is proposed, which can tackle the aforementioned challenging problem. We name it Celeb-reID because all images are collected using celebrities' street snap-shots. We choose celebrities as our target because there are tremendous resources of celebrity images on the Internet. We use street snap-shots of celebrities because they are more relevant to the real-life scenario. In this way, we can acquire a large number of people ID. Moreover, celebrities are widely seen and taken photos in various scenarios so we may obtain their images under different environments. The most important thing is that these celebrities normally wear different clothes in different scenes. It well satisfies the requirements in terms of a highly dynamic environment in the scenarios of security surveillance for long-term person re-ID.

Since the reliability of appearance (*e.g.*, color and texture) is reduced greatly due to the change of cloth, current re-ID approaches, which mainly rely on the appearance, may not be suitable for the long-term scenario. However, existing person re-ID approaches are mainly designed on the condition of unchanged clothes [10], [11], [12], [13], [14], [15], [16]. Although these approaches have achieved compelling performance for short-term re-ID, we argue that simply concentrate on appearance and overlook the change of clothes may not be suitable for the long-term re-ID scenario. Specifically, existing approaches try to distinguish the inter-class clothes changes, but do not pay attention to the capability of perceiving and differentiating the large intra-class cloth changes. Therefore, existing re-ID methods are hard to directly apply to the long-term re-ID scenario.

Compared with traditional approaches which use the Scalar Neuron (SN), we propose to use Vector-Neuron (VN) capsules [17] to perceive the cloth change of the same person. In common CNNs, the value of each scalar reflects the likelihood of a neuron belonging to an existing people ID. However, if the cloth is changed, the one-dimensional SN cannot further perceive the cloth change information. Therefore, we integrate the two-dimensional VN capsules in our network. Our idea is inspired by [17]: the length (the first dimension) of each capsule (denoted as $C_{\mathcal{L}}$) expresses the "existence of the entity"; the orientation (the second dimension) of each capsule (denoted as $C_{\mathcal{O}}$) is forced to represent "the properties of the entity". In our case, we expect $C_{\mathcal{L}}$ to reflect the likelihood of an existing people ID while $C_{\mathcal{O}}$ represents different clothes of the same ID. Fig. 1 shows the difference between SNs and VNs. With two-dimensional perception capability, VN capsules can discriminate different IDs (through the length of capsules) with confused appearance (perceived by the orientation of capsules), while the SN may make a wrong decision. We call our model as "ReIDCaps". Unlike the original application of capsules in [17] (*e.g.*, learning the property of affine transformation for handwritten digits recognition), our ReIDCaps network considers the cloth change of the same person should be further perceived and represented by $C_{\mathcal{O}}$. In the meantime, $C_{\mathcal{L}}$ is used as common SN to distinguish different people ID. Finally, to further enhance the discrimination and generalization power of ReIDCaps, we integrate Soft Embedding Attention (SEA) mechanism and Feature Sparse Representation (FSR) mechanism in our network.

In general, the main contributions of this paper can be summarized in three-fold:

- We propose a large-scale long-term person re-ID dataset Celeb-reID. Approximately more than 70% of the images of each person are in different clothes. There are 1,052 IDs and 34,186 person images in Celeb-reID to make it is the largest cloth change re-ID dataset so far.
- A well-designed ReIDCaps network is introduced to deal with the challenge of cloth change in long-term person re-ID. In addition, we integrate SEA and FSR mechanisms to enhance the discrimination and generalization power of our network.

- A comprehensive evaluation is given to verify the effectiveness of our model. We outperform state-of-the-art methods by a large margin (more than 10% in rank-1 accuracy) in the long-term re-ID scenario. In addition, the robustness of our model is further verified on two traditional short-term re-ID datasets. Finally, a robustness score is defined to compare the performance of re-ID approaches across different re-ID scenarios.

This paper is organized as follows. We first review related works in Section II. In Section III, we introduce the Celeb-reID dataset. Then, the architecture of the ReIDCaps network is given in Section IV. Experiments are given in Section VI. The conclusion is in Section VII.

II. Related Work

In this section, we will review the current publicly available person re-ID datasets, the existing person re-ID approaches, and the capsule network.

A. Existing Person Re-ID Datasets

In past years, several person re-ID datasets are proposed. A majority of them concentrate on the short-term scenario. We will first discuss these short-term re-ID datasets in this section. In addition, some cloth change re-ID datasets are introduced.

Existing short-term person re-ID datasets. Several person re-ID datasets have been proposed along with the development of re-ID researches. In general, most of them belong to the short-term scenario. These dataset do not consider the change of clothes. These datasets are mainly proposed to facilitate the study of disturbances produced by the changes of illuminations, poses, viewpoints, background clutters, and occlusions. Amongst them, VIPeR [2], CUHK03 [3], Market1501 [4], DukeMTMC-reID [5], *etc.*, are most widely used datasets. Compelling performance has achieved on these datasets.

However, current approaches are mainly designed for these short-term scenarios, which may not be suitable for the long-term scenario. To achieve a sophisticated re-ID system, the cloth change problem should be explicitly considered for real-world applications. Unfortunately, existing short-term re-ID datasets do not involve such characteristics. Therefore, in this paper, we introduce a long-term person re-ID dataset to facilitate future researches. Tab. I shows the comparison between our Celeb-reID dataset and other common short-term re-ID datasets. It can be observed that the scale of our dataset (particularly the number of images) is close to two popular large-scale re-ID datasets: Market1501 [4] and DukeMTMC-reID [5]. Currently, one huge-scale short-term re-ID dataset MSMT17 [18] has been proposed. Compared with other short-term re-ID datasets, MSMT17 contains the largest number of IDs and images with more complex lighting variations. However, a person in this dataset also does not change clothes. We do not involve MSMT17 in the comparison because our motivation is to introduce a new long-term re-ID dataset, but not to compare the scale with any re-ID datasets.

To highlight certain special cases, a bike-person re-ID dataset (BPReid) has been proposed [19]. In BPReid, a majority of persons are riding bikes rather than walking. The BPReid dataset demonstrates a valuable research case for the short-term scenario. However, it also does not consider the change of clothes for the long-term re-ID scenario.

Existing cloth change re-ID datasets. To the best of our knowledge, five small-scale datasets consider the cloth change problem for person re-ID. They can be categorized into two types: RGB-D and normal RGB video-based datasets. To handle the challenge of cloth changes, the depth information has been leveraged to extract additional 3D soft-biometric beyond the RGB color cue. Accordingly, RGB-D datasets such as PAVIS [6], BIWI [7], IAS-Lab [7], and DPI-T [8] have been proposed using the Kinect camera under controlled environments. We agree that these datasets are helpful on proof of concept by exploiting 3D information using depth information. However, RGB-D is not the mainstream in surveillance applications. The proposed dataset is aligned with the mainstream of person re-ID dataset setting for supporting the research on long-term re-ID.

On the other hand, a video-based reID dataset [9] was proposed to use gait cue for person re-ID. However, the scale of this dataset is also too small, and the environment is controlled under an indoor camera. Tab. I shows the comparison between our Celeb-reID dataset and other datasets with cloth changes. The number of ID in our Celeb-reID dataset outperforms other cloth change datasets by a large margin. In addition, these cloth change datasets are hard to satisfy the requirement of practical long-term re-ID datasets since they collected under a constrained environment rather than a highly diverse environment with multiple camera views.

B. Person Re-ID Approaches

Deep learning has achieved great success in solving the person re-ID problem [10], [11], [12], [13], [14], [15], [20], [21], [22], [23], [16]. However, current CNN-based re-ID approaches are mainly focused on the short-term scenario. According to different granularities, we simply categorize previous CNN-based re-ID methods into two types: coarse- and fine-grained learning approaches.

In coarse-grained learning methods, a CNN network receives the whole images as inputs. Amongst them, one of the classic ones is the ID-Discriminative Embedding (IDE) model [10]. The IDE model considered person re-ID problem as an ID classification task. The ImageNet-trained resnet-50 was used as a feature extractor connecting with a cross-entropy loss to classify different IDs in training. Following this work, an IDE+ model was proposed to further improve the re-ID accuracy by adding bottleneck layers [11], [12]. Moreover, in order to fully exploit middle and high semantic level features, [13] extracted and fused both middle- and high-level features in the ID classification architecture. Beyond classification models, a two-stream CNN network was proposed by considering the classification and verification signals simultaneously [14]. A multi-level factorization model that factorized the visual appearance of a person into latent discriminative factors was given in [15].

Currently, many fine-grained learning approaches which utilize the body part cue are widely used in person re-ID [21],

TABLE I
COMPARISON BETWEEN CELEB-REID AND OTHER PERSON RE-ID DATASETS.

Dataset	#IDs	#Images or #sequences	Environments	Type of cameras
Traditional Short-Term Person Re-ID Datasets (images)				
VIPeR [2]	632	1,264	indoor+outdoor	common RGB cameras
CUHK03 [3]	1467	13,164	campus	common RGB cameras
Market1501 [4]	1501	32,217	campus	common RGB cameras
DukeMTMC-reID [5]	1404	36,441	campus	common RGB cameras
Cloth Change Datasets (video sequences)				
PAVIS [6]	79	316	under controlled	Kinect
BIWI [7]	50	50	under controlled	Kinect
IAS-Lab [7]	11	33	under controlled	Kinect
DPI-T [8]	12	300	under controlled	Kinect
RGB Video-Based Dataset [9]	30	240	under controlled	common RGB cameras
Ours (images)				
Celeb-reID	**1,052**	**34,186**	**highly diverse**	**common RGB cameras**

[22], [23], [20], [16]. Huang *et al.* [21] proposed three subnets to learn differences between corresponding body parts on original images, feature maps, and regional spatial variations, respectively. Sah *et al.* [22] introduced a two-stream network in which one stream was used for appearance map extraction, and the other one was used for body part map extraction. Part-aligned feature maps were obtained from the corresponding local appearance. Li *et al.* [23] formulated a novel Harmonious Attention CNN (HACNN) to joint learning of soft pixel attention and hard regional attention on different local body regions. Sun *et al.* [20] considered the content consistency within each body part for the precise part location. The outliers of different body parts were re-assigned in [20], and the refined body parts resulted in performance gains. Recently, a multiple granularities network was proposed to combine global and local body information which integrated different regional feature representations for person re-ID [16].

In addition to these deep learning methods, some metric learning approaches also shown their feasibility in person re-ID. Zhang *et al.* [24] embedded constraints on linear transformation matrix and proposed a weight learning strategy to learn the relationships between different variables. Sun *et al.* [25] introduced a pair of transformation matrices to capture the intrinsic relationship of the same person under different camera views for video-based person re-ID.

In experiments, we employ the above-mentioned deep learning methods on our new dataset Celeb-reID to verify the robustness and feasibility of existing methods in the long-term re-ID scenario. A comprehensive evaluation is given to demonstrate the challenge exposed in our Celeb-reID dataset.

C. Capsule Network

In 2017s, capsule network was proposed for handwritten digits recognition with a dynamic routing mechanism [17]. Capsule network has been successfully used in many computer vision tasks, *e.g.*, object segmentation [26], data generation [27], object classification [17], *etc.* Sabour *et al.* used a group of neurons (vector) to represent a capsule whose length expressed the existence of the entity and orientation represented the properties of the entity. The proposed DigitCaps model in [17] was moderately robust to small affine transformations of the training data. In 2018s, another capsule-based network was introduced that performed the routing-by-agreement dynamic routing mechanism using expectation maximization (EM) [28]. Unlike the previous work [17], the newly proposed capsule in [28] is represented by pose matrices. These capsules are specifically used to define rotations and translations of objects for viewpoint changes. Our Celeb-reID dataset is mainly suffered from the impact of cloth changes. Compared with the pose matrix capsule, the vector-based capsule proposed in [17] should be more suitable for our case since it can perceive the properties cue of each entity. Therefore, our ReIDCaps model is mainly motivated by DigitCaps [17]. Beyond DigitCaps, ReIDCaps integrates the ImageNet-trained CNN to extract the low-level visual representations since CNN pre-trained on ImageNet has performed promisingly on many re-ID tasks. To the best of our knowledge, this is the first work that integrates capsules with the ImageNet-trained CNN on complex data, particularly for person re-ID.

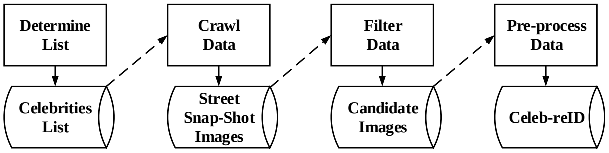


Fig. 2. The pipeline of our data acquisition. Four main steps are included.

III. CELEB-REID DATASET

In this section, we will introduce our Celeb-reID dataset. Fig. 2 shows the pipeline of the data acquisition process. It can be summarized into four steps:

1) **Determine Name List.** Since we use the street snap-shots of celebrities to build our dataset, the first step is to determine the name list of celebrities. We initially collect 2,600 popular celebrities' name from all over the world on the Internet.
2) **Crawl Data.** Given the name of celebrities, we crawl data of each celebrity on Google, Bing, and Baidu Images using keywords: name + street snap-shot (*e.g.*, Justin Bieber street snap-shot) with usage rights of free

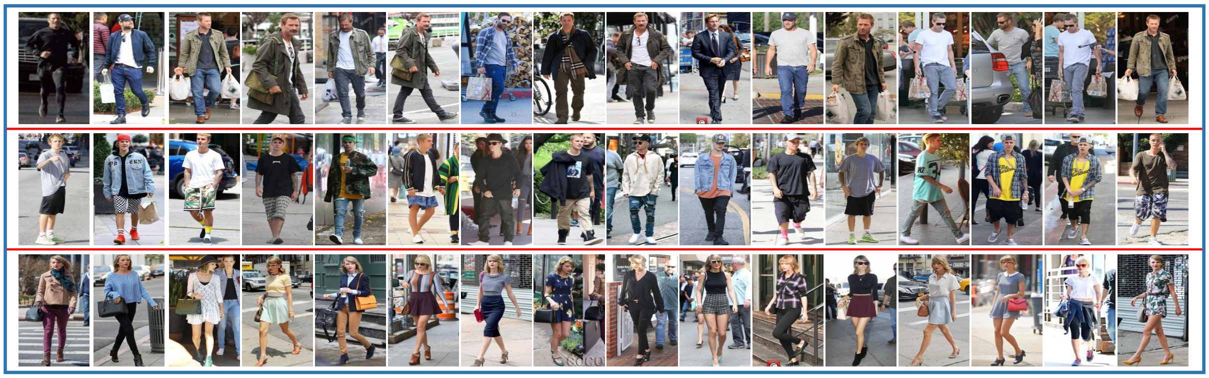


Fig. 3. The three rows represent three different IDs in our Celeb-reID dataset. For each ID, a great number of cloth changes can be found.

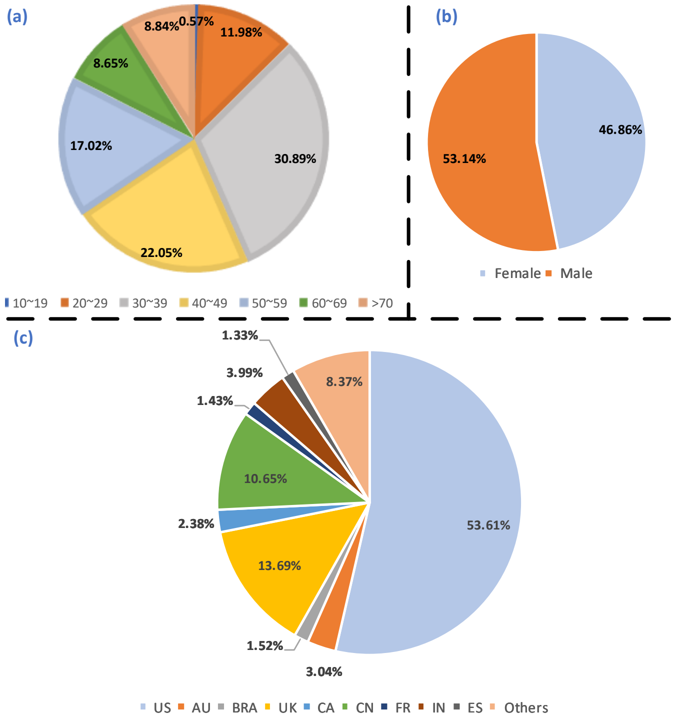


Fig. 4. Statistic information of our Celeb-reID dataset. (a), (b), and (c) respectively show the distributions of age, gender, and nationality.

to use and share. The first 100 images of each celebrity are crawled. Finally, we obtain 2,600×100 candidate street snap-shot images.

3) **Filter Data.** In this step, we filter images from the crawled images to construct a candidate image set. The annotation staff verifies the identity of images for each celebrity. Any wrong returned result is discarded.
4) **Pre-process Data.** The last step is to crop the bounding box of person body from the original image. We use Mask-RCNN [29] to detect the bounding box. Pixel paddings from the original image are used to ensure that the ratio of height and width of each image is identical to 2:1. We resize the final image size to 256×128. Finally, 1,052 celebrities with a total of 34,186 images are retained.

As the traditional setting of person re-ID datasets, we split Celeb-reID into three subsets, including training, gallery, and query (see Tab. II). The query and gallery sets are used for testing. Fig. 3 shows three IDs in our dataset. It is clear to see that the cloth change brings about large appearance changes. Notably, a person may wear the same cloth twice. Specifically, more than 70% of the images of each person show different clothes on average. In addition, people commonly show the front view or profile. Only a few back view images are included in our dataset. This is because we can only crawl a few back view street snap-shots on the Internet. Fig. 4 gives the statistic information of our dataset, including the distribution of age, gender, and nationality of celebrities.

TABLE II
DATA SPLIT OF OUR CELEB-REID DATASET. IN THE TESTING SET, AROUND 30% OF IMAGES OF THE 420 IDS BELONG TO THE QUERY SET, THE OTHER 70% OF IMAGES BELONG TO THE GALLERY SET.

split	training	testing		total
subsets	training	query	gallery	total
#ID	632	420	420	1,052
#images	20,208	2,972	11,006	34,186

IV. ARCHITECTURE OF REIDCAPS

Our ReIDCaps can be divided into three main modules: **1) Visual feature extraction module**: We adopt the ImageNet-trained model to extract the low-level visual features of each image. These features are then fed into the capsule layers to perceive the ID and dressing information. This step follows the common practice of Capsule network in [17] which uses several CNN layers to extract visual features of image before these features forwarding to the capsule layers. **2) ID and dressing perception module**: We adopt capsule layers to perceive the ID and dressing information of images. The ID information is mainly perceived by the length of each VN. The dressing information of each person is perceived by the orientation of each VN. **3) Auxiliary module**: The auxiliary modules are used to further improve the discriminability of features learned from our ReIDCaps modules. The architecture of our ReIDCaps network is shown in Fig. 5. This section will introduce each module in details.

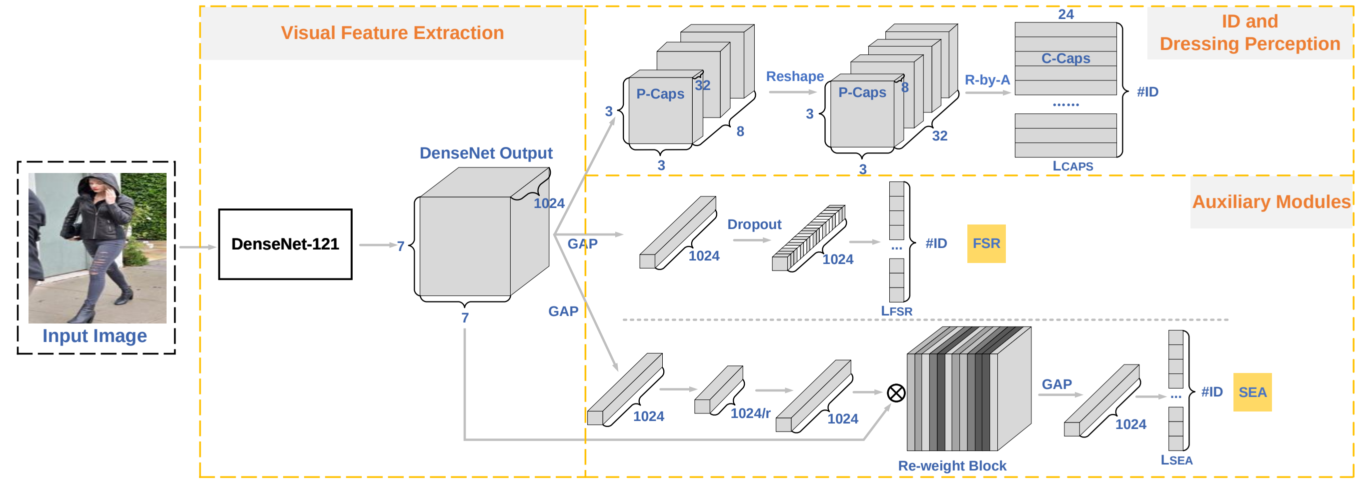


Fig. 5. Architecture of the proposed ReIDCaps network. Given an input image, an ImageNet-trained CNN backbone network (*i.e.*, DenseNet-121 [30]) is used to extract low-level visual features. The output of the backbone network is fed to three branches, including capsule modules (ID and dressing perception), FSR and SEA (two auxiliary modules).

A. *Visual Feature Extraction Module*

Given an input image I_n^x, where n is the index of an ID, x represents the x-th image of ID n, an ImageNet-trained CNN backbone network is used to extract low-level visual features from the I_n^x. We select the Densenet-121 [30] as the backbone network in our experiment[1]. We denote the output of the backbone as $\mathcal{O}(I_n^x) \in \mathbb{R}^{7\times7\times1024}$. The rest parts of our ReIDCaps are divided into three branches: Capsule layers (main), the FSR, and the SEA. We denote the loss of the three branches as $\mathcal{L}_{CAPS}$, $\mathcal{L}_{FSR}$, and $\mathcal{L}_{SEA}$ respectively. Amongst the three branches, FSR and SEA are two auxiliary branches. Both FSR and SEA dedicate to enhancing the performance of CNN-based visual feature. Then, these well-learned features can be utilized by the VN-based layers in a more efficient way. The objective function of our ReIDCaps network is given as follows:

$$\mathcal{L} = \mathcal{L}_{CAPS} + \gamma * (\mathcal{L}_{FSR} + \mathcal{L}_{SEA}), \tag{1}$$

where γ is used to balance the weight between contributions raised from capsule layers and auxiliary modules. We will respectively introduce the three different branches in details:

B. *ID and Dressing Perception Module*

The Capsule layers. We propose to use VN capsules to learn the property of cloth changes. In training, we expect the length of vector capsule $C_{\mathcal{L}}$ to reflect the likelihood of an existing people ID while its orientation $C_{\mathcal{O}}$ represents different types of clothes of the same people ID. To achieve this, we adopt two different capsule-based layers after $\mathcal{O}(I_n^x)$, including a Primary Capsules (P-Caps) layer and a Classification Capsules (C-Caps) layer [17]. Both layers are proposed in [17] for digital recognition. But in our design, we change the parameter setting on both layers to adapt the re-ID task and the ImageNet-trained CNN architecture. Specifically, given $\mathcal{O}(I_n^x)$, eight 32-channel convolutional operations (kernel size 2×2 and stride 2) are used to construct the P-Caps. Then, a reshaped operation is used to concatenate the corresponding channel of each block (8 blocks in total) in P-Caps. After that, we can get 288 ($3\times3\times32$) VN capsules with 8D in P-Caps layer. Finally, a non-linear Squashing Function is used to ensure the length of each VN capsule being normalized:

$$v_k^{8D} = \frac{\left\|v_k^{8D}\right\|^2}{1+\left\|v_k^{8D}\right\|^2}\cdot\frac{v_k^{8D}}{\left\|v_k^{8D}\right\|}, \tag{2}$$

where v_k^{8D} represents k-th 8D VN capsule in P-Caps, $k \in [1, 288]$.

The C-Caps layer is followed by the P-Caps layer. There are N ID capsules in the C-Caps layer; N represents the number of ID in the training set. Each ID capsule in C-Caps is the combination of all VN capsules in the P-Caps layer. Given an 8D VN v_k^{8D} in P-Caps, we first map its dimension to 24D by:

$$v_k^{24D} = W_k \cdot v_k^{8D}, \tag{3}$$

where $W_k \in \mathbb{R}^{24\times8}$ is a weight matrix; v_k^{24D} is a 24D VN capsule after mapping. Then an ID capsule (C_n^{24D}) in the C-Caps layer can be calculated by:

$$C_n^{24D} = \sum_{k=1}^{K} u_k^n \cdot v_k^{24D}, \tag{4}$$

where $n \in [1, N]$, u_k^n represents the coupling coefficient which is determined by a routing-by-agreement (R-by-A) process between the P-Caps and the C-Caps layers. Notably, all the C_n^{24D} are normalized by the Squashing Function (refer to Eq. 2).

The R-by-A process is a key technique to build the relationship between the P-Caps and C-Caps layers. The details of R-by-A can refer to [17]. Our R-by-A process is similar to [17]. The only difference is that we set the number of routing iteration to four rather than three in [17]. We found four iterations achieve better re-ID accuracy in experiments.

[1] Other alternatives also can be used.

Given an input image I_n^x, we use the Margin Loss for ID existence:

$$\begin{aligned} L_{CAPS} = \sum_{n=1}^{N} \{ & y_n \cdot max(0, m^+ - \|C_n^{24D}\|)^2 \\ & + \lambda \cdot (1 - y_n) \cdot max(0, \|C_n^{24D}\| - m^-)^2\}, \end{aligned} \tag{5}$$

where y_n represents the existence of ID for the input image I_n^x; $y_n = 1$ if I_n^x belongs to ID n, otherwise $y_n = 0$. $\lambda = 0.5$ is used to balance the weight between the two parts in L_{CAPS}. m^+ and m^- are used to control the length of C_n^{24D}. If the ID is present in I_n^x, we expect the length of C_n^{24D} should be long, otherwise it should be short. We set $m^+ = \frac{N-1}{N}$ and $m^- = \frac{1}{N}$.

C. Auxiliary module

FSR and SEA mechanisms. Before formally introducing the two mechanisms, the Global Average Pooling (GAP) is adopted to the output of the backbone (*i.e.*, $\mathcal{O}(I_n^x)$) to obtain a CNN representation $\mathcal{F}(I_n^x)$ of the input image I_n^x:

$$\mathcal{F}(I_n^x) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{O}(I_n^x)(i, j), \tag{6}$$

where H and W respectively represent the height and width of $\mathcal{O}(I_n^x)$ ($H = 7$ and $W = 7$ in our case). After GAP, we denote the input of FSR and SEA as $\mathcal{F}_{FSR}(I_n^x)$ and $\mathcal{F}_{SEA}(I_n^x)$, respectively.

In the FSR mechanism, we expect to enhance the generalization power of our CNN-based output to affect the learning capability of the whole network. As is well-known, Dropout has demonstrated its effectiveness to prevent the CNN network from over-fitting by randomly dropping out neurons (set the value to zero) in network training [31]. We simply adopt the Dropout to make the $\mathcal{F}_{FSR}(I_n^x)$ to become a sparse representation. The generalization capability of $\mathcal{F}_{FSR}(I_n^x)$ can be transferred to $\mathcal{O}(I_n^x)$ in backward propagation, which potentially improves the robustness of the whole network (particularly for all branches after $\mathcal{O}(I_n^x)$). We set the dropout rate to 0.75 in FSR.

In the SEA mechanism, we aim to boost the discriminative power of the backbone network by applying the Squeeze-and-Excitation (SE) block [32] on $\mathcal{O}(I_n^x)$. By doing so, we can substantially highlight informative features and suppress the useless ones in $\mathcal{O}(I_n^x)$ which is the key link between the CNN-based architecture and the capsule layers. In this way, we can use the prior knowledge learned by Densenet-121 in a more efficient way when the knowledge is fed into the Caps Modules. We first use the GAP after $\mathcal{O}(I_n^x)$ to obtain $\mathcal{F}_{SEA}(I_n^x)$. Then, we squeeze $\mathcal{F}_{SEA}(I_n^x)$ by a fully-connected layer to obtain a low-dimensional ($1024/r$-dim) representation, where r is the reduction rate. After that, another 1024-dim fully-connected layer ($\mathcal{F}'_{SEA}(I_n^x)$) is linked after the $1024/r$-dim layer. The final re-weight block is obtained by rescaling $\mathcal{O}(I_n^x)$ with $\mathcal{F}'_{SEA}(I_n^x)$:

$$\mathcal{O}'(I_n^x) = \mathcal{F}'_{SEA}(I_n^x) \otimes \mathcal{O}(I_n^x), \tag{7}$$

where $\otimes$ represents channel-wise multiplication between $\mathcal{F}'_{SEA}(I_n^x)$ and $\mathcal{O}(I_n^x)$. In SEA, we set the $r = 16$ (the same as [32]). Relu and Sigmoid activation functions are respectively used after the two fully-connected layers followed by $\mathcal{F}_{SEA}(I_n^x)$.

Finally, we use Cross-Entropy loss on both FSR ($\mathcal{L}_{FSR}$ in Eq. 1) and SEA ($\mathcal{L}_{SEA}$ in Eq. 1) branches to classify the ID of persons in the training set.

V. Experiments

In this section, we first introduce four different datasets we used in our experiments. Then, an experimental setup is given. An ablation study of ReIDCaps is provided. Moreover, quantitative and qualitative analyses will be given to verify the effectiveness of VN capsules comparing with the traditional SN. Finally, a comprehensive evaluation is carried out by comparing our methods with other state-of-the-art methods. Our experiments are mainly conducted on Celeb-reID since the proposed method is particularly designed to tackle the cloth change challenge exposed in the long-term person re-ID scenario.

A. Person Re-ID Datasets

In this paper, we use four different datasets to evaluate the proposed method mentioned in Section IV. Two of them are built for the long-term person re-ID study. We also use two traditional short-term person re-ID dataset Market1501 [4] and DukeMTMC-reID [5] to evaluate our ReIDCaps network. Although our method is specifically designed for the long-term re-ID scenario, we also use these two datasets to further verify its robustness in the common re-ID scenario. Notably, amongst all the short-term re-ID datasets, we select Market1501 and DukeMTMC-reID for evaluation because of two reasons. First, both of them are widely used datasets to the community. Second, many recently published state-of-the-art approaches use the two datasets for evaluation.

Celeb-reID is introduced in Section III. We use this dataset for long-term re-ID evaluation. Details of Celeb-reID can refer to Tab. II. Notably, a person may wear the same clothes (the ratio is less than 30% within each ID) in Celeb-reID.

Celeb-reID-light is a light version of Celeb-reID. Unlike Celeb-reID, a person in Celeb-reID-light will not wear the same cloth twice. We collect Celeb-reID-light from the candidate image set after data filtering (see Fig. 2). There are 590 persons. 490 IDs with 9,021 images are used for training, and 100 IDs with 1,821 images are used for testing. In the testing set, 887 images are used as queries, and 934 images are used as galleries. Although the scale of Celeb-reID-light is smaller than Celeb-reID, it can be used to testify the robustness of re-ID methods when a person is entirely in different clothes.

Market1501 is one of the widely used short-term person re-ID dataset. There are 751 (12,936) and 750 (19,732) IDs (images) in the training and testing sets respectively.

DukeMTMC-reID is another widely used short-term person re-ID dataset. There are 702 IDs in the training and testing sets respectively. The number of training images is 16,522 while the number of testing images is 19,889.

B. Experimental Setup

The PyTorch package is used to implement our ReIDCaps network. To be consistent, we do not change the network design, but use two different training strategies for short-term and long-term re-ID scenarios respectively. This is because, for long-term re-ID, the new model needs to produce more learnable parameters to accommodate additional information about the cloth change. Thus, it may cause certain overfitting issue when it is applied to a relatively simple case of short-term re-ID where there is no cloth change. For the long-term re-ID scenario, the initial learning rate is set to 1e-4 in the ImageNet-trained DenseNet-121 and 1e-3 in new layers after the DenseNet output (see Fig. 5). We name this training strategy as TS1. In order to mitigate the possible overfitting issue, for the short-term re-ID scenario, we pre-train the backbone network (DenseNet-121) on common SN (refer to Fig. 8). The SN-trained DenseNet-121 backbone network is adopted in ReIDCaps. The initial learning rate is set to 1e-5 in the SN-trained DenseNet-121 and 1e-4 in new layers. We name this training strategy as TS2.

For both scenarios, all input images are resized to 224×224 and randomly flipped before training. The Adam stochastic optimization [33] is used with parameters $\beta 1 = 0.9$, $\beta 2 = 0.999$. The learning rate is decayed by 10 times after 40 training epochs, and we stop training after 50-th epochs. The number of training IDs N = 632, 490, 751, and 702 for Celeb-reID, Celeb-reID-light, Market1501, and DukeMTMC-reID respectively.

In testing, the same procedure as previous works (*e.g.*, [13], [14], [20], [16]) is adopted. We use Eq. 6 to extract 1,024D feature as the person's description. This feature is used to compare the similarity between any two images in query and gallery sets using the Euclidean distance. The standard rank-N and mAP re-ID evaluation protocols are used in our experiments. The single query setting is adopted over the four datasets in testing.

C. Ablation Study of ReIDCaps

An ablation study is given to verify some important settings of ReIDCaps. The proposed Celeb-reID dataset is used.

1) Hyper-parameter Setting: We evaluate the parameter γ in Eq. 1. The results are shown in Fig. 6 and Fig. 7, respectively. When $\gamma = 0$ (*i.e.*, without the help of auxiliary modules), our ReIDCaps reduces to the baseline. It can be observed that our method achieves the best performance on both mAP and rank-1 accuracy when the $\gamma = 0.5$. Therefore, we set $\gamma = 0.5$ in our experiment setting.

2) Number of Iterations by R-by-A: To get a proper number of iterations between P-Caps and C-Caps by R-by-A, we use the Caps modules and the backbone network of ReIDCaps. In the original capsule network design [17], the R-by-A algorithm uses 3 iterations between the P-Caps and C-Caps. We also follow the same setting in our ReIDCaps network (see Caps$_{iter=3}$ in Tab. III). When we change the length of C-Caps capsules to 16 (Caps$_{iter=3,C_n^{16D}}$, as the same setting in [17]) instead of 24 (our setting, see Fig. 5), we can get a baseline performance (mAP: 7.8%, rank-1: 43.8%). After changing the length of capsules from 16 to 24, the performance can be improved from 43.8% to 46.2% in rank-1 accuracy. Thus, we use the length of C-Caps capsules to 24 in the follow-up experiments. It can be observed that compared with Caps$_{iter=3}$, we can achieve better re-ID accuracy when the number of iteration is set to 4 (Caps$_{iter=4}$, mAP=9.0%, rank-1=48.2%). It can be clear to see that when iter=5 (Caps$_{iter=5}$), the performance is dropped again. Therefore, we choose iter=4 in the R-by-A algorithm.

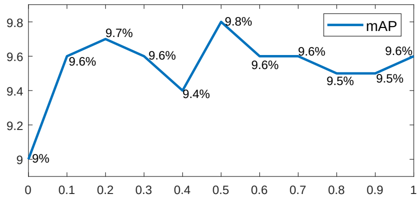


Fig. 6. Sensitivity to parameter γ in Eq. 1. The x-axis and y-axis respectively represents the γ and mAP. Experiment is conducted on Celeb-reID.

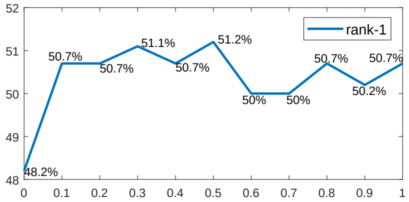


Fig. 7. Sensitivity to parameter γ in Eq. 1. The x-axis and y-axis respectively represents the γ and rank-1 accuracy. Experiment is conducted on Celeb-reID.

TABLE III
ABLATION STUDY OF OUR REIDCAPS NETWORK. MAP AND RANK-N (N=1, 5, AND 10) ARE LISTED. THE BEST PERFORMANCE IS HIGHLIGHTED IN **BLOD**.

Methods	Celeb-reID			
	mAP	rank-1	rank-5	rank-10
Caps$_{iter=3,C_n^{16D}}$	7.8%	43.8%	59.2%	67.5%
Caps$_{iter=3}$	8.7%	46.2%	62.1%	68.9%
Caps$_{iter=4}$	9.0%	48.2%	62.7%	70.0%
Caps$_{iter=5}$	8.7%	47.5%	62.4%	69.6%
Caps$_{iter=4}$+0.5*FSR	9.2%	49.1%	63.1%	70.4%
Caps$_{iter=4}$+0.5*SEA	9.5%	49.8%	63.8%	70.8%
Caps$_{iter=4}$+0.5*(FSR+SEA)	**9.8%**	**51.2%**	**65.4%**	**71.9%**

3) Combination of Different Modules: We try to combine different auxiliary modules (*i.e.*, FSR and SEA) to verify the effectiveness of them in enhancing the overall performance, shown in Tab. III. We first combine FSR (SEA) (with the weight is 0.5, refer to Sec. V-C1) with Caps$_{iter=4}$, the re-ID accuracy is improved from 48.2% to 49.1% (49.8%) in rank-1 accuracy. This combination verifies the effectiveness of different auxiliary modules. Moreover, we combine both FSR and SEA with the weight of 0.5. The rank-1 accuracy gains 3% comparing with single Caps$_{iter=4}$ module. This result verifies the combination of auxiliary modules (*i.e.*, FSR and SEA) can be useful in enhancing the overall performance.

4) Comparison of Different Training Strategies: In our experimental setup, we use two different training strategies in long-term and short-term re-ID scenarios respectively (see Sec. V-B). We try to conduct the two training strategies

TABLE IV
ABLATION STUDY OF DIFFERENT TRAINING STRATEGIES ON OUR REIDCAPS MODEL OVER TWO DIFFERENT PERSON RE-ID SCENARIOS.

Training Strategies	mAP	rank-1	rank-5	rank-10
Celeb-reID (long-term)				
TS1	9.8%	51.2%	65.4%	71.9%
TS2	9.5%	50.3%	64.9%	71.7%
Market1501 (short-term)				
TS1	62.5%	84.8%	94.7%	98.1%
TS2	72.7%	89.0%	95.5%	96.9%

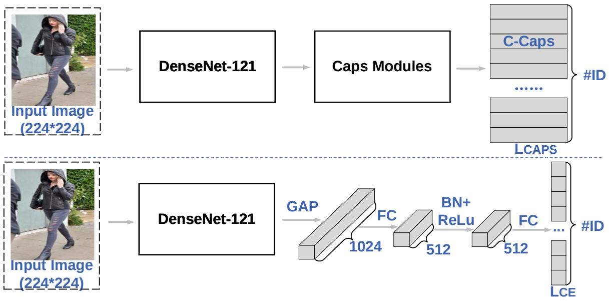


Fig. 8. The SN-based and VN capsule networks. The upper network uses VN capsules. The lower one uses the tradition CNN layers. Both networks use the same input image size and backbone network (*i.e.*, the DenseNet-121). FC, BN, ReLU, and L_{CE} respectively represent the fully-connected layer, batch normalization, ReLu activation function, and the Cross-Entropy loss.

TABLE V
PERFORMANCE COMPARISON BETWEEN THE SN-BASED IDE+ MODEL (THE LOWER NETWORK IN FIG. 8) AND OUR CAPS$_{iter=4}$ MODEL (THE UPPER NETWORK IN FIG. 8).

Methods	Celeb-reID			
	mAP	rank-1	rank-5	rank-10
IDE+	5.9%	42.9%	56.4%	63.4%
Caps$_{iter=4}$	**9.0%**	**48.2%**	**62.7%**	**70.0%**

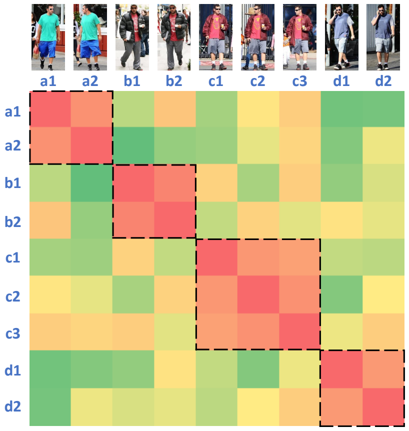


Fig. 9. Intra-class variation visualization using C-Caps. We select four types of clothes (9 images) belonging to the same person (76 images in total) in the training set of Celeb-reID. ‘a’ to ‘d’ represent different clothes and ‘1, 2, ...’ represents the index of sample images. The cosine similarity is used to calculate the similarity between two images using the VN capsules in C-Caps where the ID is presented. We use an activation map to represent the similarity between any two images. The red and green color respectively represent the most and the least similar pairs. Elements in the diagonal are self-similarity.

over different re-ID scenarios to verify the function of them. It can be observed in Tab. IV that, compared with TS1, TS2 can improve the performance on Market1501 noticeably. This result verify that using the SN-trained backbone can effectively mitigate the overfitting problem produced by more learnable parameters from VN. On the contrary, when different training strategies are applied for the case of long-term re-ID, the performance is barely affected. This comparison shows that our method is robust to the long-term re-ID scenario under different training strategies. Although Celeb-reID also achieves comparable performance by TS2 which is more suitable for the short-term re-ID scenario, we still follow our experimental setup since Celeb-reID and Market1501 achieve the best performance under TS1 and TS2 respectively.

D. Scalar Neuron vs. Vector Neuron: Brief Quantitative and Qualitative Analyses

Network Architecture. We compare our VN capsules with a SN solution. Network architectures are given in Fig. 8. Both of them share the same input and backbone network (*i.e.*, ImageNet-trained DenseNet-121). The VN capsules (Caps modules in Fig. 2) and SN-based layers are followed by the output of DenseNet-121 respectively. The SN-based network (also called as IDE+, see Sec. II) is a widely used benchmark for short-term person re-ID scenario. The result of SN-based IDE+ is achieved by the Deep-Person-reID package[2].

Quantitative Evaluation. Tab. V shows the quantitative evaluation between our Caps$_{iter=4}$ and the IDE+ models. By using the VN capsules, our Caps$_{iter=4}$ outperforms the IDE+ by +3.1% (9.0% *vs.* 5.9%) on mAP and +5.3% (48.2% *vs.* 42.9%) in rank-1 accuracy. It is clear to see that the VN capsules show their superiority in dealing with the cloth change challenge for person re-ID. This is because in addition to distinguish the inter-class variation, VN capsule can potentially perceive the intra-class variation when clothes of the same person are changed.

Qualitative Evaluation. We visually illustrate the intra-class variation of the same people ID using our Caps$_{iter=4}$. The cosine similarity is used to calculate the orientation distance between two images. Specifically, given two images (*e.g.*, $I_{n_{id}}^{x_1}$ and $I_{n_{id}}^{x_2}$) belonging to the same ID (e.g., n_{id}), we can extract VN capsules for this ID (*i.e.*, $C_{n_{id}}^{24D}(I_{n_{id}}^{x_1})$ and $C_{n_{id}}^{24D}(I_{n_{id}}^{x_2})$) obtained by Eq. 4 (after training is completed) to calculate the cosine distance between the two 24D VNs. An activation map of similarity is used to illustrate the effectiveness of our Caps Modules in perceiving the knowledge of intra-class clothes changes. It can be observed in Fig. 9 that our Caps$_{iter=4}$ can potentially learn the cloth changes of the same person. Although we do not use any clothing type labels in training and add explicit constraints in our network, the same types of clothes show higher similarity than different types to some extent. This result demonstrates that the orientation of VN can carry the information of clothes change of the same person for the long-term person re-ID.

[2] https://github.com/KaiyangZhou/deep-person-reid

TABLE VI
Performance by using different weights on different body parts. The part partition can refer to Fig. 10. We mainly evaluate the result of Celeb-reID. The weight assigned to Celeb-reID-light is similar to Celeb-reID since both belong to the long-term re-ID scenario. We simply use another group of weights on Market1501 by considering the contribution of different body parts.

Methods	Celeb-reID		
	mAP	rank-1	rank-5
$P_{11}+P_{12}+P_{13}+P_{21}+P_{22}$	13.6%	60.7%	74.2%
$P_{11}+P_{12}+P_{13}+P_{21}+P_{22}+G$	14.2%	62.2%	75.2%
$P_{11}+0.9*(P_{12}+P_{13})+P_{21}+0.9(P_{22})+G$	14.9%	62.6%	75.4%
$P_{11}+0.7*(P_{12}+P_{13})+P_{21}+0.7(P_{22})+G$	15.4%	62.9%	76.3%
$P_{11}+0.5*(P_{12}+P_{13})+P_{21}+0.5(P_{22})+G$	**15.8%**	**63.0%**	**76.3%**
$P_{11}+0.3*(P_{12}+P_{13})+P_{21}+0.3(P_{22})+G$	15.8%	62.3%	75.7%
$P_{11}+0.1*(P_{12}+P_{13})+P_{21}+0.1(P_{22})+G$	15.4%	61.4%	75.1%
	Celeb-reID-light		
$P_{11}+0.5*(P_{12}+P_{13})+P_{21}+0.5(P_{22})+G$	**19.0%**	**33.5%**	**63.3%**
	Market1501		
$0.25*(P_{11}+P_{12}+P_{13})+0.5(P_{21}+P_{22})+G$	**78.0%**	**92.8%**	**97.6%**

E. *Comparison with State-of-The-Art Methods*

We compare our proposed ReIDCaps method with several state-of-the-art re-ID approaches. Our method is mainly designed for the long-term re-ID scenario. It may not be a universal solution to both long-term and short-term re-ID scenarios. But we also evaluate the proposed method on the case of short-term re-ID to verify its robustness under different re-ID scenarios. Tab. VII lists the results. Four different coarse-grained re-ID approaches (*i.e.*, IDE+ [11], [12], ResNet-Mid [13], Two-Stream [14], MLFN [15]), and four different fine-grained re-ID approaches (*i.e.*, HACNN [23], Part-Bilinear [22], PCB [20], and MGN [16]) are picked up to evaluate the challenge exposed in the long-term re-ID scenario. Coarse-grained methods leverage the whole image as inputs while fine-grained methods take both global and body parts information into consideration. All the methods we used are recently published state-of-the-art methods which perform promisingly in the traditional short-term re-ID scenario. In these methods, the IDE+, ResNet-Mid, MLFN, and HACNN are implemented by Deep-Person-Reid package with solid performance. The rest methods are released by the authors of original papers (*i.e.*, Two-Stream, Part-Bilinear, PCB, and MGN).

Comparison with coarse-grained methods. It can be observed in Tab. VII, by only using the global body cue, our method outperforms other coarse-grained methods on both Celeb-reID and Celeb-reID-light datasets. Compared with the second best method ResNet-Mid, our ReIDCaps outperforms it on Celeb-reID by a large margin in rank-1 accuracy (51.2% *vs.* 43.3%). We also achieve the best performance on Celeb-reID-light (mAP: 11.2%, rank-1: 20.3%). This result verifies our ReIDCaps can best tackle the extreme case where a person does not wear the same clothes twice. MLFN achieves the best performance on Market1501 (mAP: 74.3%, rank-1: 90.0%). However, it only obtains 6.0% mAP and 41.4% rank-1 accuracy on Celeb-reID. Although our method is not designed for the short-term re-ID scenario, we also try to experiment with Market1501 (DukeMTMC-reID) and obtain 89.0% (81.2%) in rank-1 accuracy, which is a comparable result comparing with other coarse-grained approaches.

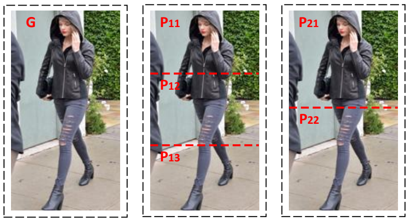


Fig. 10. Body parts partitions. The whole image is denoted as G. P_{11}, P_{12}, and P_{13} (also P_{21} and P_{22}) are parts equally divided from G.

Comparison with fine-grained methods. To compare with fine-grained learning approaches, the body parts are utilized in our experiments. We follow the same setting in [16] which equally divides a person image into three and two parts respectively. Fig. 10 shows the partition result. We simply train and test our ReIDCaps on the five parts respectively. Tab. VII shows the re-ID accuracy using different parts on four re-ID datasets. The long-term re-ID datasets such as Celeb-Reid and its light version involve appearance changes dramatically. Parts such as P_{12}, P_{13}, and P_{22} show much lower re-ID accuracy than P_{11} and P_{21} where present more robust appearance cues, *i.e.*, the upper body. However, short-term re-ID dataset, *i.e.*, Market1501 and DukeMTMC-reID illustrate opposite results. The P_{11} and P_{21} on Market1501 even show lower re-ID accuracy comparing with other body parts. This is because, without substantial appearance changes, cues such as color and texture are more likely to be re-identified as significant factors. When clothes are completely changed (Celeb-reID-light), the appearance cues become even less reliable. However, with the benefits of our ReIDCaps network, even the most difficult part (P_{12}) can play a certain role (mAP: 3.5%, rank-1: 5.3%) in Celeb-reID-light.

As in [16], we integrate the five body parts and the global body cue to get the final result (denoted as ReIDCaps*). We try to assign different weights to each part according to their contributions. Tab. VI shows the performance. Considering the discriminability of different body parts, we simply fix the weight of P_{11}, P_{21}, and G to 1 because the three parts are visually more recognizable than other parts (P_{12}, P_{13}, and P_{22}, refer to Fig. 10). Under this setting, the search space (*i.e.*, different combinations) can be greatly reduced. We use the same weight on parts P_{12}, P_{13}, and P_{22}, the weight varies from 0.1 to 1 (gap 0.2). We find when the weight equals to 0.5, our method achieves the best performance. Since the Celeb-reID and Celeb-reID-light are all long-term re-ID datasets, we use the same setting. We change the weight on Market1501 and DukeMTMC-reID because each person in both datasets does not change clothes. With the help of color information on clothes, we can easily recognize each person on Market1501 and DukeMTMC-reID with more body part information. Therefore, we simply set the weight of the whole body to 1, 0.5 for large body parts (P_{21} and P_{22}), and 0.25 for small body parts (P_{11}, P_{12}, and P_{13}).

It is clear to see in Tab. VII that our ReIDCaps* outperforms other methods by a large margin on Celeb-reID and Celeb-

TABLE VII
COMPARISON OF OUR RESULTS WITH THE PUBLISHED STATE-OF-THE-ART METHODS. THE BEST RESULT IS SHOWN IN **BOLD**. RANK-N ACCURACY AND mAP ARE LISTED.

Methods	Long-Term Person Re-ID						Short-Term Person Re-ID			
	Celeb-reID (New)			Celeb-reID-light (New)			Market1501		DukeMTMC-reID	
	mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	mAP	rank-1
Coarse-Grained Learning Approaches (without human body parts partition)										
IDE+ (DenseNet-121)	5.9%	42.9%	56.4%	5.3%	10.5%	24.8%	68.0%	87.8%	57.5%	77.9%
ResNet-Mid [13] ArXiv17	5.8%	43.3%	54.6%	6.0%	10.3%	28.0%	75.6%	89.9%	**64.0%**	**81.6%**
Two-Stream [14] TOMM18	7.8%	36.3%	54.5%	-	-	-	60.9%	80.3%	51.4%	72.6%
MLFN [15] CVPR18	6.0%	41.4%	54.7%	6.3%	10.6%	31.0%	**74.3%**	**90.0%**	63.2%	81.1%
ReIDCaps (Ours)	**9.8%**	**51.2%**	**65.4%**	**11.2%**	**20.3%**	**48.2%**	72.7%	89.0%	62.6%	81.2%
Fine-Grained Learning Approaches (with human body parts partition)										
HACNN [23] CVPR18	9.5%	47.6%	63.3%	11.5%	16.2%	42.8%	75.7%	91.2%	63.2%	80.1%
Part-Bilinear [22] ECCV18	6.4%	19.4%	40.6%	-	-	-	74.5%	88.8%	64.2%	82.1%
PCB [20] ECCV18	8.2%	37.1%	57.0%	-	-	-	77.4%	92.3%	66.1%	81.8%
MGN [16] ACMMM18	10.8%	49.0%	64.9%	13.9%	21.5%	47.4%	**86.9%**	**95.7%**	**78.4%**	**88.7%**
ReIDCaps* (Ours)	**15.8%**	**63.0%**	**76.3%**	**19.0%**	**33.5%**	**63.3%**	78.0%	92.8%	67.8%	83.8%
Performance on Different Body Part using ReIDCaps										
ReIDCaps: P_{11}	11.5%	53.6%	69.2%	14.4%	24.5%	54.9%	29.2%	56.2%	38.7%	59.4%
ReIDCaps: P_{12}	3.6%	33.7%	43.2%	3.5%	5.3%	19.2%	40.1%	65.3%	34.5%	56.6%
ReIDCaps: P_{13}	3.7%	32.3%	43.2%	4.0%	7.5%	22.4%	22.9%	41.6%	20.8%	35.0%
ReIDCaps: P_{21}	10.2%	50.9%	66.5%	13.1%	25.3%	51.8%	37.7%	65.4%	45.0%	66.6%
ReIDCaps: P_{22}	4.3%	36.0%	48.0%	4.4%	8.5%	23.8%	45.5%	67.3%	32.9%	52.2%

reID-light. We obtain 15.8% mAP and 63.0% rank-1 accuracy on Celeb-reID. The second best method MGN only achieves 10.8% mAP and 49.0% rank-1 accuracy. Even without using any body part cue, our method (ReIDCaps) can outperform MGN in rank-1 accuracy (51.2% *vs.* 49.0%). In the meantime, we outperform MGN on Celeb-reID-light by 12.0% in rank-1 accuracy (33.5% *vs.* 21.5%), which further verifies the effectiveness of our method in dealing with the long-term re-ID scenario.

Robustness Evaluation. To further verify the robustness of re-ID approaches between the long-term and short-term re-ID scenarios. In this paper, we define a Robustness Score (RS$\in [0,1]$). It can be observed that MGN achieves the state-of-the-art re-ID performance on both short-term re-ID datasets (*i.e.*, Market1501 and DukeMTMC-reID). This is because MGN is elaborately designed for short-term re-ID based on the appearance information of clothes. However, MGN is still hard to tackle long-term re-ID challenge when people can change their clothes. To get a quantitative comparison of the robustness between different re-ID approaches, the RS is defined as follows:

$$RS(\mathbb{P}) = \frac{\sqrt{score_L(\mathbb{P}) \times score_S(\mathbb{P})}}{|score_L(\mathbb{P}) - score_S(\mathbb{P})| + 1}, \tag{8}$$

where $score_L$ and $score_S$ represent the long-term and short-term re-ID accuracy respectively; $\mathbb{P} \in \{mAP, rank-N\}$ represents different evaluation indexes. Given a re-ID model, RS returns a large value if the score of both $score_L$ and $score_S$ are high and close to each other. Tab. VIII shows the robustness comparison results by using Eq. 8. We can observe that our method has better robustness than MGN when the evaluation is conducted on different re-ID scenarios.

Reliability of Head Information. Normally, person re-ID (in existing researches) would like to use overall body information rather than head/face information. In this experiment, we want to clarify that the proposed dataset does not present high-quality head/face information. Otherwise, any method of person re-ID developed on such dataset may have bias caused by strong head/face information. In order to verify this point, the 2D human pose estimation approach [34] is adopted to localize anatomical keypoints on the body of persons. According to the keypoint of the neck, we can extract the head part from the body. As shown in Fig. 11, we obtain the head part (including the face) from our Celeb-reID dataset and Market1501. It can be clear to see that, in terms of head information, the proposed dataset shares the same characteristics and presumption (*i.e.*, head information not useful) as other existing datasets for person re-ID research such as Market1501. The Two-Stream [14] model, which contains both identification and verification signals, is used to train and test the re-ID performance when only the head images are adopted. Tab. IX shows the result. It shows that the quality of head information in the proposed dataset is even worse than Market1501. This is because celebrities in our dataset have a great chance to wear sunglasses or hats, which makes them hard to be recognized.

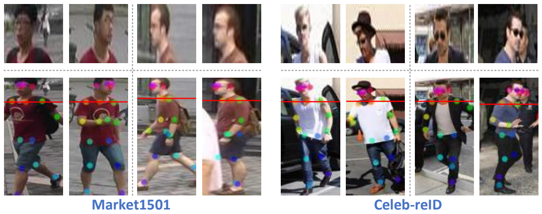


Fig. 11. Pose estimation results (the bottom row). The head images (the top row) are extracted according to the location of keypoint of the neck.

VI. CONCLUSION

This paper introduces a new long-term person re-ID dataset called “Celeb-reID” to the community. This dataset uses the street snap-shots of celebrities as the resource. Compared with

TABLE VIII
ROBUSTNESS SCORE EVALUATION BETWEEN LONG-TERM (CELEB-REID (C) OR CELEB-REID-LIGHT (C-L)) AND SHORT-TERM (MARKET1501 (M) OR DUKEMTMT-REID (D)) RE-ID SCENARIOS USING THE RS (SEE EQ. 8). THE MAP AND RANK-1 (R1) ACCURACY (FROM TAB. VII) ARE USED AS EVALUATION INDEXES TO THE RS.

Methods	RS(mAP)				RS(rank-1)			
	C & M	C & D	C-l & M	C-l & D	C & M	C & D	C-l & M	C-l & D
MGN	0.17	0.17	0.20	0.20	0.47	0.47	0.26	0.26
ReIDCaps* (Ours)	**0.22**	**0.22**	**0.24**	**0.24**	**0.59**	**0.60**	**0.35**	**0.35**

TABLE IX
COMPARISON BETWEEN RE-ID PERFORMANCE WHEN ONLY HEAD IMAGES ARE USED.

Methods	rank-1	mAP
Celeb-reID (head)		
Two-Stream [14]	27.9%	11.6%
Market1501 (head)		
Two-Stream [14]	29.0%	13.2%

previous datasets, our dataset is the largest re-ID dataset with the cloth change of the same people ID. A ReIDCaps model is designed to tackle the cloth change challenge exposed in this paper. Compared with the common SN-based CNN, we use VN capsules to perceive the cloth change of the same person. We integrate the capsule layers with an ImageNet-trained CNN on complex person re-ID data. A comprehensive experiment is given to demonstrate the superiority of our method in the long-term re-ID scenario.

ACKNOWLEDGMENT

We thank Fucheng Guo and Fanfan Liu for working on the new dataset. This research is supported by an Australian Government Research Training Program Scholarship and Beijing Institute of Technology Research Fund Program for Young Scholars.

REFERENCES

[1] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, 2014.

[2] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 262–275, 2008.

[3] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 152–159, 2014.

[4] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1116–1124, 2015.

[5] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3754–3762, 2017.

[6] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 433–442, 2012.

[7] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *Proc. Eur. Conf. Rob. Autom. (ICRA)*, pages 4512–4519, 2014.

[8] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1229–1238, 2016.

[9] Peng Zhang, Qiang Wu, Jingsong Xu, and Jian Zhang. Long-term person re-identification using true motion from videos. In *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 494–502, 2018.

[10] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1367–1376, 2017.

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017.

[12] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3800–3808, 2017.

[13] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv:1711.08106*, 2017.

[14] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)*, 14(1):13, 2018.

[15] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, volume 1, page 2, 2018.

[16] Wang Guanshuo, Yuan Yufeng, Chen Xiong, Li Jiwei, and Zhou Xi. Learning discriminative features with multiple granularities for person re-identification. In *Proc. ACM Int. Conf. Multimedia. (ACMMM)*, pages 274–282, 2018.

[17] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pages 3856–3866, 2017.

[18] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 79–88, 2018.

[19] Yuan Yuan, Jian'an Zhang, and Qi Wang. Bike-person re-identification: A benchmark and a comprehensive evaluation. *IEEE Access*, 6:56059–56068, 2018.

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 501–518, 2018.

[21] Yan Huang, Hao Sheng, Yanwei Zheng, and Zhang Xiong. Deepdiff: learning deep difference features on human body parts for person re-identification. *Neurocomputing*, 241:191–203, 2017.

[22] Suh Yumin, Wang Jingdong, Tang Siyu, Mei Tao, and Mu Lee Kyoung. Part-aligned bilinear representations for person re-identification. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 418–437, 2018.

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, volume 1, page 2, 2018.

[24] Jian'an Zhang, Qi Wang, and Yuan Yuan. Metric learning by simultaneously learning linear transformation matrix and weight matrix for person re-identification. *IET Comput. Vis.*, 13(4):428–434, 2019.

[25] Lingchuan Sun, Zhuqing Jiang, Hongchao Song, Qishuo Lu, and Aidong Men. Semi-coupled dictionary learning with relaxation label space transformation for video-based person re-identification. *IEEE Access*, 6:12587–12597, 2018.

[26] Rodney LaLonde and Ulas Bagci. Capsules for object segmentation. *arXiv:1804.04241*, 2018.

[27] Ayush Jaiswal, Wael AbdAlmageed, Yue Wu, and Premkumar Natarajan. Capsulegan: Generative adversarial capsule network. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 526–535, 2018.

[28] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 2980–2988, 2017.

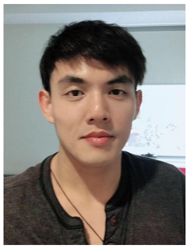
[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4700–4708, 2017.

[31] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res. (JMLR)*, 15(1):1929–1958, 2014.

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7132–7141, 2018.

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980.*, 2014.

[34] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7291–7299, 2017.



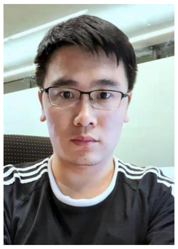
Yan Huang received B.S. and M.S. degrees from the School of Computer Science and Engineering in Sichuan University (SCU), China, and Beihang University (BUAA), China, respectively. He is currently a Ph.D. student with the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), NSW, Australia. His research interests include deep learning and person re-identification.

Jingsong Xu received the B.S. degree in computer science and the Ph.D. degree in pattern recognition from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2014, respectively. He is currently a Research Fellow with the Global Big Data Technologies Center, University of Technology Sydney (UTS), Ultimo, NSW, Australia. His research interests include computer vision, pattern recognition, and machine learning.

Qiang Wu received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Australia, in 2004. He is currently an Associate Professor and a Core Member of the Global Big Data Technologies Centre, University of Technology Sydney.

His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. The application fields where the research outcomes are applied span over video security surveillance, biometrics, video data analysis, and human–computer interaction. His research outcomes have been published in many premier international conferences, including ECCV, CVPR, ICCV, ICIP, and ICPR and the major international journals, such as the IEEE TIP, IEEE TSMC-B, IEEE TCSVT, IEEE TIFS, PR, PRL, and Signal Processing.

Yi Zhong received the Ph.D. degree in Information and Communication Engineering from the Beijing University of Posts and Telecommunications, Beijing, China, and in Computer Science from the University of Technology Sydney, Sydney, NSW, Australia, in 2017 and 2019, respectively. She is currently an Assistant Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. Her research interests include pattern recognition, machine learning, deep learning, and signal processing.



Peng Zhang received B.S. and M.S. degrees from the School of Information Science and Engineering in Shandong University (SDU), China. He is currently a Ph.D. student with the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), NSW, Australia. His research interests include gait recognition, person reidentification and deep learning.

Zhaoxiang Zhang received the B.S. degree in electronic science and technology from the University of Science and Technology of China (HUST), Hefei, China, in 2004, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009.

In 2009, he joined the School of Computer Science and Engineering, Beihang University, Beijing, China, as an Assistant Professor from 2009 to 2011, an Associate Professor from 2012 to 2015, and as the Vice-Director of the Department of Computer Application Technology from 2014 to 2015. In 2015, he returned to the Institute of Automation, Chinese Academy of Sciences, as a Full Professor. His current research interests include computer vision, pattern recognition, machine learning, and brain-inspired neural network and brain-inspired learning.

Prof. Zhang is the Associate Editor or Guest Editor of some internal journals, like Neurocomputing, Pattern Recognition Letters, and IEEE ACCESS.