3D Shape Temporal Aggregation for Video-Based Clothing-Change Person Re-Identification

Ke Han[1,2], Yan Huang[1,2]*, Shaogang Gong[3], Yan Huang[1,2], Liang Wang[1,2], and Tieniu Tan[1,2]

[1] Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] University of Chinese Academy of Sciences (UCAS)
[3] Queen Mary University of London (QMUL)
{ke.han, yan.huang}@cripac.ia.ac.cn,
{yhuang, wangliang, tnt}@nlpr.ia.ac.cn, s.gong@qmul.ac.uk

Abstract. 3D shape of human body can be both discriminative and clothing-independent information in video-based clothing-change person re-identification (Re-ID). However, existing Re-ID methods usually generate 3D body shapes without considering identity modelling, which severely weakens the discriminability of 3D human shapes. In addition, different video frames provide highly similar 3D shapes, but existing methods cannot capture the differences among 3D shapes over time. They are thus insensitive to the unique and discriminative 3D shape information of each frame and ineffectively aggregate many redundant framewise shapes in a videowise representation for Re-ID. To address these problems, we propose a 3D Shape Temporal Aggregation (3STA) model for video-based clothing-change Re-ID. To generate the discriminative 3D shape for each frame, we first introduce an identity-aware 3D shape generation module. It embeds the identity information into the generation of 3D shapes by the joint learning of shape estimation and identity recognition. Second, a difference-aware shape aggregation module is designed to measure inter-frame 3D human shape differences and automatically select the unique 3D shape information of each frame. This helps minimise redundancy and maximise complementarity in temporal shape aggregation. We further construct a Video-based Clothing-Change Re-ID (VCCR) dataset to address the lack of publicly available datasets for video-based clothing-change Re-ID. Extensive experiments on the VCCR dataset demonstrate the effectiveness of the proposed 3STA model. The dataset is available at https://vhank.github.io/vccr.github.io.

Keywords: Clothing-change person re-identification · 3D body shape · temporal aggregation.

1 Introduction

Person re-identification (Re-ID) aims to match the same person across non-overlapping cameras. Short-term Re-ID methods [30, 16, 14, 8] consider the prob-

* Corresponding author.

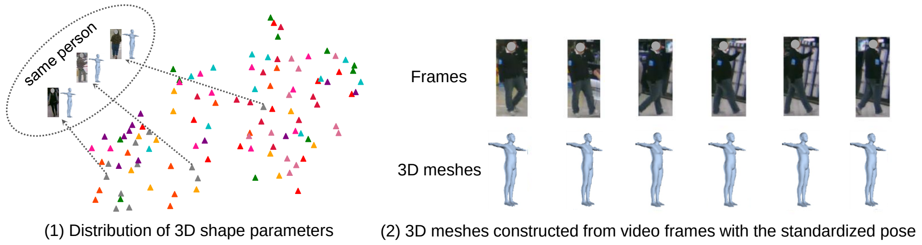


Fig. 1. The motivation of this paper. (1) 3D shape parameters are not discriminative, as shown by randomly sampled 100 images of 10 different persons (color indicated) on the VCCR dataset, with their corresponding 3D shape parameters estimated by a 3D human estimation model [21], visualized by t-SNE. (2) The 3D shapes captured in successive video frames are usually very similar and redundant, as shown in the 3D mesh examples by SMPL [29] constructed with standardized pose parameters.

lem within a short time period assuming no changes of clothing, and therefore are mostly clothing-dependent. In practice, Re-ID over a longer time period in a more general setting, *e.g.*, over several days, probably includes clothing changes. In certain situations, a suspect may even deliberately change clothes to avoid being found. To this end, a number of methods have been proposed to address the challenge of clothing-change person Re-ID (CC Re-ID) [38, 44, 34, 26, 15].

Since clothing is less reliable in CC Re-ID, it is necessary to explore other clothing-independent characteristics, *e.g.*, body shape. Some methods consider 2D body shape features by human contours/silhouettes extraction [44, 13], key-points detection [34] or body shape disentanglement [26]. However, human body actually exists in a 3D space. 2D body shape information is not only view-dependent but also lacking of 3D depth information, which has been shown to be discriminative in Re-ID [42, 37]. Some works [49, 1] therefore have explored 3D shape information for Re-ID, but they suffer from two major limitations.

First, these methods [49, 1] generate 3D human shapes without considering discriminative identity modelling. They usually directly employ a 3D human estimation model to generate 3D shape parameters, which are then used to construct 3D meshes by SMPL [29] for discriminating different persons. As illustrated in Fig.1 (1), due to lack of identity modelling, the 3D shape parameters of the same person can be dispersive (especially when the same person wears different styles of clothes), while those of different persons can be close, making such 3D shape information not discriminative enough for Re-ID. **Second**, as shown in Fig.1 (2), although a video provides richer information a single frame, the 3D shapes captured by successive frames are mostly highly similar. Existing temporal aggregation models [5, 41, 47] in Re-ID are usually designed for appearance instead of 3D shape information. They are insensitive to the differences among 3D shapes over time, and cannot select the unique shape information of each frame. Consequently, many redundant shapes from different frames are aggregated in a videowise shape representation, while some unique and discriminative shape information of each frame is suppressed.

To solve these problems, in this work we propose a 3D Shape Temporal Aggregation (3STA) Re-ID model for video-based CC Re-ID. Our 3STA model consists of three main modules: Identity-aware 3D Shape Generation (ISG), Difference-aware Shape Aggregation (DSA) and Appearance and Shape Fusion (ASF). 1) In order to generate the discriminative 3D shape for each video frame, the ISG embeds identity information into the 3D shape generation. This is realized by combing shape supervision of an auxiliary 3D human dataset with identity supervision of a Re-ID dataset in a joint learning framework. 2) The DSA is formulated to adaptively aggregate videowise shape representations from frames by referring to the intra-frame and inter-frame shape differences. The intra-frame shape difference enables our model to compare the changes of all the shape parameters in each frame for framewise spatial attention learning. The inter-frame shape difference is used to capture the change of each shape parameter over time to learn temporal attention. By considering both differences, DSA is sensitive to the unique and discriminative shape information in each frame and selectively aggregates into videowise shape representations with suppressed redundancy and enhanced complementarity. 3) We also exploit appearance information to model visual similarities unaffected by clothing changes, which can complement shape information especially when the target person only partially changes clothes. The ASF module is presented to fuse appearance and shape information adaptively into the final identity representation.

Another significant challenge to video-based CC Re-ID is that there is no **publicly available** dataset. For this purpose, we introduce a Video-based Clothing-Change Re-ID (VCCR) dataset in this work. Built on the attribute recognition dataset RAP [24] collected in a large indoor shopping mall, VCCR covers rich variations in clothing, cameras and viewpoints. To our best knowledge, it is currently the largest video-based CC Re-ID dataset with 4,384 tracklets of 232 clothing-change persons and 160 distractors, compared to the other dataset [3].

Our contributions are summarized as follows. 1) To our best knowledge, our 3D Shape Temporal Aggregation (3STA) model is the first attempt to explore temporal 3D shape information for video-based CC Re-ID. 2) To generate discriminative 3D shapes for Re-ID, we introduce Identity-aware 3D Shape Generation (ISG) that enforces 3D shape parameters to be person-specific. 3) The proposed Difference-aware Shape Aggregation (DSA) can be sensitive to temporal shape differences, and help minimise the redundancy of shape aggregation. 4) We construct a VCCR dataset for video-based CC Re-ID research. Extensive comparative evaluations show the effectiveness of our method against the state-of-the-art methods.

2 Related Work

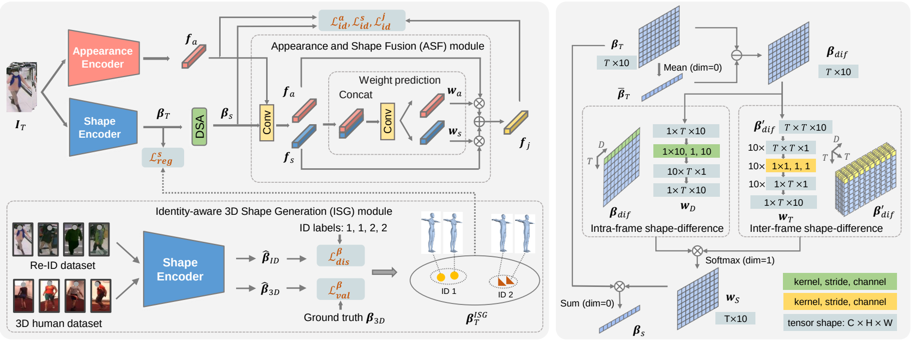
Short-Term Re-ID. Short-term person Re-ID includes image-based [9, 32, 31, 10, 17] and video-based [20, 33, 25] Re-ID. This research primarily relies on clothing information for discriminative person representation learning. Compared with image-based Re-ID methods, video-based Re-ID methods can leverage temporal information in video sequences to explore richer identity features. However, the performance of these short-term Re-ID methods suffers a sharp drop when a person changes clothes.

Imaged-Based Clothing-Change Re-Id. To handle clothing variations for Re-ID, many methods have been proposed to learn clothing-independent shape information, which can be categorized into 2D and 3D shape based methods. For 2D shape learning, Qian *et al.* [34] present a clothing-elimination and shape-distillation model for structural representation learning from human keypoints. Yang *et al.* [44] directly learn feature transformation from human contour sketches. Hong *et al.* [13] transfer shape knowledge learned from human silhouettes to appearance features by interactive mutual learning. Li *et al.* [26] remove clothing colors and patterns from identity features by adversarial learning and shape disentanglement. Shu *et al.* [35] force the model to learn clothing-irrelevant features automatically by randomly exchanging clothes among persons.

In contrast to 2D shape that is confined to a plane, 3D shape can introduce human depth information to facilitate Re-ID. A few works have attempted to employ 3D human estimation models to construct 3D meshes for Re-ID. Zheng *et al.* [49] learn shape features directly with these 3D meshes instead of RGB images as inputs. However, due to the lack of consideration of identity information, the discriminability of constructed 3D meshes is not guaranteed. Chen *et al.* [1] propose an end-to-end framework to recover 3D meshes from original images. This method is supervised in a 2D manner by reprojecting the recovered 3D meshes back into a 2D plane again, which is lack of supervision of 3D shapes. Unlike them, our method unifies ground-truth 3D shape signals from a 3D human dataset with identity signals from a Re-ID dataset in a joint learning framework. In this way, our method can generate reliable and discriminative 3D shapes to boost shape learning for Re-ID.

Video-Based Clothing-Change Re-Id. Compared with image-based CC Re-ID, video-based CC Re-ID is rarely studied and still in the initial stage. Zhang *et al.* [46] make the first attempt on video-based CC Re-ID based on hand-crafted motion features from optical flow, assuming that people have constant walking patterns. Fan *et al.* [3] study video-based CC Re-ID with radio frequency signals reflected from human body instead of RGB color signals, thus completely removing clothing information. Different from them, we take advantage of temporal 3D shape information as a discriminative cue, which is more stable than walking patterns and easier to obtain than radio frequency signals.

Single-View 3D Human Estimation. Single-view 3D human estimation aims to construct human 3D meshes, including 3D shape and pose, from a single image. Current methods [21, 23] typically predict shape and pose parameters with the supervision of 3D ground truths, and then construct 3D meshes by the SMPL model [29]. However, the 3D shape parameters estimated by these models are usually not discriminative enough and cannot fully reflect the differences of body shapes among persons, making these methods not well applied to Re-ID. To this end, our proposed ISG module combines 3D shape estimation and Re-ID in a joint learning framework to generate more discriminative shape parameters.



(1) 3D Shape Temporal Aggregation (3STA) model

(2) Difference-aware Shape Aggregation (DSA) module

Fig. 2. An overview of the proposed 3D Shape Temporal Aggregation (3STA) model.

3 Method

In this paper, we propose a 3D Shape Temporal Aggregation (3STA) Re-ID model to learn temporal 3D shape representations for video-based CC Re-ID. As shown in Fig. 2, our model includes three main modules. The Identity-aware 3D Shape Generation (ISG) is first performed to generate discriminative 3D shape parameters for each video frame. The Difference-aware Shape Aggregation (DSA) exploits the differences across intra-frame and inter-frame 3D shape parameters to aggregate videowise shape parameters. The Appearance and Shape Fusion (ASF) further exploits appearance information to complement shape and adaptively fuses them into final representations for CC Re-ID. Let us start with an introduction to parametric 3D human estimation, which is the basis for our discriminative 3D shape learning.

3.1 Parametric 3D Human Estimation

3D human estimation models [21, 23, 29] usually parameterize 3D human body by shape parameters and pose parameters that are irrelevant to each other. Typically, SMPL [29] is modeled as a function of the pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{24\times3}$ representing the rotation vectors of 24 human joints and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$. Given the two parameters, SMPL can construct the 3D mesh with the corresponding pose and shape. Since pose is not person-specific, we focus only on shape parameters $\boldsymbol{\beta}$ in our modelling.

The SMPL model predefines a human shape template, and uses shape parameters to formulate the shape offset to the template by principal component analysis (PCA). Then the body shape is represented by a linear function B_S

$$B_S = \sum_{n=1}^{|\boldsymbol{\beta}|} \beta^n \boldsymbol{S}^n, \tag{1}$$

where the shape parameters $\boldsymbol{\beta} = [\beta^1, \ldots, \beta^{|\boldsymbol{\beta}|}]^{\mathrm{T}}$, and $|\boldsymbol{\beta}|$ is the number of parameters ($|\boldsymbol{\beta}| = 10$). $\boldsymbol{S}^n \in \mathbb{R}^{3N}$ represents orthonormal principal components of shape offsets, where N is the number of vertices on the predefined human mesh, and $3N$ is the number of 3D coordinates of N vertices. The function B_S thus formulates all the shape offsets to the shape template. When $\boldsymbol{S}^n$ is shared by all the people's shapes, the shape parameters $\boldsymbol{\beta}$ reflect the difference among these shapes. Each parameter β^n (n=1,$\cdots$,10) usually controls some specific aspects of body shape, *e.g.*, the body size, waistline or leg length. We thus can formulate the change of 3D shapes over time in a video by the change of shape parameters.

3.2 Identity-Aware 3D Shape Generation

One challenge of learning 3D shape information for Re-ID is that Re-ID datasets do not contain annotations of 3D shape parameters $\boldsymbol{\beta}$. In fact, it is very difficult and has to rely on special equipments to collect individual 3D shape parameters in real-world scenarios. Existing CC Re-ID methods [49, 1] directly utilize 3D human estimation models to estimate shape parameters which are nevertheless not discriminative. To overcome this problem, we introduce an Identity-aware 3D Shape Generation (ISG) module that embeds identity information into the generation of shape parameters.

In the ISG module (Fig. 2 (1)), a shape encoder is modeled as a function: $\mathbb{R}^{C\times H\times W}\rightarrow\mathbb{R}^{10}$, to predict 10D shape parameters for a given image, where C, H and W are the number of channels, height and width of the image, respectively. The generated 3D shape parameters need to satisfy two requirements. (1) Validity: shape parameters are valid and close to the ground truths so that they can formulate true 3D body shape. (2) Discriminability: shape parameters of the same person are close while those of different persons are away from each other in the parameter space.

To meet the requirement (1), we introduce a 3D human dataset [19] as an auxiliary dataset with ground truths of shape parameters. Images from both the Re-ID and 3D datasets are input into the shape encoder to estimate shape parameters $\hat{\boldsymbol{\beta}}_{ID}$ and $\hat{\boldsymbol{\beta}}_{3D}$, respectively. We introduce a shape validity loss $\mathcal{L}^{\beta}_{val}$:

$$\mathcal{L}^{\beta}_{val}=\|\hat{\boldsymbol{\beta}}_{3D}-\boldsymbol{\beta}_{3D}\|^{2}, \tag{2}$$

where $\boldsymbol{\beta}_{3D}$ is ground-truth 3D shape parameters from the 3D human dataset. To meet the requirement (2), we further introduce a shape discrimination loss $\mathcal{L}^{\beta}_{dis}$ on $\hat{\boldsymbol{\beta}}_{ID}$.

$$\mathcal{L}^{\beta}_{dis}=\mathcal{L}^{\beta}_{ce}+\mathcal{L}^{\beta}_{tri}, \tag{3}$$

where $\mathcal{L}^{\beta}_{ce}$ and $\mathcal{L}^{\beta}_{tri}$ are the cross-entropy and triplet losses, respectively, which are enforced by pairwise positive and negative identity labels from the Re-ID dataset. The total loss for ISG is

$$\mathcal{L}_{ISG}=\mathcal{L}^{\beta}_{val}+\alpha\mathcal{L}^{\beta}_{dis}, \tag{4}$$

where α is a weight factor. After performing ISG, the generated shape parameters for the Re-ID dataset are kept as pseudo labels ($\boldsymbol{\beta}^{ISG}_{T}$) to train the 3STA model.

When we train the 3STA model, the shape encoder is retrained from scratch on the Re-ID dataset and does not share the weights with the shape encoder in ISG. This can decrease the interference of the 3D human data distribution bias in training the shape encoder. We denote a random tracklet composed of T frames in the Re-ID dataset as $\boldsymbol{I}_T$. The shape encoder generates shape parameters $\boldsymbol{\beta}_T \in \mathbb{R}^{T\times 10}$ and is optimised by a shape regression loss $\mathcal{L}^s_{reg}$:

$$\mathcal{L}^s_{reg} = \|\boldsymbol{\beta}_T - \boldsymbol{\beta}^{ISG}_T\|^2, \tag{5}$$

where $\boldsymbol{\beta}^{ISG}_T$ is the corresponding shape parameters generated by ISG. In this way, the shape encoder learns to regress discriminative 3D shape parameters for each video frame.

3.3 Difference-Aware Shape Aggregation

Existing temporal aggregation methods [5, 41, 47] are usually proposed for aggregating appearance information, and insensitive to the shape differences over time in a video. They thus aggregate much redundant shape information of different frames and suppress valuable unique shape information of each frame. To this end, we propose a Difference-aware 3D Shape Aggregation (DSA) module, which takes advantage of the shape differences among frames to drive the shape aggregation with suppressed redundancy and enhanced complementarity.

To make our method more sensitive to inter-frame shape differences over time, we use relative shape instead of absolute shape per frame in video shape aggregation. As shown in Section. 3.1, each shape parameter $\boldsymbol{\beta}^d$ controls some specific aspects of body shape, so we can formulate the subtle difference of 3D shapes among frames by the difference values of shape parameters. As shown in Fig. 2. (2), we first compute the mean shape parameters $\bar{\boldsymbol{\beta}}_T$ of a tracklet as a reference, and then obtain the shape-difference map $\boldsymbol{\beta}_{dif} = \boldsymbol{\beta}_T - \bar{\boldsymbol{\beta}}_T$ ($\boldsymbol{\beta}_{dif} \in \mathbb{R}^{T\times 10}$). We denote the value at the coordinate (t, d) as $(\boldsymbol{\beta}_{dif})^d_t$ indicating the shape difference between the d-th shape parameter of the t-th frame and the corresponding mean parameter. We introduce the intra-frame and inter-frame shape-difference references to jointly decide the weight for each shape parameter of each frame.

Intra-frame shape-difference reference. We consider a shape parameter with a larger difference to the mean parameter to be more informative than other parameters in that frame. But if most shape parameters have larger differences, their importance should be balanced because it is possibly caused by the body occlusion or shape estimation error. Therefore we introduce an intra-frame shape-difference reference $\boldsymbol{w}_D \in \mathbb{R}^{T\times 10}$ to consider all the shape parameters within a frame to balance the weight of each one. $\boldsymbol{w}_D$ for the t-th frame is formulated as

$$(\boldsymbol{w}_D)^1_t, \cdots, (\boldsymbol{w}_D)^{10}_t = \text{Sigmoid}(\text{Conv}[(\boldsymbol{\beta}_{dif})^1_t, \cdots, (\boldsymbol{\beta}_{dif})^{10}_t]), \tag{6}$$

where Sigmoid is the Sigmoid function, Conv is a convolutional layer, of which the kernel size is 1×10 to span all of 10 shape parameters of a frame. The output channel is 10 for 10 different parameters. Details are illustrated in Fig. 2. (2).

Inter-frame shape-difference reference. To reduce the temporal redundant information, we also introduce an inter-frame shape-difference reference $\boldsymbol{w}_T \in \mathbb{R}^{T\times 10}$ to compare each shape parameter across frames to assign temporal attention. Concretely, we compute an inter-frame shape-difference map

$\boldsymbol{\beta}'_{dif} \in \mathbb{R}^{T\times T\times 10}$, on which the value at the coordinate (t_1,t_2,d) is formulated as

$$(\boldsymbol{\beta}'_{dif})^d_{t_1,t_2} = (\boldsymbol{\beta}_{dif})^d_{t_1} - (\boldsymbol{\beta}_{dif})^d_{t_2}, \tag{7}$$

where t_1, t_2=1,$\cdots$,T; d=1,$\cdots$,10. $\boldsymbol{\beta}'_{dif}$ indicates the shape difference of the d-th shape parameter between the t_1-th and t_2-th frames. $\boldsymbol{w}_T$ is then formulated by 10 convolutional layers as

$$(\boldsymbol{w}_T)^d_t = \text{Sigmoid}(\text{Conv}_d([(\boldsymbol{\beta}'_{dif})^d_{t,1}, \cdots, (\boldsymbol{\beta}'_{dif})^d_{t,T}])), \tag{8}$$

where Conv_d is the d-th convolutional layer with a kernel size of 1 × 1, which considers all of T frames on the d-th parameter to determine the reference weight.

The reference weights $\boldsymbol{w}_D$ and $\boldsymbol{w}_T$ make our model sensitive to the shape changes both in respect to each shape parameter over time and all shape parameters in each frame at a time. They thus impose selective aggregation by minimising redundant spatial-temporal shape information across frames in a video. The final weight $\boldsymbol{w}_S = \boldsymbol{w}_T \odot \boldsymbol{w}_D$, where $\odot$ is the elementwise product, and then is normalized by the Softmax function. The aggregated videowise shape parameters $\boldsymbol{\beta}_s$ are the sum of $\boldsymbol{\beta}_T$ weighted by $\boldsymbol{w}_S$, where $\boldsymbol{\beta}_s$ is optimised by a shape-based identity loss $\mathcal{L}^s_{id}$, same as in Eq.(3), *i.e.*, the sum of the cross-entropy loss and triplet loss to learn discriminative videowise shape parameters.

3.4 Appearance and Shape Fusion

Appearance remains useful in complementing some visual similarities to shape for Re-ID, *e.g.*, when a person only changes partial clothes and/or with certain aspects of appearance unaffected by clothing changes, such as gender, age, etc. To that end, we formulate a joint appearance and shape fused representation that is adaptively learned in model training.

The appearance encoder extracts videowise appearance features $\boldsymbol{f}_a$, to be combined with videowise shape parameters $\boldsymbol{\beta}_s$. A fusion module includes two steps, *i.e.*, feature transformation and weight prediction. The feature transformation projects two feature vectors into a common feature space, defined as

$$\boldsymbol{f}_a \leftarrow \text{Sigmoid}(\text{Conv}_a(L_2(\boldsymbol{f}_a))), \boldsymbol{f}_s \leftarrow \text{Sigmoid}(\text{Conv}_s(L_2(\boldsymbol{\beta}_s))), \tag{9}$$

where L_2 is L_2 normalization, Conv_a and Conv_s are two independent convolutional layers with the kernel size of 1 × 1.

The weight prediction aims to estimate the weights for the two feature vectors by making them refer to each other and jointly optimise the weight for each one. This process is defined as

$$\boldsymbol{w}_a = \text{Conv}_a([\boldsymbol{f}_a, \boldsymbol{f}_s]), \quad \boldsymbol{w}_s = \text{Conv}_s([\boldsymbol{f}_a, \boldsymbol{f}_s]), \tag{10}$$

where $\boldsymbol{f}_a$ and $\boldsymbol{f}_s$ are concatenated as a tensor, which then separately goes forward through two convolutional layers Conv_a and Conv_s. They both have the kernel size of 1×2 and thus output two weight vectors $\boldsymbol{w}_a$ and $\boldsymbol{w}_s$. A fused joint feature vector $\boldsymbol{f}_j = \boldsymbol{w}_a \odot \boldsymbol{f}_a + \boldsymbol{w}_s \odot \boldsymbol{f}_s$, where $\odot$ is element-wise product, with $\boldsymbol{f}_a$ and $\boldsymbol{f}_j$ being optimised by the appearance-based loss $\mathcal{L}^a_{id}$ and fusion-based loss $\mathcal{L}^j_{id}$. Each of them is the sum of a cross-entropy loss and a triplet loss as in Eq.(3). The overall 3STA model is jointly trained by an overall loss

$$\mathcal{L}_{all} = \mathcal{L}^{\beta}_{reg} + \lambda_1\mathcal{L}^s_{id} + \lambda_2\mathcal{L}^a_{id} + \lambda_3\mathcal{L}^j_{id}, \tag{11}$$

Table 1. Comparison among CC Re-ID datasets. Some data are unclear because of being publicly unavailable.

Dataset	NKUP [40]	LTCC [34]	PRCC [44]	COCAS [45]	RRD-Campus [3]	Motion-ReID [46]	VCCR (Ours)
Type	**image**	**image**	**image**	**image**	**radio frequency**	**video**	**video**
CC IDs\Distractors	107\0	91\61	221\0	5,266\0	100\0	30\0	232\160
Cameras	15	12	3	30	5	2	23
Cropping	detection	detection	manual	manual	detection	manual	manual
Tracklets	-	-	-	-	863	240	4,384
Images (Frames)	9,738	17,119	33,698	62,383	unclear	24,480	152,610
Clothes/ID	2∼3	2∼14	2	2∼3	unclear	unclear	2∼10
Publicly Available	Y	Y	Y	N	N	N	Y (to be released)

where λ_1, λ_2 and λ_3 are weight factors. Discriminative appearance and shape representations, which are optimised by $\mathcal{L}_{id}^{a}$, $\mathcal{L}_{id}^{s}$ and $\mathcal{L}_{reg}^{\beta}$, are the foundations of contributing to a more discriminative joint representation optimised by $\mathcal{L}_{id}^{j}$.

4 VCCR Dataset

Given that there is no publicly available dataset for video-based CC Re-ID model learning and evaluation, we introduce a new Video-based Clothing-Change Re-ID (VCCR) dataset to be released for open access to the research community.

4.1 Collection and Labelling

We collect data from the Richly Annotated Pedestrian (RAP) dataset [24] for reducing the collection and annotation cost. Moreover, this does not lead to additional privacy issues by not introducing new data. RAP contains person images captured in an indoor shopping mall over 3 months. It was originally built for attribute recognition and annotated with identity labels. We select 232 persons that change clothes and 160 distractors that do not change clothes from the RAP dataset, with access to the corresponding videos given the permission and consent by the authors. Then we manually crop person patches from video frames. Apart from the identity and camera labels from RAP, we additionally annotate each tracklet with a clothing label. Two tracklets with the same identity label are given two different clothing labels only if there is a visible clothing change. A change of carrying items, such as bottles, books and boxes, does not affect the clothing label.

4.2 Statistics and Comparison

We compare VCCR with other CC Re-ID datasets in Table 1 in four aspects.
1) Type. Most existing datasets, *e.g.*, NKUP [40], LTCC [34], PRCC [44] and COCAS [45], are image-based. RRD-Campus [3] collects radio frequency signals. Only Motion-ReID [46] is video-based, but not publicly available.
2) Scale. Motion-ReID includes 240 tracklets, while VCCR has 4,087 tracklets of 232 clothing-change persons and 297 tracklets of 160 distractors. Each tracklet

has 5 to 130 frames with an average of 35. VCCR is thus the currently largest video-based CC Re-ID dataset.
3) Number of persons. Since it is much more challenging to collect and label clothing-change data in videos than images, VCCR has a smaller number of clothing-change persons (232) than the image-based COCAS (5,266). But VCCR still contains more clothing-change persons than the video-based Motion-ReID (30), and also competitive compared with most image-based datasets, such as NKUP (107), LTCC (91) and PRCC (221).
4) Number of clothes. In contrast to NKUP, PRCC and COCAS with 2∼3 clothes/ID, VCCR contains 2∼10 clothes/ID with an average of 3.3.

4.3 Protocol

The training set includes 2,873 tracklets of 150 clothing-change persons. For test, 496 tracklets of 82 clothing-change persons constitute the query set, while 718 tracklets of these 82 persons along with 297 tracklets of 160 distractors build the gallery set. We make sure that the training and test sets have close statistics and diversity in samples. We adopt two test modes like [34], *i.e.*, the cloth-changing (CC) and standard modes, to evaluate the performance of CC Re-ID models. In the **clothing-change (CC)** mode, all the ground-truth gallery tracklets have different clothing labels to the query. In the **standard** mode, the ground-truth gallery tracklets can have either same or different clothing labels to the query. When evaluating Re-ID performances, we use the average cumulative match characteristic and report results at ranks 1, 5 and 10.

5 Experiments

5.1 Implementation Details

The appearance encoder adopts the Resnet-50 [11] backbone pretrained on ImageNet [2] to extract framewise appearance features and average pooling to produce 2048D videowise features $\boldsymbol{f}_a$. The shape encoder is composed of a pretrained Resnet-50 backbone and three fully-connected layers of 1024, 1024 and 10 dimensions, respectively. All images are scaled to 224×112 and randomly flipped.

We first run the ISG module with the VCCR dataset and 3D human dataset Human3.6M [19]. All the tracklets of VCCR are broken into 152,610 images in total. We randomly sample 16 persons with 4 images per person from VCCR, and 64 random images from Human3.6M in each training batch. ISG is performed for 20,000 iterations with the Adam optimiser [22] (β_1=0.9 and β_2=0.999). The learning rate is set to 0.00001 and the weight factor α is set to 500. After training ISG, we keep the generated shape parameters $\boldsymbol{\beta}_T^{ISG}$ for VCCR to train the overall 3STA model.

The overall 3STA model is trained on VCCR. We randomly choose 8 different persons, 4 tracklets for each person and 8 successive frames for each tracklet in each training batch. We also use the Adam optimiser, with the learning rates of the shape encoder and other modules initialized at 0.00001 and 0.0001, respectively, and decayed by 0.1 after 20,000 iterations. The 3STA model is jointly trained over 30,000 epochs. The weight factor λ_1=1, λ_2=10, λ_3=0.05, and the margin parameters of all the used triplet losses are set to 0.3. The dimension of the projected feature space in the appearance and shape fusion module is 2048, *i.e.*, $\boldsymbol{f}_a, \boldsymbol{f}_s, \boldsymbol{f}_j \in \mathbb{R}^{2048}$.

5.2 Evaluation on CC Re-Id Datasets

We compare our 3STA model and four types of state-of-the-art methods on the VCCR dataset in Table 2. In terms of the deep learning based methods, the results show a general trend that the performance is incrementally improved from image-based short-term, image-based CC, video-based short-term to video-based CC Re-ID methods. Specifically, first, **image-based short-term Re-ID** methods have the lowest accuracies, because they primarily make use of clothing information to discriminate persons and inevitably lose some discriminability under clothing-change situations. Second, **image-based CC Re-ID** methods reduce the reliance on clothing by exchanging clothes among persons [35] or using vector-neuron capsules to perceive clothing change of the same person [18]. Third, **video-based short-term** Re-ID methods have more robust Re-ID capacities due to exploiting temporal information, but they are still sensitive to clothing changes. Overall, our **video-based CC Re-ID** model 3STA achieves the highest accuracies in both CC and standard modes. The reasons are two-fold. 1) Discriminative temporal 3D shape information in videos is modelled as clothing-independent person characteristics. 2) Complementary appearance information is jointly modelled with 3D shape, resulting in the joint representation more robust to both clothing-change and clothing-consistent situations.

For completeness, we also list the released results of STFV3D [27], DynFV [6] and FITD [46] on the Motion-ReID dataset [46]. All of these methods are based on hand-crafted features. We are unable to compare them with other methods on Motion-ReID because the dataset is not publicly available, but we include a comparison on the video-based short-term Re-ID dataset PRID.

5.3 Evaluation on Short-Term Re-Id Datasets

We also conduct evaluations on the video-based short-term Re-ID datasets MARS [48] and PRID [12]. MARS is a large-scale dataset containing 1,261 persons with 20,715 tracklets. PRID includes 200 persons captured by two cameras, and only 178 persons with more than 25 frames are used, following the previous work [46].

Our 3STA model can perform clearly better than **image-based short-term and CC Re-ID** methods on both MARS and PRID, benefiting from modelling temporal 3D shape apart from clothing information. But the **video-based short-term Re-ID** methods can surpass our 3STA model, due to enhancing clothing based temporal information for better discriminating clothing-consistent persons. For the two **video-based CC Re-ID** models, 3STA significantly outperforms FITD on PRID. FITD utilizes motion cues for Re-ID with the

Table 2. Results on video-based CC Re-ID datasets VCCR and Motion-ReID, and short-term Re-ID datasets MARS and PRID (%). Image-based methods produce videowise features by average pooling on framewise features. Appe., Shape and Joint denote that appearance features $\boldsymbol{f}_a$, shape parameters $\boldsymbol{\beta}_s$ and joint features $\boldsymbol{f}_j$ in the 3STA model are used for evaluation (the same below).

Method Type	Methods	Features	VCCR (CC Mode)			VCCR (Standard Mode)			Motion-ReID	MARS	PRID
			mAP	Rank1	Rank5	mAP	Rank1	Rank5	Rank1	Rank1	Rank1
Image-Based Short-Term Re-ID	PCB [36]	Deep Learning	15.6	18.8	38.6	36.6	55.6	75.2	-	85.2	89.1
	MGN [39]	Deep Learning	22.6	23.6	44.9	42.7	64.4	81.9	-	86.4	90.6
	HPM [4]	Deep Learning	19.4	23.1	42.9	39.5	58.3	78.7	-	87.9	90.3
Video-Based Short-Term Re-ID	STFV3D [27]	Hand-Crafted	-	-	-	-	-	-	29.1	-	42.1
	DynFV [6]	Hand-Crafted	-	-	-	-	-	-	32.3	-	17.6
	MGH [43]	Deep Learning	30.7	34.6	54.5	51.6	76.3	87.2	-	90.0	94.8
	AP3D [7]	Deep Learning	31.6	35.9	55.8	52.1	78.0	88.4	-	**90.1**	94.6
	GRL [28]	Deep Learning	31.8	35.7	55.3	51.4	76.9	88.2	-	89.8	**95.1**
Image-Based Clothing-Change Re-ID	ReIDCaps [18]	Deep Learning	29.9	33.4	53.6	48.4	75.1	86.3	-	83.2	88.0
	SPS [35]	Deep Learning	30.5	34.5	54.1	50.6	76.5	85.5	-	82.8	87.4
Video-Based Clothing-Change Re-ID	FITD [46]	Hand-Crafted	-	-	-	-	-	-	**43.8**	-	58.7
	Appe. (3STA)	Deep Learning	29.3	32.8	52.0	46.7	74.3	84.5	-	83.7	87.8
	Shape (3STA)	Deep Learning	20.6	21.3	36.9	39.2	62.8	82.4	-	74.0	76.3
	Joint (3STA)	Deep Learning	**36.2**	**40.7**	**58.7**	**54.3**	**80.5**	**90.2**	-	89.1	93.4

assumption that people keep constant motion patterns, which does not always hold in practice. In contrast, 3STA explores discriminative 3D shape together with appearance, which is more stable and robust than motion cues.

5.4 Ablation Study

Appearance v.s. Shape v.s. Joint Representations. We compare the performance of appearance, 3D shape and joint representations in Table 2. The results show two phenomenons that deserve the attention. **1)** Appearance can achieve higher performance than 3D shape in both test modes, due to two reasons. First, when people do not change or just slightly change clothes, appearance remains more competitive than 3D shape by exploiting visual similarities for Re-ID. Second, 3D shape parameters only have 10 dimensions and they have a limited capacity of modelling complex body shape. Overall, 3D shape is best complemented with appearance instead of being used alone. **2)** The joint representations outperform both appearance and 3D shape by a significant margin. This demonstrates that our model can exploit the complementarity of two information to adaptively fuse more discriminative information, which can adapt to both cloth-changing and clothing-consistent situations better.

Identity-Aware 3D Shape Generation (ISG). The ISG module ensures the validity and discriminability of the generated 3D shape parameters by the loss $\mathcal{L}^{\beta}_{val}$ and $\mathcal{L}^{\beta}_{dis}$ in Eq. (4), respectively. We remove either of the two losses during performing ISG and show the results in Table 3 (top two lines). **1)** Removing $\mathcal{L}^{\beta}_{val}$ degenerates the accuracy of shape representations from 21.3%/62.8% to 12.7%/57.4% in the CC/starndard mode. Losing the supervision from the 3D human dataset in validity makes the shape parameters not formulate true 3D

Table 3. Rank 1 accuracy on the VCCR dataset for the ablation study of losses.

ISG module		3STA model				CC Mode			Standard Mode		
$\mathcal{L}^{\beta}_{val}$	$\mathcal{L}^{\beta}_{dis}$	$\mathcal{L}^{s}_{reg}$	$\mathcal{L}^{a}_{id}$	$\mathcal{L}^{s}_{id}$	$\mathcal{L}^{j}_{id}$	Appe.	Shape	Joint	Appe.	Shape	Joint
✗	✓	✓	✓	✓	✓	32.2	12.7	31.5	73.5	57.4	72.6
✓	✗	✓	✓	✓	✓	32.7	6.3	24.1	73.9	18.7	67.3
✓	✓	✗	✓	✓	✓	31.9	14.6	26.3	73.6	44.7	64.3
✓	✓	✓	✗	✓	✓	14.5	19.7	17.6	46.2	62.5	55.7
✓	✓	✓	✓	✗	✓	32.6	15.8	29.8	73.4	57.6	72.3
✓	✓	✓	✓	✓	✗	**33.5**	**21.6**	30.3	**75.2**	62.6	73.5
✓	✓	✓	✓	✓	✓	32.8	21.3	**40.7**	74.3	**62.8**	**80.5**

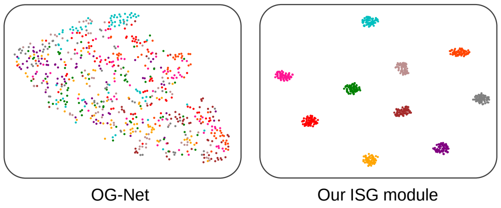


Fig. 3. Comparing 3D shape parameters generated by the OG-Net [49] and our ISG module. The visualisation shows 700 images of 10 persons from the VCCR dataset.

body shapes. Only using identity supervision from the Re-ID dataset for training, the model implicitly relies on appearance instead of 3D shape in minimising the loss $\mathcal{L}^{\beta}_{dis}$. **2)** Removing $\mathcal{L}^{\beta}_{dis}$ decreases the rank 1 of shape representations to 6.3%/18.7% in the CC/standard mode. In Fig. 3, we visualize the shape parameters generated by the ISG and the OG-Net [49]. OG-Net does not embed identity information, similar to removing $\mathcal{L}^{\beta}_{dis}$ from ISG. ISG enables 3D shape parameters to be separable for different persons, attributed to introducing $\mathcal{L}^{\beta}_{dis}$ to significantly improve discriminative 3D shape learning for Re-ID.

Difference-Aware 3D Shape Aggregation (DSA). 1) Intra-frame and inter-frame shape-difference references. The weight $\boldsymbol{w}_S$ in DSA is decided jointly by the intra-frame and inter-frame shape-difference references $\boldsymbol{w}_D$ and $\boldsymbol{w}_T$. As shown in the 2nd and 3rd rows in Table 4, using $\boldsymbol{w}_T$ alone degenerates the rank 1 accuracy of shape representations from 21.3%/62.8% to 18.2%/60.7% in the CC/standard mode. The joint representations are affected similarly. The 1st and 3rd rows suggest that using $\boldsymbol{w}_D$ alone degrades the rank 1 accuracy of shape representations by 3.4%/4.0% in the CC/standard mode. The joint use of $\boldsymbol{w}_D$ and $\boldsymbol{w}_T$ makes DSA sensitive to the changes of spatial and temporal shape information and reduces the redundancy of shape aggregation over time.

2) Shape differences. DSA uses the shape differences among frames ($\boldsymbol{\beta}_{dif} = \boldsymbol{\beta}_T - \bar{\boldsymbol{\beta}}_T$) to guide the weight prediction. To validate the effectiveness, we also show the result of directly using $\boldsymbol{\beta}_T$ to replace $\boldsymbol{\beta}_{dif}$ when computing $\boldsymbol{w}_D$ by Eq. (6). Comparing the 1st and 4th, or 3rd and 5th rows in Table 4, we can observe

Table 4. Rank 1 accuracy of different temporal aggregation methods on the VCCR dataset. $\boldsymbol{\beta}_{dif} = \boldsymbol{\beta}_T - \bar{\boldsymbol{\beta}}_T$.

Methods	Shape	Weight	CC Mode		Standard Mode	
			Shape	Joint	Shape	Joint
Ours	$\boldsymbol{\beta}_{dif}$	$\boldsymbol{w}_D$	17.9	36.6	58.8	78.4
	$\boldsymbol{\beta}_{dif}$	$\boldsymbol{w}_T$	18.2	39.0	60.7	79.7
	$\boldsymbol{\beta}_{dif}$	$\boldsymbol{w}_D \odot \boldsymbol{w}_T$	**21.3**	**40.7**	**62.8**	**80.5**
	$\boldsymbol{\beta}_T$	$\boldsymbol{w}_D$	16.2	35.8	58.5	77.3
	$\boldsymbol{\beta}_T$	$\boldsymbol{w}_D \odot \boldsymbol{w}_T$	17.5	36.4	59.4	77.8

using $\boldsymbol{\beta}_{dif}$ improves the rank 1 accuracy obviously over $\boldsymbol{\beta}_T$. The advantage of $\boldsymbol{\beta}_{dif}$ lies in helping the DSA module explicitly perceive the subtle 3D shape differences among frames in a form of relative shape, and reduce the reuse of redundant shape information better.

Losses. We perform the ablation study on the losses $\mathcal{L}_{reg}^s$, $\mathcal{L}_{id}^a$, $\mathcal{L}_{id}^s$ and $\mathcal{L}_{id}^j$ in training the 3STA model and show the results in Table 3. Taking off $\mathcal{L}_{reg}^s$ greatly decreases the accuracy of shape representations from 21.3%/62.8% to 14.6%/44.7% in the CC/standard mode, and also affects adversely the joint representations in a similar way. This is because $\mathcal{L}_{reg}^s$ can enable effective framewise shape learning, which is the basis of temporal shape aggregation. Our model also suffers from performance degradation in different degrees if trained without $\mathcal{L}_{id}^a$, $\mathcal{L}_{id}^s$ or $\mathcal{L}_{id}^j$. This reveals that the discriminative joint representations have to be built on discriminative appearance and shape representations.

6 Conclusion

To our best knowledge, for the first time this paper has formulated a model to learn discriminative temporal 3D shape information for video-based CC Re-ID. First, our proposed 3STA model has included an ISG module, which considers identity modelling to generate the discriminative 3D shape for each frame. Then, a DSA module that is sensitive to the shape differences among frames has been proposed to aggregate framewise shape representations into videowise ones. It selectively exploits the unique shape information of each frame to reduce the redundancy of shape aggregation. Moreover, we have also contributed a VCCR dataset for the video-based CC Re-ID research community.

Acknowledgements This work was jointly supported by National Key Research and Development Program of China Grant No. 2018AAA0100400, National Natural Science Foundation of China (62236010, 62276261, 61721004, and U1803261), Key Research Program of Frontier Sciences CAS Grant No. ZDBS-LY- JSC032, Beijing Nova Program (Z201100006820079), CAS-AIR, the fellowship of China postdoctoral science foundation (2022T150698), China Scholarship Council, Vision Semantics Limited, and the Alan Turing Institute Turing Fellowship.

References

1. Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., Zheng, W.S.: Learning 3d shape feature for texture-insensitive person re-identification. In: CVPR (2021)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
3. Fan, L., Li, T., Fang, R., Hristov, R., Yuan, Y., Katabi, D.: Learning longterm representations for person re-identification using radio signals. In: CVPR (2020)
4. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.S.: Horizontal pyramid matching for person re-identification. In: AAAI (2019)
5. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104 (2018)
6. Gou, M., Zhang, X., Rates-Borras, A., Asghari-Esfeden, S., Sznaier, M., Camps, O.: Person re-identification in appearance impaired scenarios. arXiv preprint arXiv:1604.00367 (2016)
7. Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X.: Appearance-preserving 3d convolution for video-based person re-identification. In: ECCV (2020)
8. Han, K., Huang, Y., Chen, Z., Wang, L., Tan, T.: Prediction and recovery for adaptive low-resolution person re-identification. In: ECCV (2020)
9. Han, K., Huang, Y., Song, C., Wang, L., Tan, T.: Adaptive super-resolution for person re-identification with low-resolution images. PR (2021)
10. Han, K., Si, C., Huang, Y., Wang, L., Tan, T.: Generalizable person re-identification via self-supervised batch norm test-time adaption (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: SCIA (Scandinavian Conference on Image Analysis) (2011)
13. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: CVPR (2021)
14. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Sbsgan: Suppression of inter-domain background shift for person re-identification. In: ICCV (2019)
15. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: ICCV (2021)
16. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Unsupervised domain adaptation with background shift mitigating for person re-identification. IJCV (2021)
17. Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z., Zhang, J.: Multi-pseudo regularized label for generated data in person re-identification. TIP (2018)
18. Huang, Y., Xu, J., Wu, Q., Zhong, Y., Zhang, P., Zhang, Z.: Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. TCSVT (2019)
19. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI (2013)
20. Isobe, T., Zhu, F., Wang, S.: Revisiting temporal modeling for video super-resolution. arXiv preprint arXiv:2008.05765 (2020)
21. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

23. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
24. Li, D., Zhang, Z., Chen, X., Huang, K.: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. TIP (2018)
25. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: CVPR (2019)
26. Li, Y.J., Luo, Z., Weng, X., Kitani, K.M.: Learning shape representations for clothing variations in person re-identification. arXiv preprint arXiv:2003.07340 (2020)
27. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: ICCV (2015)
28. Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X.: Watching you: Global-guided reciprocal learning for video-based person re-identification. In: CVPR (2021)
29. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG(Transactions On Graphics) (2015)
30. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. TIP (2020)
31. Niu, K., Huang, Y., Wang, L.: Fusing two directions in cross-domain adaption for real life person search by language. In: ICCV Workshops (2019)
32. Niu, K., Huang, Y., Wang, L.: Textual dependency embedding for person search by language. In: ACM MM (2020)
33. Pathak, P., Eshratifar, A.E., Gormish, M.: Video person re-id: Fantastic techniques and where to find them. arXiv preprint arXiv:1912.05295 (2019)
34. Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Long-term cloth-changing person re-identification. arXiv preprint arXiv:2005.12633 (2020)
35. Shu, X., Li, G., Wang, X., Ruan, W., Tian, Q.: Semantic-guided pixel sampling for cloth-changing person re-identification. IJIS (2021)
36. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV (2018)
37. Uddin, M.K., Lam, A., Fukuda, H., Kobayashi, Y., Kuno, Y.: Fusion in dissimilarity space for rgb-d person re-identification. Array (2021)
38. Wan, F., Wu, Y., Qian, X., Chen, Y., Fu, Y.: When person re-identification meets changing clothes. In: CVPRW (2020)
39. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM (2018)
40. Wang, K., Ma, Z., Chen, S., Yang, J., Zhou, K., Li, T.: A benchmark for clothes variation in person re-identification. IJIS (2020)
41. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid spatial-temporal aggregation for video-based person re-identification. In: ICCV (2021)
42. Wu, A., Zheng, W.S., Lai, J.H.: Robust depth-based person re-identification. TIP (2017)
43. Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., Shao, L.: Learning multi-granular hypergraphs for video-based person re-identification. In: CVPR (2020)
44. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. TPAMI (2019)
45. Yu, S., Li, S., Chen, D., Zhao, R., Yan, J., Qiao, Y.: Cocas: A large-scale clothes changing person dataset for re-identification. In: CVPR (2020)
46. Zhang, P., Wu, Q., Xu, J., Zhang, J.: Long-term person re-identification using true motion from videos. In: WACV (2018)

47. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In: CVPR (2020)
48. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV (2016)
49. Zheng, Z., Zheng, N., Yang, Y.: Parameter-efficient person re-identification in the 3d space. arXiv preprint arXiv:2006.04569 (2020)