Long-Term Cloth-Changing Person Re-identification

Xuelin Qian[1], Wenxuan Wang[1], Li Zhang[2], Fangrui Zhu[2], Yanwei Fu[⋆2], Tao Xiang[3], Yu-Gang Jiang[1], and Xiangyang Xue[1,2]

[1] School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
{xlqian15,wxwang19,ygj,xyxue}@fudan.edu.cn
[2] School of Data Science, and MOE Frontiers Center for Brain Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
{lizhangfd,18210980021,yanweifu}@fudan.edu.cn
[3] University of Surrey
t.xiang@surrey.ac.uk

Abstract. Person re-identification (Re-ID) aims to match a target person across camera views at different locations and times. Existing Re-ID studies focus on the short-term cloth-consistent setting, under which a person re-appears in different camera views with the same outfit. A discriminative feature representation learned by existing deep Re-ID models is thus dominated by the visual appearance of clothing. In this work, we focus on a much more difficult yet practical setting where person matching is conducted over long-duration, *e.g.*, over days and months and therefore inevitably under the new challenge of changing clothes. This problem, termed Long-Term Cloth-Changing (LTCC) Re-ID is much understudied due to the lack of large scale datasets. The first contribution of this work is a new LTCC dataset containing people captured over a long period of time with frequent clothing changes. As a second contribution, we propose a novel Re-ID method specifically designed to address the cloth-changing challenge. Specifically, we consider that under cloth-changes, soft-biometrics such as body shape would be more reliable. We, therefore, introduce a shape embedding module as well as a cloth-elimination shape-distillation module aiming to eliminate the now unreliable clothing appearance features and focus on the body shape information. Extensive experiments show that superior performance is achieved by the proposed model on the new LTCC dataset. The dataset is available on the project website: https://naiq.github.io/LTCC_Perosn_ReID.html.

1 Introduction

Person re-identification (Re-ID) aims at identifying and associating a person at different locations and times monitored by a distributed camera network. It underpins many crucial applications such as multi-camera tracking [1], crowd counting [2], and multi-camera activity analysis [3].

⋆ corresponding author

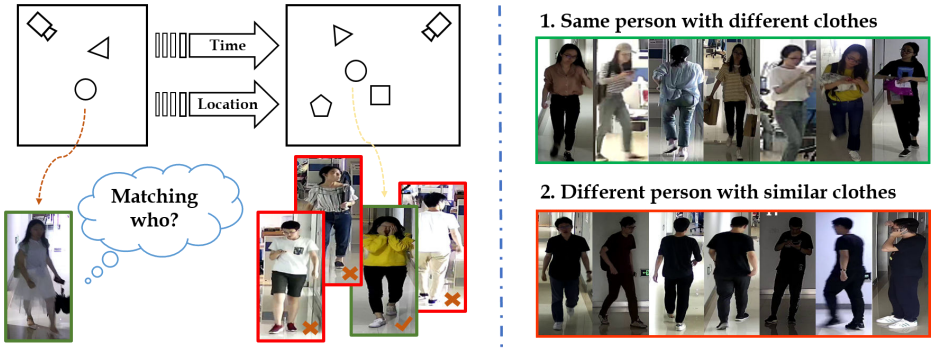

Fig. 1. Illustration of the long-term cloth-changing Re-ID task and dataset. The task is to match the same person under cloth-changes from different views, and the dataset contains same identities with diverse clothes.

Re-ID is inherently challenging as a person's appearance often undergoes dramatic changes across camera views. Such a challenge is captured in existing Re-ID datasets, including Market-1501 [4], DukeMTMC [5] and CUHK03 [6], where the appearance changes are caused by changes in body pose, illumination [7], occlusion [8–10] and camera view angle [11, 12]. However, it is noted that in these popular datasets each person is captured across different camera views within a short period of time on the same day. As result, each wears the same outfit. We call the Re-ID under this setting as short-term cloth-consistent (STCC) Re-ID. This setting has a profound impact on the design of existing Re-ID models – most existing models are based on deep neural networks (DNNs) that extract feature representations invariant to these changes. Since faces are of too low resolutions to be useful, these representations are dominated by the clothing appearance.

In this paper, we focus on a more challenging yet practical new Long-Term Cloth-Changing (LTCC) setting (see Fig. 1 (a)). Under this setting, a person is matched over a much longer period of time, *e.g.*, days or even months. As a result, clothing changes are commonplace, though different parts of the outfit undergo changes at different frequencies – the top is more often than the bottom and shoe changes are the least common. Such a setting is not only realistic but also crucial for certain applications. For instance, if the objective is to capture a criminal suspect after a crime is committed, he/she may change outfits even over short-term as a disguise. However, LTCC Re-ID is largely ignored so far. One primary reason is the lack of large-scale datasets featuring clothing changes.

The first contribution of this paper is thus to provide such a dataset to fill in the gap between research and real-world applications. The new Long-Term Cloth-Changing (LTCC) dataset (see Fig. 1 (b)) is collected over two months. It contains 17, 138 images of 152 identities with 478 different outfits captured from 12 camera views. This dataset was completed under pedestrian detection and

careful manual labeling of both person identities and clothing outfit identities. Due to the long duration of data collection, it also features drastic illumination, viewing angle, and pose changes as in existing STCC datasets, but additionally clothing and carrying changes and occasionally hairstyle changes. Furthermore, it also includes many persons wearing similar clothing, as shown in Fig.1 (b).

With cloth-changing now commonplace in LTCC Re-ID, existing Re-ID models are expected to struggle (see Tab. 1) because they assume that the clothing appearance is consistent and relies on clothing features to distinguish people from each other. Instead, it is now necessary to explore identity-relevant biological traits (*i.e.*, soft-biometrics) rather than cloth-sensitive appearance features. A number of naive approaches can be considered. First, can a DNN 'does the magic' again, *i.e.*, figures out automatically what information is cloth-change-invariant? The answer is no, because a) the soft-biometrics information is subtle and hard to compute without any network architectural change to assist in the extraction; and b) in a real-world, some people will wear the same outfit or at least keep part of the outfit unchanged (*e.g.*, shoes) even over a long duration. The network thus receives a mixed signal regarding what information is discriminative. Second, would adding a cloth-changing detector be a solution so that models can be switched accordingly? The answer is also negative since detecting changes for the same person needs person Re-ID to be solved in the first place.

To overcome these difficulties, we propose a novel DNN for LTCC Re-ID. The key idea is to remove the cloth-appearance related information completely and only focus on view/pose-change-insensitive body shape information. To this end, we introduce a Shape Embedding (SE) to help shape feature extraction and a Cloth-Elimination Shape-Distillation (CESD) module to eliminate cloth-related information. Concretely, the SE module aims to extract body pose features. This is achieved by encoding the position/semantic information of human body joints, and leveraging the relation network [13] to explore the implicit correlations between each pair of body joints. The CESD module, on the other hand, is designed to learn identity-relevant biological representation. Based on the features extracted from SE module, we adaptively distill the shape information by re-scaling the original image feature. To better disentangle the identity-relevant features from the residual (*i.e.*, the difference between the original information and the re-scaled shape information), we explicitly add the clothing identity constrain to ensure the identity-irrelevant clothing feature to be eliminated.

Contribution. Our contributions are as follows: (1) We introduce a new Long-Term Cloth-Changing (LTCC) dataset, designed to study the more challenging yet practical LTCC Re-ID problem. (2) We propose a novel model for LTCC Re-ID. It contains a shape embedding module that efficiently extracts discriminative biological structural feature from keypoints, and a cloth-elimination shape-distillation module that learns to disentangle identity-relevant features from the cloth-appearance features. (3) Extensive experiments validate the efficacy of our Re-ID models in comparison with existing Re-ID models.

2 Related Work

Short-Term Cloth-Consistent Re-ID. With the popularization of surveillance system in the real-world, person re-identification task attracts more and more attention. As mentioned earlier, almost all existing Re-ID datasets [14, 15, 4] were captured during short-period of time. As a result, for the same person, the clothing appearances are more or less consistent. In the deep-learning era, more efforts have been made in developing approaches for automatic person re-identification by learning discriminative features [16, 17] or robust distance metrics [18, 19]. These models are robust against changes caused by pose, illumination and view angles as they are plenty in those datasets. However, they are vulnerable to clothing changes as the models are heavily reliant on the clothing appearance consistency.

Long-Term Cloth-Changing Re-ID Datasets. Barbosa *et al.* [20] proposed the first cloth-changing Re-ID dataset. However, the dataset is too small to be useful for deep learning based Re-ID. More recently iQIYI-VID [21] is introduced which is not purposefully built for LTCC but does contain some cloth-changing images. However, it is extracted from on-line videos, unrepresentative of real-world scenarios and lack of challenges caused illumination change and occlusion. Similarly, Huang *et al.* [22, 23] collect Celebrities-ReID dataset containing clothing variations. Nevertheless, the celebrity images are captured in high quality by professional photographers, so unsuited for the main application of Re-ID, i.e. video surveillance using CCTV cameras. Another related work/dataset is Re-ID in photo albums [24, 25]. It involves person detection and recognition with high-resolution images where face is the important clue to solve this task. However, this setting is not consistent with canonical Person Re-ID, where pedestrians are captured by non-overlapped surveillances with low resolution. Note that a concurrent work [26] also introduced an LTCC dataset, called PRCC which is still not released. Though featuring outfit changes, PRCC is a short-term cloth-changing dataset so it contains less drastic clothing changes and bare hairstyle changes. Further, with only 3 cameras instead of 12 in our LTCC dataset, it is limited in view-angle and illumination changes. In contrast, our LTCC Re-ID aims at matching persons over long time from more cameras, with more variations in visual appearance, *e.g.*, holding different bags or cellphones, and wearing hats as shown in Fig. 1(b).

Long-Term Cloth-Changing Re-ID Models. Recently, Xue *et al.* [27] particularly address the cloth-change challenge by downplaying clothing information and emphasizing face. In our dataset, face resolution is too low to be useful and we take a more extreme approach to remove clothing information completely. Zheng *et al.* [17] propose to switch the appearance or structure codes and leverage the generated data to improve the learned Re-ID features. Yang *et al.* [26] introduce a method to capture the contour sketch representing the body features. These two works mainly focus on learning invariant and discriminative Re-ID features by taking clothing information into account. In contrast, our model focuses solely on extracting soft-biometrics features and removes the model dependency on clothing information completely.

Fig. 2. Examples of one person wearing the same and different clothes in LTCC dataset. There exist various illumination, occlusion, camera view, carrying and pose changes.

3 Long-Term Cloth-Changing (LTCC) Dataset

Data Collection. To facilitate the study of LTCC Re-ID, we collect a new Long-Term Cloth-Changing (LTCC) person Re-ID dataset. Different from previous datasets [4, 28, 5, 6], this dataset aims to support the research of long-term person re-identification with the added challenge of cloth changes. During dataset collection, an existing CCTV network is utilized which is composed of twelve cameras installed on three floors in an office building. A total of 24-hour videos are captured over two months. Person images are then obtained by applying the Mask-RCNN [29] detector.

Statistics. We release the first version of LTCC dataset with this paper, which contains 17,138 person images of 152 identities, as shown in Fig. 2. Each identity is captured by at least two cameras. To further explore the cloth-changing Re-ID scenario, we assume that different people will not wear identical outfits (however visually similar they may be), and annotate each image with a cloth label as well. Note that the changes of the hairstyle or carrying items, *e.g.*, hat, bag or laptop, do not affect the cloth label. Finally, dependent on whether there is a cloth-change, the dataset can be divided into two subsets: one cloth-change set where 91 persons appearing with 417 different sets of outfits in 14,756 images, and one cloth-consistent subset containing the remaining 61 identities with 2,382 images without outfit changes. On average, there are 5 different clothes for each cloth-changing person, with the numbers of outfit changes ranging from 2 to 14.

Comparison with Previous Datasets. Comparing to the most widely-used cloth-consistent Re-ID datasets such [4, 6, 5], our dataset is gathered primarily to address the new LTCC Re-ID task, which challenges the Re-ID under a more realistic yet difficult setting. Comparing to few cloth-changing datasets [30, 22],

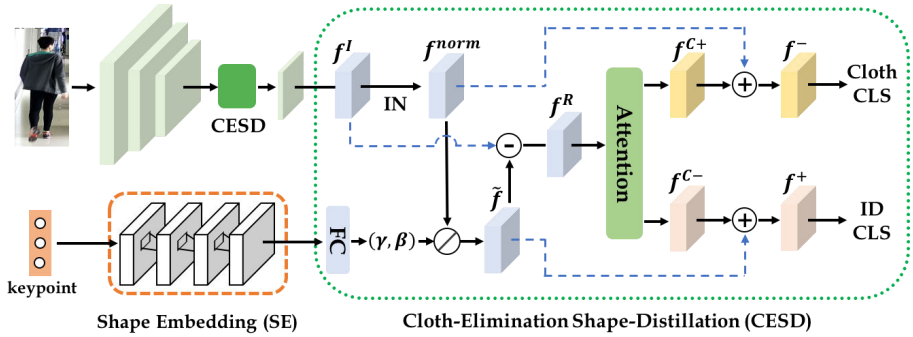

Fig. 3. Illustration of our framework and the details of Cloth-Elimination Shape-Distillation (CESD) module. Here, we introduce Shape Embedding (SE) module to extract structural features from human keypoints, followed by learning identity-sensitive and cloth-insensitive representations using the CESD module. There are two CESD modules (green solid box and green dash box). Both share the same SE module, but we only show the detail of CESD in the second one. '⊘' denotes the operation of re-scale in the Eq. 4.

ours is collected in a natural and long-term way without any human intervention. Concretely, our dataset includes not only various cloth-changing (*e.g.*, top, bottom, shoe-wear), but also diverse human pose (*e.g.*, squat, run, jump, push, bow), large changes of illumination (*e.g.*, from day to night, indoor lighting) and large variations of occlusion (*e.g.*, self-occlusion, partial), as shown in Fig. 2. More details can be found in Supplementary Material.

4 Methodology

Under the LTCC Re-ID setting, the clothing appearance becomes unreliable and can be considered as a distraction for Re-ID. We thus aim to learn to extract biological traits related to soft biometrics, with a particular focus on body shape information in this work. Unfortunately, learning the identity-sensitive feature directly from body shape [31] is a challenging problem on its own. Considering the recent success in body parsing or human body pose estimation [32] from RGB images, we propose to extract identity-discriminative shape information whilst eliminating clothing information with the help of an off-the-shelf body pose estimation model. Specifically, given a generic DNN backbone, we first introduce a Shape Embedding (SE) module to encode shape information from human body keypoints. Then we propose a Cloth-Elimination Shape-Distillation (CESD) module, which is able to utilize shape embedding to adaptively distill the identity-relevant shape feature and explicitly disentangle the identity-irrelevant clothing information.

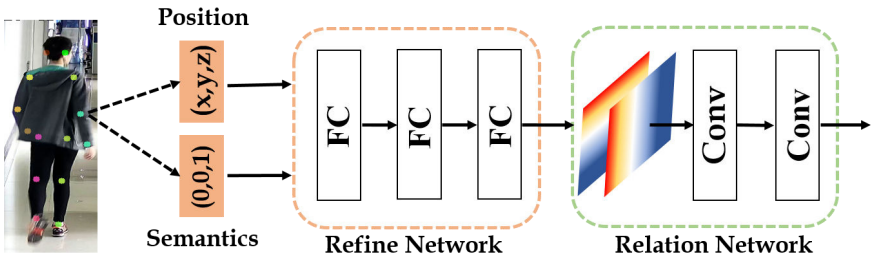

Fig. 4. The detailed structure of the Shape Embedding (SE) module.

4.1 Shape Embedding (SE)

Humans can easily recognize an old friend in an unseen outfit from the back (no face). We conjecture that this is because the body shape information is discriminative enough for person identification. Here, 'shape' is a general term referring to several *unique* biological traits, *e.g.*, stature and body-part proportion. One intuitive way to represent body shape is to employ joints of human body and model the relations between each pair of points. For example, the relation between points of left-shoulder and left-hip reflects the height of the upper body.

Representation. We employ an off-the-shelf pose detector [33, 34], which is trained without using any re-id benchmark data, to detect and localize n human body joints as shown in Fig. 4. Each point n_i is represented by two attributes, position P_i and semantics S_i (*i.e.*, corresponding to which body joint). Concretely, $P_i = (\frac{x_i}{w}, \frac{y_i}{h}, \frac{w}{h})$ where (x_i, y_i) denotes the coordinates of joint i in the person image and (w, h) represents the original width and height of the image. S_i is a n-dimensional one-hot vector to index the keypoint i (*e.g.*, head and knee). If one of the points is undetectable, we set P_i as $(-1, -1, \frac{w}{h})$.

Embedding. The two keypoint representation parts are first encoded with learnable embedding weights W individually. Then, we employ a refinement network to integrate the two parts and improve the representation of each keypoint, which can be formulated as

$$f_i = \mathcal{F}\left(W_p P_i^{\mathsf{T}} + W_s S_i^{\mathsf{T}}\right) \in \mathbb{R}^{d_2}, \tag{1}$$

where $W_p \in \mathbb{R}^{d_1 \times 3}$ and $W_s \in \mathbb{R}^{d_1 \times 13}$ are two different embedding weights; $\mathcal{F}(\cdot)$ denotes the refinement network with several fully-connected layers to increase the dimension from d_1 to d_2. In this paper, we have two hidden layers and set $d_1 = 128$, $d_2 = 2048$.

After the embedding, we now obtain a set of keypoint features $f \in \mathbb{R}^{n \times d_2}$. Intuitively, the information of body proportion cannot be captured easily by the feature of a single joint. We still need to represent the relation between each pair of keypoints. To this end, we propose to leverage a relation network for exploiting relations between every two points. As illustrated in Fig. 4, our relation network concatenates features of two different keypoints from all combinations

and feeds them into two convolution layers for relation reasoning. The final shape embedding feature of f^P is obtained by maximizing[4] over all outputs. The whole formulations can be expressed as follows,

$$H_{ij} = [f_i \ ; \ f_j], \qquad f^p = \mathrm{GMP}\left(Conv_\theta(H)\right) \tag{2}$$

where $f^P \in \mathbb{R}^{d_2}$ and $H \in \mathbb{R}^{n \times n \times d_2}$; GMP and θ denote global max pooling and parameters of convolution layers, respectively. $[*;*]$ denotes the operation of concatenation.

4.2 Cloth-Elimination Shape-Distillation (CESD)

Person images contain clothing and shape information. Disentangling the two and focusing on the latter is the main idea behind our model for LTCC Re-ID. Recently, many works on visual reasoning [35, 36] or style transfer [37, 38] are proposed to solve a related similar problem with adaptive normalization. Inspired by these works, we introduce a cloth-elimination shape-distillation module. It is designed to utilize the shape embedding to explicitly disentangle the clothing and discriminative shape information from person images[5]. The final extracted feature thus contains the information transferred from shape embedding and the ones from a task-driven attention mechanism. Specifically, we denote both the input image and shape features by $f^I \in \mathbb{R}^{h \times w \times c}$ and $f^P \in \mathbb{R}^{d_2}$, where h, w, c indicate the height, width and number of channel, respectively.

Shape Distillation. Inspired by adaptive instance normalization [37], we first try to reduce the original style information in the input by performing instance normalization [39]. Then, we distill the shape information by re-scaling the normalized feature with the parameters γ and β calculated from f^P. These two steps can be expressed as,

$$f^{norm} = \frac{f^I - \mathrm{E}\left[f^I\right]}{\sqrt{\mathrm{Var}\left[f^I\right] - \epsilon}}, \tag{3}$$

$$\begin{aligned} \widetilde{f} &= (1 + \Delta\gamma)\, f^{norm} + \beta, \\ &where\ \Delta\gamma = \mathcal{G}_s\left(f^P\right), \quad \beta = \mathcal{G}_b\left(f^P\right) \end{aligned} \tag{4}$$

where $\mathrm{E}\left[\cdot\right]$ and $\mathrm{Var}\left[\cdot\right]$ denote the mean and variance of image features calculated per-dimension separately for each sample; $\mathcal{G}_s\left(\cdot\right)$ and $\mathcal{G}_b\left(\cdot\right)$ are both one fully-connected layer to learn new parameters of scale and bias. Particularly, rather than directly predicting γ, we output the offset $\Delta\gamma$ in case of re-scaling factor value for the feature activation being too low.

[4] Max pooling is found to be more effective than alternatives such as avg pooling; possible reason is that it is more robust against body pose undergoing dramatic changes.

[5] We found empirically that directly using shape embeddings as Re-ID features leads to worse performance. A likely reason is that the detected 2D keypoints may be unreliable due to occlusion. So, we treat them as intermediate ancillary features.

Cloth Elimination. To further enhance the representation of identity-sensitive but cloth-insensitive feature, we propose to eliminate the identity-irrelevant clothing clue from the final image features. As shown in Fig. 3, given the image feature f^I and the transferred feature $\widetilde{f}$, we first obtain the residual feature $f^R \in \mathbb{R}^{h\times w\times c}$ by computing the difference between f^I and $\widetilde{f}$,

$$f^R = f^I - \widetilde{f} \tag{5}$$

For f^R, it inevitably includes some discriminative features (*e.g.*, contour) and features that are sensitive to cloth changes. Since reducing intra-distance of the same identity with different clothing is the primary objective here, we propose to leverage the self-attention mechanism to explicitly disentangle the residual feature into two parts, the cloth-irrelevant feature $f^{C-} \in \mathbb{R}^{h\times w\times c}$ and the cloth-relevant feature $f^{C+} \in \mathbb{R}^{h\times w\times c}$, which can be formulated as follows,

$$\alpha = \phi\left(\mathcal{G}^2\left(\mathrm{GAP}\left(f^R\right)\right)\right), \tag{6}$$

$$f^{C+} = \alpha f^R, \quad f^{C-} = (1-\alpha) f^R \tag{7}$$

where $\mathcal{G}^i$ denotes the i-th layer of a convolution neural network with ReLU activation function, GAP is the global average pooling operation, ϕ is a sigmoid activation function and $\alpha \in \mathbb{R}^{1\times 1\times c}$ is the learned attention weights.

By adding the cloth-irrelevant feature f^{C-} to the re-scaled shape feature $\widetilde{f}$, we add one more convolutional layer to refine it and obtain the identity-relevant feature of f^+. Analogously, we sum the cloth-relevant feature f^{C+} and the normalized feature f^{norm} followed by a different convolutional layer to get the final cloth-relevant feature f^-, that is,

$$f^+ = \mathrm{Conv}_\theta\left(\widetilde{f} + f^{C-}\right), \quad f^- = \mathrm{Conv}_\phi\left(f^{norm} + f^{C+}\right). \tag{8}$$

4.3 Architecture Details

Figure 3 gives an overview of our framework. Particularly, a widely-used network of ResNet-50 [40] is employed as our backbone for image feature embedding. We insert the proposed CESD module after *res3* and *res4* blocks. For each CESD module, we apply two classification losses (person ID and cloth ID) to support the learning of identity-relevant feature f^+ and cloth-relevant feature f^- respectively. Therefore, the overall loss of our framework is,

$$\mathcal{L} = \sum_{i=1}^{2} \lambda_i \mathcal{L}^i_{clothing} + \sum_{i=1}^{2} \mu_i \mathcal{L}^i_{id}, \tag{9}$$

where $\mathcal{L}^i_{clothing}$ and $\mathcal{L}^i_{id}$ denote the cross-entropy loss of clothing and identity classification from the i-th CESD module, respectively; λ_i and μ_i are both coefficients which control the contribution of each term. Intuitively, the feature at the deeper layer is more important and more relevant to the task, thus, we empirically set λ_1, λ_2 to $0.3, 0.6$ and μ_1, μ_2 to $0.5, 1.0$ in all experiments.

5 Experiment

5.1 Experimental Setup

Implementation details. Our model is implemented on the Pytorch framework, and we utilize the weights of ResNet-50 pretrained on ImageNet [41] for initialization. For key points, we first employ the model from [33, 34] for human joints detection to obtain 17 key points. Then, we average 5 points from face (*i.e.*, nose, ear, eye) as one point of 'face', given us the 13 human key joints and their corresponding coordinates. During training, the input images are resized to 384×192 and we only apply random erasing [42] for data augmentation. SGD is utilized as the optimizer to train networks with mini-batch 32, momentum 0.9, and the weight decay factor for L2 regularization is set to 0.0005. Our model is trained on one NVIDIA TITAN Xp GPU for total 120 epochs with an initial learning rate of 0.01 and a reduction factor of 0.1 every 40 epochs. During testing, we concatenate features f^+ from all CESD modules as the final feature.
Evaluation settings. We randomly split the LTCC dataset into training and testing sets. The training set consists of 77 identities, where 46 people have cloth changes and the rest of 31 people wear the same outfits during recording. Similarly, the testing set contains 45 people with changing clothes and 30 people wearing the same outfits. Unless otherwise specified, we use all training samples for training. For better analyzing the results of long-term cloth-changing Re-ID in detail, we introduce two test settings, as follows, (1) *Standard Setting*: Following the evaluation in [4], the images in the test set with the same identity and the same camera view are discarded when computing evaluation scores, *i.e.*, Rank-k and mAP. In other words, the test set contains both cloth-consistent and cloth-changing examples. (2) *Cloth-changing Setting*: Different from [4], the images with same identity, camera view and clothes are discarded during testing. This setting examines specifically how well a Re-ID model copes with cloth changes.

5.2 Comparative Results

LTCC dataset. We evaluate our proposed method on the LTCC dataset and compare it with several competitors, including hand-crafted features of LOMO [44] + KISSME [45], LOMO [44] + XQDA [44], LOMO [44] + NullSpace [46], deep learning baseline as ResNet-50 [40], PCB [47], and strong Re-ID methods MuDeep [49], OSNet [47] and HACNN [6] [48]. In addition, we try to leverage human parsing images, which are semantic maps containing body structure information [50]. Please find more quantitative and qualitative studies in Supplementary Material.

A number of observations can be made from results in Tab. 1. (1) Our method beats all competitors with a clear margin under both evaluation settings. As expected, the re-identification results under the cloth-changing setting are much

[6] OSNet is trained with the size of 384×192 and the cross-entropy loss as ours for a fair comparison. HACNN is trained with 160×64 as required by the official code.

Table 1. Results comparisons of our model and other competitors. 'Standard' and 'Cloth-changing' mean the standard setting and cloth-changing setting, respectively. '(Image)' or '(Parsing)' represents that the input data is person image or body parsing. '†' denotes that only identities with clothes changing are used for training. '§' indicates our simple implementation of the siamese network using face images detected by [43].

Methods	Standard		Cloth-changing		Standard†		Cloth-changing†	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
LOMO [44] + KISSME [45]	26.57	9.11	10.75	5.25	19.47	7.37	8.32	4.37
LOMO [44] + XQDA [44]	25.35	9.54	10.95	5.56	22.52	8.21	10.55	4.95
LOMO [44] + NullSpace [46]	34.83	11.92	16.45	6.29	27.59	9.43	13.37	5.34
ResNet-50 (Image) [40]	58.82	25.98	20.08	9.02	57.20	22.82	20.68	8.38
PCB (Image) [47]	65.11	30.60	23.52	10.03	59.22	26.61	21.93	8.81
HACNN [48]	60.24	26.71	21.59	9.25	57.12	23.48	20.81	8.27
MuDeep [49]	61.86	27.52	23.53	10.23	56.99	24.10	18.66	8.76
OSNet [47]	67.86	32.14	23.92	10.76	63.48	29.34	24.15	10.13
ResNet-50 (Parsing) [40]	19.87	6.64	7.51	3.75	18.86	6.16	6.28	3.46
PCB (Parsing) [47]	27.38	9.16	9.33	4.50	25.96	7.77	10.54	4.04
ResNet-50 + Face§ [27]	60.44	25.42	22.10	9.44	55.37	22.23	20.68	8.99
Ours	**71.39**	**34.31**	**26.15**	**12.40**	**66.73**	**31.29**	**25.15**	**11.67**

lower than the standard setting, which verify the difficulty of LTCC Re-ID. Under cloth-changing setting, where the training and testing identities both with different clothes, our method surpass the general baseline 'ResNet-50 (Image)' with 4.47%/3.29% of Rank-1/mAP, and it also outperforms strong baseline 'PCB (Image)' by 3.22%/2.86%. (2) Meanwhile, our model also achieves better results than those advanced Re-ID methods, where some of them are designed with multi-scale or attention mechanism to extract more discriminative features under the general Re-ID problem. Such results indicate that our proposed method can better eliminate negative impact of cloth changes and explore more soft-biometrics identity-relevant features. (3) Following the idea of [27], we build a siamese network applying the information of face for LTCC Re-ID. Intuitively, more information, *i.e.,* face, leads to better performance. However, face feature is sensitive to factors of illumination, occlusion and resolution, so it is not the optimal way for the cloth-changing Re-ID. (4) Comparing the performance of 'ResNet-50 (Parsing)' and 'ResNet-50 (Image)', as well as 'PCB (Parsing)' and 'PCB (Image)', we can find that the models trained on person images have superior accuracy than those using parsing. This suggests that the identity-sensitive structural information is hard to capture directly from images. As a result, applying it alone for re-identification is unhelpful.

BIWI dataset. Additional experimental results on BIWI [30] dataset are further presented and discussed. Please refer to Supplementary Material for more details about dataset introduction and implementation details. From the results shown in Table 2, we can make the following observations: (1) Our method achieves the highest performance on Rank-1 at both settings (Rank-1 of 49.51% and 39.80% compared to the closest competitor OSNet [51] which gives 47.14%

Table 2. Results of the multi-shot setting on BIWI dataset. '$*$' denotes the results are reported from original paper [26]. Note that, the same baselines (*i.e.*, ResNet-50 [40], PCB [47]) implemented by us yield better results than those reported in [26].

Methods	Still Setting		Walking Setting	
	Rank-1	Rank-5	Rank-1	Rank-5
ResNet-50 [40]	41.26	65.13	37.08	65.01
PCB [47]	46.36	72.37	37.89	**70.21**
SPT+ASE* [26]	21.31	66.10	18.66	63.88
HACNN [48]	42.71	64.89	37.32	64.27
MuDeep [49]	43.46	**73.80**	38.41	65.35
OSNet [51]	47.14	72.52	38.87	61.22
Ours	**49.51**	71.44	**39.80**	66.23

and 38.87% at Still and Walking setting respectively). (2) Our method beats the state-of-the-art competitor SPT+ASE [26], which is designed for cloth-changing re-ID specifically, by over 28% and 21% on Rank-1 under the Still and Walking setting, respectively. Note that we adopt the same training/test split and pretrain strategy as SPT+ASE [26]. (3) We also notice that the performance of Still setting are generally higher than the Walking setting. This confirms the conclusion drawn in the main paper as 'Still' subset involve fewer variations of pose and occlusions.

5.3 Ablation Study

Table 3. Comparing the contributions of each component in our method on LTCC dataset. 'w/ Cloth label' means the model is also trained with clothes labels. 'RN' and 'Attn' refer to the relation network and the design of self-attention in CESD module, respectively. 'w/ single CESD' indicates that we only insert one CESD module after *res4* block. Note that all models of variants are trained only with images of identities who have more than one outfit.

Methods	Standard Setting			Cloth-changing Setting		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
ResNet-50 [40]	57.20	71.19	22.82	20.68	31.84	8.38
ResNet-50 [40] w/ Cloth label	63.08	75.05	29.16	20.89	31.64	9.46
Ours w/o SE	65.92	76.29	29.84	22.10	31.86	10.28
Ours w/o RN	65.51	75.89	29.37	21.29	31.67	10.05
Ours w/o Attn	64.09	76.70	28.82	22.31	34.12	9.78
Ours w/ single CESD	64.21	**77.51**	29.46	23.73	34.30	10.19
Ours	**66.73**	77.48	**31.29**	**25.15**	**34.48**	**11.67**

Effectiveness of shape embedding module. Our shape embedding (SE) module is designed to exploit the shape information from human joints, so that it can facilitate the task of long-term cloth-changing Re-ID. Here, we compare

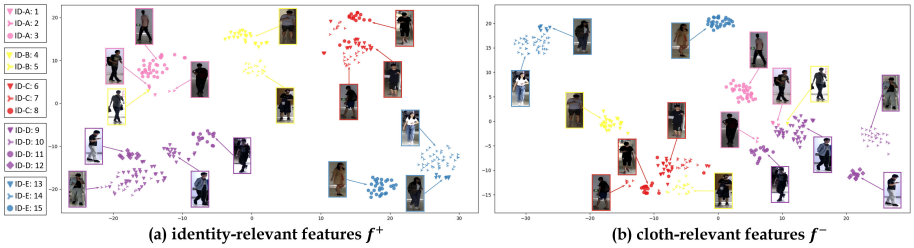

Fig. 5. Visualization of the identity-relevant features and cloth-relevant features learned from CESD using t-SNE [52]. In (a) and (b), each color represents one identity which is randomly selected from the testing set, and each symbol (circle, rhombus, triangle, cross, *etc.*) with various color indicates different clothes. In the legend, the capital letters behind the 'ID' indicate the identity labels, and the numbers are used to represent the clothing category. Best viewed in color and zoomed in.

several variants to evaluate its effectiveness. 'Ours w/o SE': a model without the SE module. 'Ours w/o RN': a model where the keypoint feature in SE module is encoded directly with several FC layers instead of the relation network.

We have three observations from the results in Tab. 3. (1) The variants of 'Ours w/o SE' and 'Ours w/o RN' achieve very similar results and both are clearly inferior to the full model. It clearly verifies that the shape information can be better encoded with the relations between two different points rather than the feature from an individual point. (2) With the information from human body shape, 'Ours' obtains a better performance than 'Ours w/o SE' (about 3% higher on Rank-1), which demonstrates the effectiveness of our shape embedding module. (3) Comparing the gap of our model with/without SE module between two different settings, we notice that the shape information contributes about 1% to the accuracy of Rank-1 under the standard setting, but 3 points under the cloth-changing setting, which shows that under the LTCC ReID task, the biological traits, *e.g.*, body shape, is more important and helpful.

Effectiveness of cloth-elimination shape-distillation module. Our proposed CESD module aims to distill the shape features and remove the cloth-relevant features from the input. We consider three variants in the analysis. Concretely, we first compare our full model with 'ResNet50 w/ Cloth label', whereby a vanilla ResNet-50 is trained with identity and clothes labels. The results under the two settings in Tab. 3 show that the clothes label helps the model a bit under the standard setting but not on the more challenging cloth-changing setting. Meanwhile, it is clear that it is our model rather than the clothes label that plays the main part. Secondly, we compare our full model with 'Ours w/o Attn', where a variant removes the self-attention block. It shows that 'Ours' gets about $2 \sim 3\%$ higher accuracy on Rank-1/mAP under both settings, which clearly suggests the benefit of the proposed CESD module on LTCC Re-ID by explicitly disentangling and eliminating the cloth-relevant features. Considering the features from the last two blocks of ResNet-50 are deep features, which are

more relevant to the specific task, we then evaluate the effect of different number of CESD modules. From Tab. 3, we can conclude that inserting one CESD module after *res3* and *res4* blocks individually is better than 'Ours w/ single CESD', which achieves around 2 points higher on Rank-1 and mAP. It further demonstrates the effectiveness of our CESD module.

Visualization of features learned from CESD. As described in Sec. 4, our CESD module can disentangle the input feature into two spaces, the identity-relevant and the cloth-relevant. To verify our motivation and better understand its objective, we visualize the learned features from the last CESD module using t-SNE [52]. Specifically, we randomly select images of five identities in the testing set, and each of them appeared in more than one outfits. We can observe that (1) with the overview of Fig. 5, the samples with the same identity are clustered using identity-relevant feature f^+, and the distances between images which have similar appearances are closer than those with large clothing discrepancy based on the cloth-relevant feature f^-. (2) As for Fig. 5 (b), images with warm color appearances are clustered at the top of feature space, and those with cold color are centered at the lower right corner. For example, the girl wearing a khaki dress (blue circle) has more similar appearance feature f^- to the boy with pink shorts (pink circle), and images at the bottom of the space are clustered due to wearing similar dark clothes. (3) In the identity-relevant feature space of Fig. 5 (a), identities denoted by blue, yellow and purple have smaller intra-class distances comparing with the distances in the cloth-relevant space in (b). (4) Interestingly, there is a special case denoted by a yellow inverted triangle in both spaces. His body shape is similar to the pink one, so in (a) he is aggregated closer to the pink cluster. Meanwhile, he, wearing black pants and dark shorts (affected by illumination), is surrounded by the images having dark clothes in (b). In conclusion, our model can successfully disentangle the identity-relevant features with the cloth-relevant features.

6 Conclusion

In this paper, to study the Re-ID problem under more realistic conditions, we focus on Long-Term Cloth-Changing (LTCC) Re-ID, and introduce a new large-scale LTCC Dataset, which has no cloth-consistency constraint. LTCC has 152 identities with 478 different clothes of 17,128 images from 12 cameras, and among them, there are 91 persons showing clothing changes. To further solve the problem of dramatic appearance changes, we propose a task-driven method, which can learn identity-sensitive and cloth-insensitive representations. We utilize the relation between the human keypoints to extract biological structural features and apply attention mechanism to disentangle the identity-relevant features from clothing-related information. The effectiveness of proposed method is validated through extensive experiments.

Acknowledgment. This work was supported in part by Science and Technology Commission of Shanghai Municipality Projects (19511120700, 19ZR1471800), NSFC Projects (U62076067, U1611461).

References

1. Wang, X., Tieu, K., Grimson, W.: Correspondence-free multi-camera activity analysis and scene modeling. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)
2. Chan, A., Vasconcelos, N.: Bayesian poisson regression for crowd counting. In: IEEE International Conference on Computer Vision. (2009)
3. Berclaz, J., Fleuret, F., Fua, P.: Multi-camera tracking and atypical motion detection with behavioral maps. In: European Conference on Computer Vision. (2008)
4. Zheng, L., Shen, L., Tian, L., S.Wang, J.Wang, Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision. (2015)
5. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE International Conference on Computer Vision. (2017)
6. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2014)
7. Bak, S., Carr, P., Lalonde, J.: Domain adaptation through synthesis for unsupervised person re-identification. In: European Conference on Computer Vision. (2018)
8. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: IEEE International Conference on Computer Vision. (2015)
9. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: IEEE International Conference on Computer Vision. (2019)
10. He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., Feng, J.: Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In: IEEE International Conference on Computer Vision. (2019)
11. Qian, X., Fu, Y., Wang, W., Xiang, T., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. European Conference on Computer Vision (2018)
12. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: IEEE Conference on Computer Vision and Pattern Recognition. (2019)
13. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Neural Information Processing Systems. (2017)
14. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)
15. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision Workshop. (2016)
16. Wang, G., Yang, Y., Cheng, J., Wang, J., Hou, Z.: Color-sensitive person re-identification. In: International Joint Conference on Artificial Intelligence. (2019)
17. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2019)
18. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015)

19. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: End-to-end deep kronecker-product matching for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)
20. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: European Conference on Computer Vision. (2012)
21. Liu, Y., Shi, P., Peng, B., Yan, H., Zhou, Y., Han, B., Zheng, Y., Lin, C., Jiang, J., Fan, Y., et al.: iqiyi-vid: A large dataset for multi-modal person identification. arXiv preprint arXiv:1811.07548 (2018)
22. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In: International Joint Conference on Neural Networks. (2019)
23. 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. IEEE Transactions on Circuits and Systems for Video Technology (2019)
24. Joon Oh, S., Benenson, R., Fritz, M., Schiele, B.: Person recognition in personal photo collections. In: IEEE International Conference on Computer Vision. (2015)
25. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: Improving person recognition using multiple cues. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015)
26. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
27. Xue, J., Meng, Z., Katipally, K., Wang, H., van Zon, K.: Clothing change aware person identification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2018)
28. Li, W., Zhao, R., X.Wang: Human re-identification with transferred metric learning. In: Asian Conference on Computer Vision. (2012)
29. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision. (2017)
30. Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L.: One-shot person re-identification with a consumer depth camera. In: Person Re-Identification. (2014)
31. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: AAAI Conference on Artificial Intelligence. (2019)
32. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
33. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: IEEE International Conference on Computer Vision. (2017)
34. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose Flow: Efficient online pose tracking. In: British Machine Vision Conference. (2018)
35. Perez, E., De Vries, H., Strub, F., Dumoulin, V., Courville, A.: Learning visual reasoning without strong priors. arXiv preprint arXiv:1707.03017 (2017)
36. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence. (2018)
37. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision. (2017)
38. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. arXiv preprint arXiv:1705.06830 (2017)

39. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
40. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015)
41. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009)
42. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
43. Tang, X., Du, D.K., He, Z., Liu, J.: Pyramidbox: A context-assisted single shot face detector. In: European Conference on Computer Vision. (2018)
44. Liao, S., Hu, Y., Zhu, X., Li., S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015)
45. Kittur, A., Chi, E.H., Suh., B.: Crowdsourcing user studies with mechanical turk. In: ACM Computer-Human Interaction (CHI) Conference on Human Factors in Computing Systems. (2008)
46. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identificatio. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016)
47. 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: European Conference on Computer Vision. (2018)
48. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)
49. Qian, X., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Leader-based multi-scale attention deep architecture for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
50. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: IEEE Conference on Computer Vision and Pattern Recognition. (2017)
51. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: IEEE International Conference on Computer Vision. (2019)
52. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)