Image and Vision Computing 115 (2021) 104288

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

You look so different! Haven't I seen you a long time ago?

Ehsan Yaghoubi [a,*], Diana Borza [b], Bruno Degardin [a,c], Hugo Proença [a,c]

[a] University of Beira Interior, Portugal
[b] Babes Boylai University, Cluj-Napoca, Romania
[c] IT: Instituto de Telecomunicações, Portugal

ARTICLE INFO

Article history:
Received 25 July 2021
Received in revised form 30 August 2021
Accepted 31 August 2021
Available online 06 September 2021

Keywords:
Cloth-changing person re-identification
Long-term embedding
Biometric recognition
Visual surveillance
Identity feature extraction
Soft biometric analysis

ABSTRACT

Person re-identification (re-id) aims to match a query identity (ID) to an element in a gallery set, composed of elements collected from multiple cameras. Most of the existing re-id methods assume the short-term setting, where the query/gallery samples share the clothing style. In this setting, the optimal feature representations are based in the visual appearance of clothes, which considerably drops the identification performance for long-term settings. Having this problem in mind, we propose a model that learns long-term representations of persons by ignoring any features previously learned by a short-term re-id model, which naturally makes it invariant to clothing styles. We start by synthesizing a data set in which we distort the most relevant biometric information (based in face, body shape, height, and weight cues), keeping the short-term cues (color and texture of clothes) unchanged. This way, while the original data contains both ID-related and other varying features, the synthesized representations are composed mostly of short-term attributes. Then, the key to obtaining stable long-term representations is to learn embeddings of the original data that maximize the dissimilarity with the previously inferred short-term embeddings. In practice, we use the synthetic data to learn a model that embeds the ID-unrelated features and then learn a second model from the original data, where long-term embeddings are obtained, keeping their independence with respect to the previously obtained ID-unrelated features. Our experiments were performed on three challenging cloth-changing sets (LTCC, PRCC, and NKUP) and the results support the effectiveness of the proposed method, for both short and long-term re-id settings. The source code is available at https://github.com/Ehsan-Yaghoubi/You-Look-So-Different-Haven-t-I-Seen-You-a-Long-Time-Ago?

1. Introduction

Retrieving a query identity from a gallery of people with consistent clothing-style across a distributed camera network is known as short-term person re-identification (re-id) [1]. Being a challenging task, short-term re-id has been the topic of substantial research for more than a decade, with several datasets announced [2,3], methods proposed [4,5], and multiple surveys published [5–8]. In this problem, the major challenges are the variations in body pose, varying illumination, occlusions, camera resolution, and viewing angle. Here, SOTA methods typically obtain feature representations based on the clothing texture/color which are the most discriminative/permanent cues in this kind of data. Short-term re-id methods are known to substantially degrade their performance under cloth-changing scenarios [9]. It is commonly accepted that re-identifying people from biological traits rather than from any transient appearance characteristics is more challenging [10], which provides the main motivation for this work: to develop a re-id model that is naturally invariant to clothing features such as colors, textures, shapes, and styles.

As illustrated in Fig. 1, in long-term person re-id settings, the model should recognize instances of the same person after several weeks or months, assuming that the query subject might be wearing different clothes than any instance of the gallery. Recently, some models were proposed to learn cloth-independent features, by either generating people with different clothing patterns [9,11] or extracting shape-based body features [12,13]. Other works assumed specific constraints (e.g., constant walking patterns [14], moderate clothing changes [13], and visible facial images [15]), attempting to learn ID-sensitive embeddings by changing the clothes colors/patterns. In opposition, [13,15] exclusively focused in the body-shape or facial attribute.

Learning robust features is a key factor in long-term person re-id. Here, *robustness* refers to 1) the extraction of discriminative features from inter-person samples and 2) the invariance to intra-person dynamics. Although the cross-entropy loss function optimizes the re-id model for these criteria, high variations in the intra-person samples

* Corresponding author.
E-mail address: Ehsan.Yaghoubi@ubi.pt (E. Yaghoubi).

https://doi.org/10.1016/j.imavis.2021.104288
0262-8856/

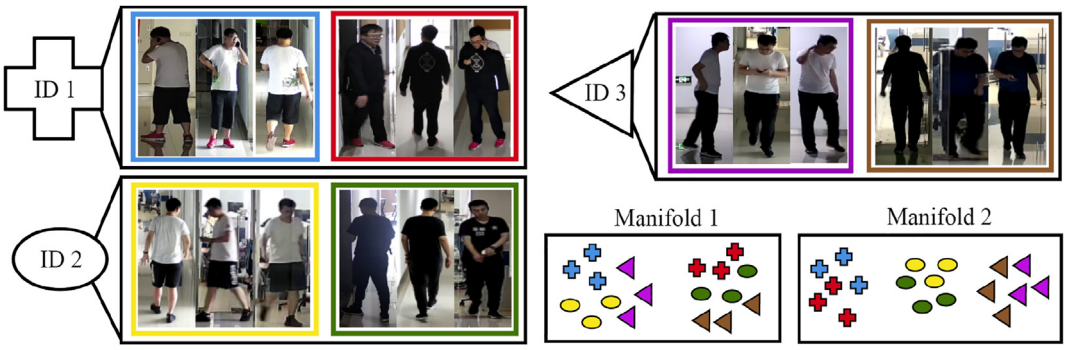

Fig. 1. Main motivation of the proposed work. Short-term person re-id methods rely on appearance features that typically converge into "Manifold 1", in which samples with similar clothes appear nearby. Instead, our goal is to obtain an embedding such as "Manifold 2", where samples of different persons appear together, regardless of their clothing styles. Each symbol points to a specific person ID, and different colors show different clothing styles. The samples are taken from the PRCC dataset (best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

hinder the model from learning useful long-term representations and lead embeddings similar to "manifold 1", illustrated in Fig. 1.

Based on our analysis, we concluded that the key to mitigate the above problems is to keep the useful visual appearance information (face, body shape, body figure, height, gender) while disregarding any other ID-unrelated features (clothing styles and background features). Accordingly, this paper proposes a framework that firstly transforms the original learning data in a way to help the model to infer ID-unrelated features (i.e., short-term). At a second step, a long-term embedding is learned by minimizing the correlation between the inferred features and the previously obtained short-term feature representations, according to a cosine similarity loss.

The main contributions of this paper are as follows:

- We propose an image transformation pipeline that induces image-based re-id models to disregard background and clothing-based features.
- We propose a framework that re-identifies people based on their face and soft biometrics cues (e.g., body shape), while automatically disregarding any changeable visual appearance features (e.g., clothes). Moreover, at the inference time, our solution does not depend on any kind of additional labeling information, such as body masks or key-points.

2. Related work

Most of the prior person re-id studies assume that the query persons wear the same outfits in the gallery set [8]. However, this assumption is not always valid and leads to poor performance when applied to long-term re-id settings. In this paper, we focus on a real-world scenario, when people may appear with different clothes. We refer the readers to [5,8] for discussions on the representative works of the short-term person re-id.

As an early study in the context of long-term re-id, Zhang *et al.* [14] proposes a video-based re-id technique based on the body motion to address the challenge of person appearance variations. In this work, the authors applied local descriptors (i.e., Histogram of Optical Flow and Motion Boundary Histogram) to capture the latent motion cues of a person's walking style and relative motion between feature points, based on the hypothesis that persons' movement follows a consistent pattern. Although this method captures some fine-grained gait features, it disregards the useful appearance features related to the body-shape and head area.

In [15], the authors focused solely on scenarios where the face is clearly visible. The proposed model processes two persons' pictures and uses the face area to yield the person ID and detects whether the subject has different clothes based on the body area or not. However, the high resolution face shots are rarely available in the surveillance data, which leads to an undesirable performance of the state-of-the-art face recognition models. So, coupling a short-term re-id model equipped to a face re-id branch cannot obtain satisfactory results [1, 13]. Later, [13] performed a case-study, in which the individuals change their clothes, such that the overall body shape is preserved. In other words, the authors proposed a re-id model based on the person's contour sketches to ignore the color-based features and demonstrate the importance of the body-shape in long-term person re-id.

In order to enhance the performance of the deep-based long-term person re-id, one strategy would be to increase the learning data, such that each subject wears numerous different clothes. As collecting such a dataset on a large scale demands expensive gathering and annotation processes, some studies proposed applying generative models. In this context, inspired by a pose-invariant generative re-id model [16], Yu *et al.* [9] proposed a clothing simulator model to synthesize more samples for each ID with several different clothing styles. The authors applied a body-parsing technique on the image to mask out the clothes area and trained a generative model to reconstruct the clothes area differently. Afterward, another model used both the original image and the reconstructed image to learn the differences (clothes area). Although this method has tried to decrease the clothing change effects, some challenges may rise: 1) segmentation clothing area is itself a challenging task in computer vision and yet cannot yield accurate results on the real-world human surveillance data, 2) this method neglects the feature similarities in the background area, 3) the shape of the clothing styles (e.g., short dress and long dress) needs to be considered as it highly affects the final feature representation of the persons.

In another generative-based study [12], the authors proposed an adversarial learning-based model to ignore the color features and focused solely on the body-shape features. To derive the body-shape representation, the authors extracted image features in RGB and grey-scale modes and fed them into a feature discriminator to distinguish between the RGB and grey-scale feature sets. Supposing that another image of the same person contains similar body-shape features, the authors concatenated the grey-scale features of a first body-pose with the RGB features of a second body-pose. Then, they trained a generator to reconstruct an RGB image with the first body-pose.

With an assumption that the body-shape is a reliable soft-biometric for long-term re-id scenarios, Qian *et al.* [1] used the human joint coordinates to model the relations among them by two scalar numbers in x-axis and y-axis directions. Next, these scalars were used to generate the shape-based features that their difference with the image-based features could result in a shape-sensitive feature representation of the input sample. [1] relies on capturing the information of the body-joints coordinates; however, [13] shows that the contour sketch of the body has useful information which cannot be inferred from the body key-points.

Most recently, [17], [18], and [19] presented frameworks based on convolutional neural networks to perform cloth-changing person re-id. In [17], the authors suggest a two-stream framework to learn fine-grained features of the body shape and transfer them to the other stream of the model to refine the appearance features such that the influence of clothing features are reduced in the final decision of the model. [18] proposed a framework for long term re-id that combines both appearance and gait cues. The training process relies on a two-stream architecture, comprising a re-id branch and an auxiliary gait recognition branch. The latter stream is used as a regularizer to ensure that the re-identification model learns unique, cloth-independent gait information representations from a single image. For efficiency, the gait branch is discarded during the inference phase. [19] designed a semantic-guided pixel sampling approach which ensures that the re-id model exploits cloth-irrelevant cues. [19] relies on a human parsing module, and changes the clothing of a subject by sampling pixels from other pedestrians. These generated samples are then used in the training process along with the original images. To exploit solely the cloth irrelevant cues, the authors use a mean squared error loss term to ensure that the learned features remain consistent before and after the change of clothes.

Based on the above-mentioned review on the recent studies, a long-term person re-id model may extract useful information from head-neck area, full-body soft biometrics, and body-shape characteristics and excludes identity-irrelevant characteristics such as clothing and background features. In the next section, we explain how our model captures these data and disregards the short-term features.

3. Proposed method

The proposed Long-term, Short-term features Decoupler (LSD) framework is an image-based person re-id network that extracts long-term discriminative representations of people that are invariant to clothes and background changes. The LSD framework is developed in four phases: 1) pre-processing, 2) learning short-term embeddings (ID-unrelated features), 3) learning the long-term embeddings (ID-related features), and 4) inferring the long-term feature representations of people. In the pre-processing phase, we generate a synthesized dataset, in which we apply several image transformations on each sample of the original learning set to distort the visual identity cues such as facial area, body figure, height, weight, and gender (see Fig. 2 and Fig. 3). Then, in the first learning phase, we train an auxiliary model, named as Short-Term Embedding Convolutional Neural Network (STE-CNN), on the *synthesized data* to extract the ID-unrelated embeddings of each instance. In the next learning phase, we use a *cosine similarity loss function* in the learning phase of a second model, called Long-Term Embedding CNN (LTE-CNN), to learn from the *original images* such that the learned embeddings are dissimilar to the ID-unrelated embeddings. This way, the LTE-CNN model captures the embeddings of the identity cues that are unchangeable during long time intervals and disregards the attributes that are more prone to change e.g., clothing style, accessories and background. In the evaluation phase, we only use the LTE-CNN model to infer the long-term representations of people. This denotes that *training the STE-CNN model and generating synthesized data are auxiliary steps that enhances the learning quality of the LTE-CNN model and are skipped in the inference phase*. Meanwhile, the evaluation process of the LTE-CNN model is similar to the typical re-id models: the gallery samples are ranked based on the similarity between the long-term representations of the gallery and query instances.

It is worth noting that the STE-CNN and LTE-CNN are regular deep architectures (e.g., resnet-50) that extract the global features of the input data, and the given names are to provide the reader with a feeling about their functionality; therefore, both the STE-CNN and LTE-CNN may have an identical architecture, but are different in terms of the input data and loss function.

3.1. Pre-processing: image transformation pipeline

In the proposed LSD model, the STE-CNN must learn the embeddings unrelated to the subject's ID, such as clothes and background features. This section describes the various image processing steps applied to the original learning set to remove the ID cues and generate the learning

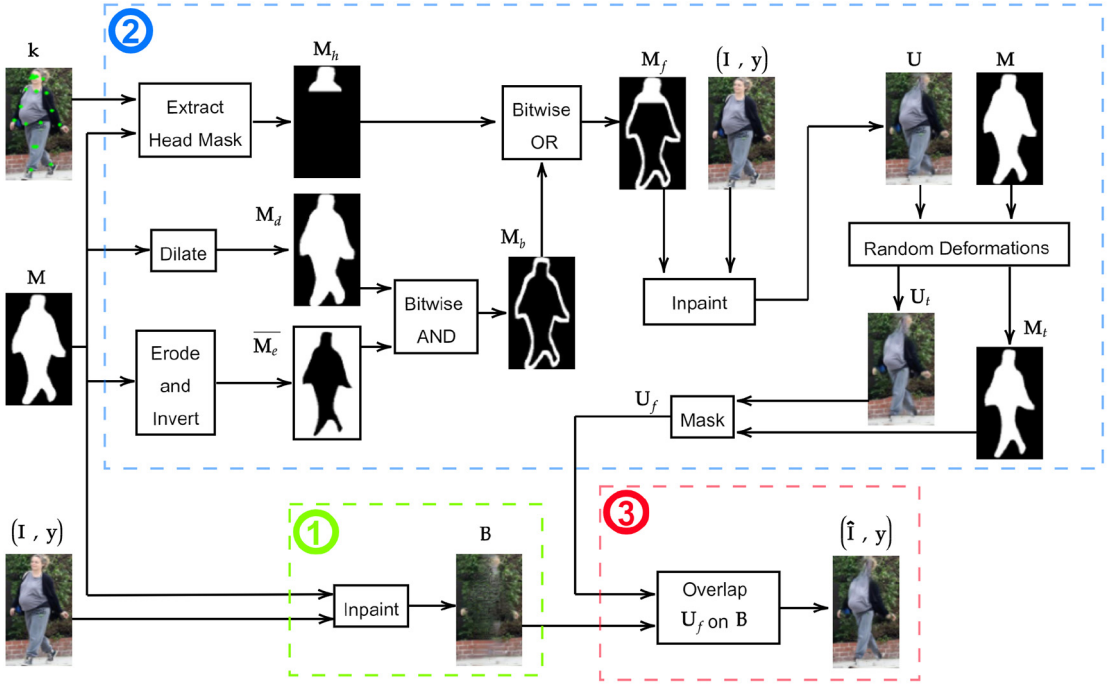

Fig. 2. Overview of the image transformation pipeline for removing the ID-related cues. ***k***, ***M***, ***I***, *y*, ***U***, and ***B*** are respectively the body keypoints, binary mask, RGB image, ID label, transformed image, and reconstructed background of the person. (1) shows the reconstruction of plain background ***B***, (2) illustrates the steps to generate the distorted foreground area $\boldsymbol{U_f}$, and (3) shows that ID-unrelated image $\hat{\boldsymbol{I}}$ is generated by overlapping $\boldsymbol{U_f}$ over ***B***. Best viewed in color.

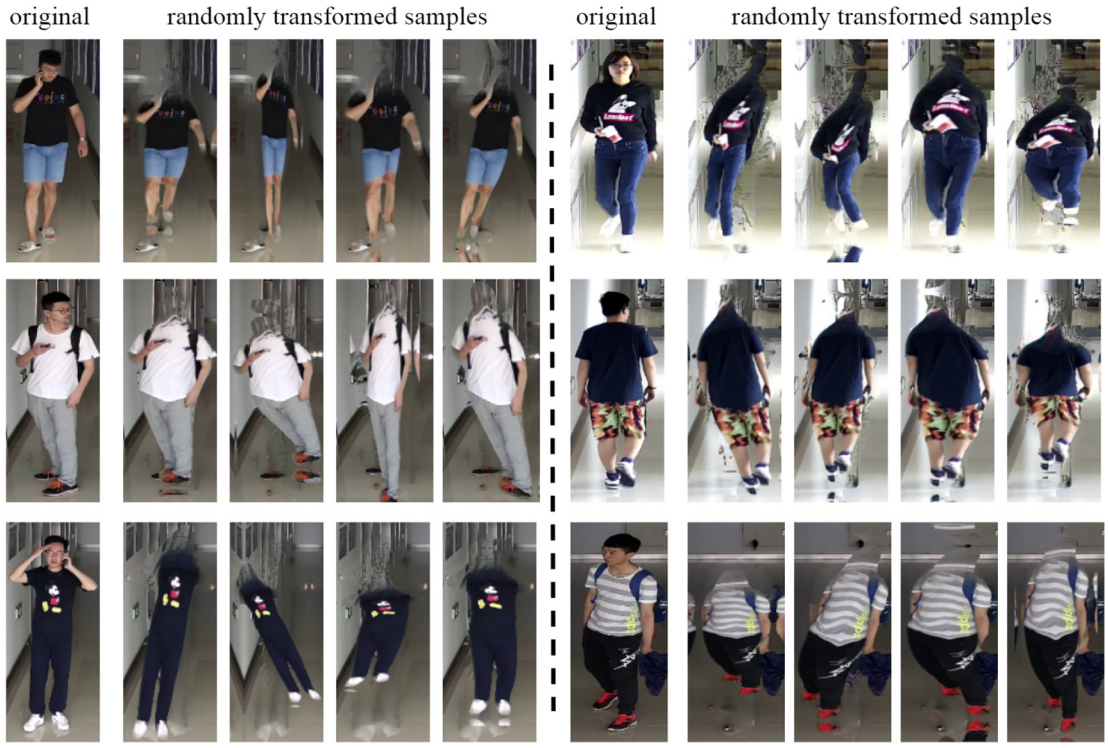

Fig. 3. Samples of the synthesized ID-unrelated data, juxtaposed with the original images. As shown, identity cues such as general body shape and face area have been distorted successfully, while the overall texture of the original samples have been remained unchanged. For example, considering the last subject in the left column, the generated samples cannot represent the height of the person.

data for the STE-CNN model. Fig. 2 gives an overview of the image transformation pipeline and Fig. 3 shows some examples of several synthesized samples. The results show that as we intended, the robust soft biometrics (such as weight, height, and body shape) have been visually distorted in the transformed images, while the background area and accessories have been unchanged approximately.

The proposed pipeline requires the input image, the segmentation mask, and the body key-points of the subject. The latter data are extracted using the state-of-the-art methods, for instance, segmentation [20] and human body key-point localization [21]. It is worth noting that our approach does not require a perfect segmentation and localization of the body parts, as these data are used to roughly establish an irregular shaped region of interest (body contour) to be removed from the input image.

We hypothesize that the head area and the overall body contour (shape) contain the most ID-related cues, while background, accessories, clothes texture, and clothes color result in temporary features. Therefore, we apply several transformations on each input image to (1) remove the subject ID from the scene and create a plain background, (2) generate the ID-unrelated foreground, for which we distort the ID-related cues of the person body and face, (3) overlap the ID-unrelated foreground on the plain background. Fig. 2 presents an overview of our strategy for generating ID-unrelated images. In the remainder of this section, we explain each of these steps in detail. For simplicity, we skip the index i and use $\boldsymbol{I}$ to denote the ith original input image, $\boldsymbol{M}$ to refer its corresponding, original body mask, and $\boldsymbol{K} = \{(\boldsymbol{k}_{x1}, \boldsymbol{k}_{y1}), (\boldsymbol{k}_{x2}, \boldsymbol{k}_{y2}), \ldots, (\boldsymbol{k}_{x17}, \boldsymbol{k}_{y17})\}$ to show the body key-points for this image.

1. To generate a plain background image $\boldsymbol{B}$, we consider the foreground area (subject body) using the mask $\boldsymbol{M}$ as the missing pixels and apply the in-painting method [22] to restore the background area (see boarder 1 in Fig. 2).
2. Next, we generate an ID-unrelated foreground area $\boldsymbol{U}_f$ that contains the short-term attributes (illustrated in boarder 2 in Fig. 2). To this end, (a) We use the body key-points $\boldsymbol{K}$ and the full-body mask $\boldsymbol{M}$ to select a head-neck mask $\boldsymbol{M}_h$ from the original mask $\boldsymbol{M}$. (b) In parallel, we should obtain a body contour mask $\boldsymbol{M}_b$, for which we use a method similar to the top-hat morphological transformation. The original body mask $\boldsymbol{M}$ is first expanded using a morphological dilation operator to obtain the mask $\boldsymbol{M}_d$: $\boldsymbol{M} \oplus B = \cup_{d \in B} \boldsymbol{M}_d$; (the size of the dilation kernel B is proportional to the size of the original mask. We used 3% of the width and height of the mask in our experiments). Then, we use an erosion morphological operator to shrink the body mask area: $\boldsymbol{M} \ominus B = \cap_{e \in B} \boldsymbol{M}_e$. Next, the body contour $\boldsymbol{M}_b$ is obtained by taking the intersection (bitwise AND operation) between the dilated mask $\boldsymbol{M}_d$ and the inversion (bitwise NOT operation) eroded body mask $\bar{\boldsymbol{M}}_e$. (c) A final mask $\boldsymbol{M}_f$ is obtained by adding (bitwise OR operation) the head-neck pixels with the body contour pixels: $\boldsymbol{M}_f = \boldsymbol{M}_b + \bar{\boldsymbol{M}}_e$. (d) ID-related pixels are then in-painted in the input image $\boldsymbol{I}$ using [22] to generate an image ($\boldsymbol{U}$) without any identity information. (e) It is important to deform the overall body shape of the person (by simulating random changes in weight, height, and clothes pattern). We apply this deformation to remove the remained ID-related features. However, to preserve the background area from deformation, we perform the same random transformations on the mask $\boldsymbol{M}$ and the in-painted image $\boldsymbol{U}$; so, in the next step, we could mask out the body area. We use [23] followed by a random stretching in height and width of the body area to apply some image deformations randomly. Precisely speaking, we impose a perturbation mesh on the mask $\boldsymbol{M}$ and image $\boldsymbol{U}$ to alter the subject's silhouette. Then, some points are selected on the mesh to distort the body shape by some random directions and strengths; this mesh deformation is applied by linear interpolation at a pixel-level on both $\boldsymbol{M}$ and $\boldsymbol{U}$. (f) Finally, the deformed foreground area $\boldsymbol{U}_f$ is obtained by masking out the image $\boldsymbol{U}_t$ with $\boldsymbol{M}_t$.
3. The last transformation step in the proposed pipeline overlaps the deformed foreground region $\boldsymbol{U}_f$ on the background $\boldsymbol{B}$, yielding ID-unrelated image $\hat{\boldsymbol{I}}$ (see boarder 3 in Fig. 2).

Fig. 3 shows some examples of the long-term cloth-changing (LTCC) data set [1] that have been transformed by our pre-processing pipeline due to the removal of their ID-related cues.

3.2. Proposed model: learning phase

Learning robust features is a key factor in long-term person re-id. In the context of this task, robustness refers to 1) the extraction of discriminative features from inter-person samples and 2) being invariant to intra-person attribute variations. Although the cross-entropy loss function optimizes these criteria, high variations in the intra-person samples and limited data hinder the model from learning useful long-term representations. *The key to enhance the quality and speed of the learning process of long-term representations of people is to focus on both distilling the identity-related features and disregarding the identity-unrelated features.*

Suppose that the learning set $G = \{(\boldsymbol{I}_i, y_i, c_j)\}$ consists of n persons with m different clothing styles for each person, where $y_.$ denotes the person-ID label, $c_.$ refers to the clothing label, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. By performing several image transformations on the learning set G, we synthesize another learning data set $\hat{G} = \{\hat{\boldsymbol{I}}_{\boldsymbol{i}}, y_i, c_j\}$ that excludes the ID-related visual features. This phase was described in the previous subsection.

As shown in the first learning phase in Fig. 4 (b), we feed the synthesized data $(\hat{\boldsymbol{I}}_{\boldsymbol{i}}, y_i, c_j)$ to the STE-CNN model $\hat{\varphi}(\hat{G};\hat{\theta})$ and learn labels y_i, c_j with a cross-entropy loss function. The label y_i, c_j refers to the person i with the ID label y_i with the clothing label c_j; in other words, this network learns to distinguish between the outfits worn by person i. The extracted features of this person are denoted as short-term features $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$ and are frozen during the next learning phase, where we feed the original image of person i to a second model. Precisely, given the original data $(\boldsymbol{I}_{\boldsymbol{i}}, y_i, c_j)$ and frozen short-term features $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$, the LTE-CNN model $\varphi(G, \theta)$ learns the long-term representations $\boldsymbol{f}_{\boldsymbol{i}}$, such that it is *mathematically dissimilar* to the ID-unrelated feature vector $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$, while simultaneously learns the ID-related features, using an aggregation loss function:

$$L_{\mathrm{LTE}} = \sum_{i=1}^{n} \frac{\boldsymbol{f}_i \hat{\boldsymbol{f}}_{\mathrm{ij}}}{\|\boldsymbol{f}_i\| \|\hat{\boldsymbol{f}}_{\mathrm{ij}}\|} + \sum_{i=1}^{n} t_i \ \log(s_i), \tag{1}$$

where n is the number of person IDs in the learning set, t_i is the ground-truth person ID (label), and s_i denotes the predicted probability score of person i. In equation 1, the cosine-similarity term minimizes the similarity between the short-term and long-term features, while the cross-entropy term helps the LTE-CNN learn the person ID.

Finally, in the inference phase, we only use LTE-CNN model $\varphi(G, \theta)$ to extract the long-term representations of the query and gallery data. Next, similar to the short-term person re-id methods, the gallery set is ordered based on the euclidean distances between the query and gallery samples. Then, the Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP) metrics are reported as the evaluation criteria.

4. Experiments and discussion

4.1. Datasets

The Long-Term cloth-changing (LTCC) Person Re-identification dataset [1] was collected using a CCTV system with 12 cameras installed on different floors in an office building. It comprises 24 hours of video recording that were collected over two months. As a result, persons were appeared with substantial changes in lighting, viewing angle, and body pose. The authors used the Mask-RCNN framework [20] to extract the person bounding boxes from video frames and then annotated each bounding box with a person ID and clothing label. The LTCC dataset comprises 17,138 images from 152 identities with 478 outfits, and on average, each person appears with five different clothing outfits. The LTCC dataset is publicly available in two subsets: 1) training subset

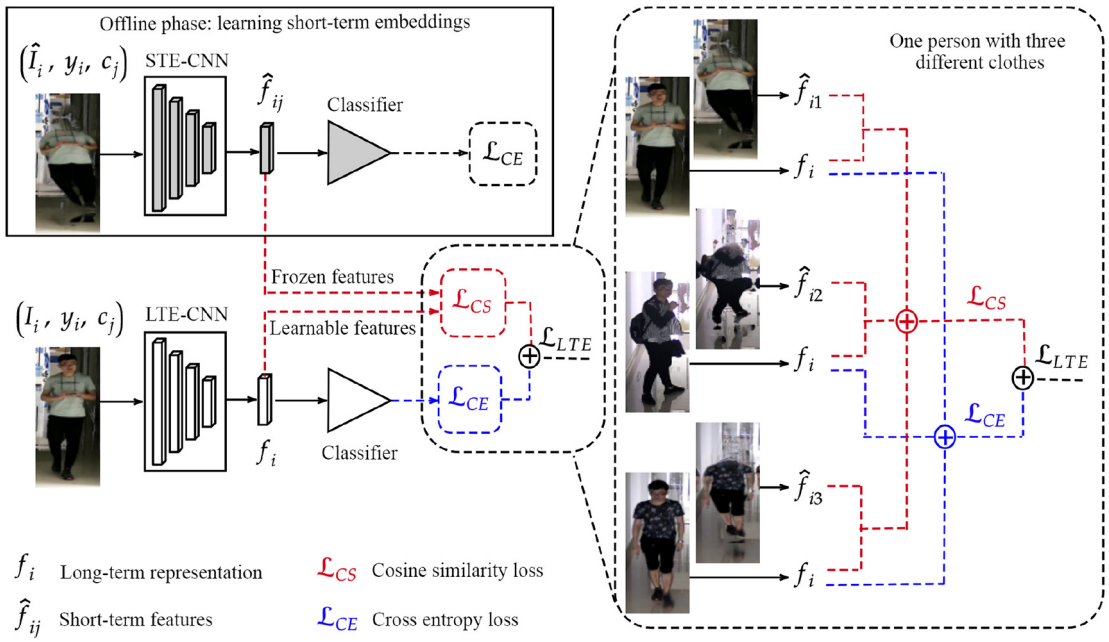

Fig. 4. Overview of the learning phase of the proposed model. In the offline learning phase, the STE-CNN model receives a transformed image $\hat{\boldsymbol{I}}_{\boldsymbol{i}}$ and extracts its short-term embeddings (ID-unrelated) $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$. Then, the long-term representation (ID-related) of the original image $\boldsymbol{I}_{\boldsymbol{i}}$ is obtained by minimizing the similarity between the long-term feature vector $\boldsymbol{f}_{\boldsymbol{i}}$ and the frozen short-term embeddings $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$. The magnified boarder shows the images of one person with three different clothes and indicates that how LTE-CNN loss function helps to learn the identity of the person (blue traces) and disregard clothing features (red traces). $\boldsymbol{I}_{\boldsymbol{i}}$ refers to the original image of person i with clothing style j, and $\hat{\boldsymbol{I}}_{\boldsymbol{i}}$ is the ID-unrelated version of $\boldsymbol{I}_{\boldsymbol{i}}$. Best viewed in color.

with 77 individuals, where 46 subjects are wearing different clothes and 31 elements appear with identical garments. 2) testing subset with 76 persons, where 46 people appear with different outfits and 30 individuals are wearing the same clothes.

The Person Re-identification by Contour Sketch (PRCC) dataset [13] was captured indoors using three cameras positioned in separate rooms. The PRCC dataset consists of 221 identities and a total of 33,698 images. In two camera views, the subjects wear the same clothes, while on the other camera, the garments change. Therefore, there are precisely two different clothes-changes per subject.

The NanKai University Pedestrian (NKUP) [24] dataset contains 9738 bounding box images of 107 pedestrians, out of which 5336 images of 40 persons were suggested as the training set, while 332 and 4070 bounding boxes of 67 identities were used as the query and gallery images, respectively. The NKUP dataset was collected from 15 cameras, 8 out of which were installed in the outdoor environment. Pedestrians in the NKUP dataset appear mostly with 2 or 3 different clothing styles. The query set includes 3 to 10 images of each person, randomly selected from certain clothing styles; then, the remaining images of the same person (with different clothing styles) were considered as the gallery images. Finally, images of individuals with one clothing style were considered distractors and added to the gallery samples.

We trained and evaluated our model on the LTCC [1], PRCC [13], and NKUP [24] long-term re-id datasets, as all of them comprise real-world data recorded with cameras and are large enough to be suitable for deep architectures. These datasets are publicly available in train and test splits, and there is no overlap between the subjects in the test and train sets. We followed the same evaluation settings in the original papers [1,13,24] to have a fair comparison. Fig. 5 shows a few samples of each dataset. As observed and indicated in [24], the resolution of the images in the LTCC dataset are much higher than the NKUP dataset [24]. Meanwhile, covered faces in the NKUP dataset makes this benchmark much harder to achieve high accuracy for long-term person re-id task.

4.2. Implementation details

We processed the original image $\boldsymbol{I}$ using the off-the-shelf Mask R-CNN [20] and Alpha-Pose [21] models with default configurations[1] and prepared the inputs of the pre-processing pipeline i.e., $\boldsymbol{K}$ and $\boldsymbol{M}$. The dilation and erosion transformations were performed using a kernel (filter), with a size that is proportional to 3% of the image width. The inpainting technique [22] was also used in its default configurations[2] using the pre-trained weights on the Places2 dataset [35].

The proposed framework, including the STE-CNN and LTE-CNN, can be implemented using any CNN architecture as feature extractors. In this paper, we implemented the proposed model based on residual CNNs using the Pytorch library to evaluate the effectiveness of our method. We started the training phases by fine-tuning the ImageNet pre-trained weights, using the Adam optimizer [36], for 250 epochs. The input images were 256×128 for both networks, i.e., STE-CNN and LTE-CNN. For more implementation details, we refer the readers to the project page[3].

4.3. Results

4.3.1. LTCC dataset

To evaluate our model on the LTCC dataset [1], we considered the two settings suggested in the original paper [1]: 1) standard setting, in which we ignore those images of the gallery that have captured from the same person and same camera. 2) cloth-changing setting, where

Fig. 5. Randomly selected samples of three cloth-changing person re-id benchmarks. The LTCC and PRCC datasets were captured in indoor locations, while the NKUP dataset includes variety of samples captured at outdoor/indoor locations at daylight/night from different distances and also includes grey scale samples, making it a much harder benchmark.

[1] https://github.com/matterport/Mask_RCNN, https://github.com/MVIG-SJTU/AlphaPose

[2] https://github.com/Atlas200dk/sample-imageinpainting-HiFill

[3] https://github.com/Ehsan-Yaghoubi/You-Look-So-Different-Haven-t-I-Seen-You-a-Long-Time-Ago

Table 1
Results on the LTCC data set. The method performance on the head patches is denoted by the * symbol. In standard setting, those images of the gallery that have captured from the same person and same camera are disregarded, while in cloth-changing setting, the images of the same person with identical clothes captured by the same camera are discarded from the gallery.

Methods	Standard setting					Cloth-changing setting				
	R-1	R-5	R-10	R-50	mAP	R-1	R-5	R-10	R-50	mAP
LOMO [25] + KISSME [26]	26.6	-	-	-	9.1	10.8	-	-	-	5.3
LOMO [25] + NullSpace [27]	34.8	-	-	-	11.9	16.5	-	-	-	6.3
resnet-50 [28]*	9.4	23.2	31.3	59.8	5.9	22.9	43.0	53.9	**77.7**	9.8
Luo *et al.* [29]*	25.8	47.5	57.2	80.6	10.2	11.7	23.8	33.4	62.9	5.9
resnet-50 [28]	49.7	64.9	70.4	86.6	19.7	18.1	32.4	38.8	59.2	8.1
se-resnext [30]	48.3	64.1	71.4	85.4	19.0	20.4	34.2	44.1	63.8	9.3
senet [30]	54.6	70.0	77.9	87.2	21.2	24.2	36.6	45.2	62.0	9.4
resnet50-ibn-a [31]	55.4	69.2	74.4	86.2	23.3	23.7	35.7	42.1	64.0	10.4
HACNN [32]	60.2	-	-	-	26.8	21.9	-	-	-	9.3
MuDeep [33]	61.9	-	-	-	27.5	23.5	-	-	-	10.2
Luo *et al.* [29]	60.2	74.0	80.1	88.8	25.6	24.2	40.6	51.5	71.2	11.3
Qian *et al.* [1]	71.4	-	-	-	34.3	26.2	-	-	-	12.4
Jin *et al.* [18]	73.6	-	-	-	36.1	28.1	-	-	-	13.2
Hong *et al.* [17]	73.2	-	-	-	35.4	38.5	-	-	-	16.2
Ours (LSD)	72.2	80.3	84.6	**91.9**	31.0	31.4	46.7	54.3	73.5	13.6
Ours + re-ranking [34]	**76.7**	**83.6**	**85.2**	**91.9**	**44.9**	**41.1**	**53.6**	**57.7**	74.0	**19.5**

The best and second-best results are in Bold and Underline styles, respectively.

the images of the same person with identical clothes captured with the same camera are discarded from the gallery before ranking the gallery elements based on the query person.

We provide a comparison between our model performance to several baselines in Table 1, based on the LTCC dataset. In general, our model shows superior performance for both evaluation metrics: mAP and CMC for ranks 1 to 50.

As shown in the middle column of Table 1, in standard evaluation setting, the hand-crafted based methods can extract better feature representations (from full-body images of persons) in comparison with simple baselines [28,29], when simple baselines are learned based on the face/head patches. In the next performance level, resnet50-ibn-a [31] achieves 55.4% and 23.3% of rank-1 and mAP, respectively; these numbers improve by the short-term re-id baselines, specifically to 61.9% and 27.5% by [33]. As a long-term re-id framework, Qian *et al.* [1] presents competitive results (71.4%/34.3% of rank-1/mAP) compared to our method without re-ranking (72.2%/31.0% of rank-1/mAP). [18, 17] methods are two most recent long-term person re-id methods that proceed the LSD model with 73.6%/36.1% and 73.2%/35.4% of rank-1/mAP, respectively. However, by applying the re-ranking technique [34] on our results, our method consistently outperforms the other competitors ([1,18,17]) and achieves 76.7%/44.9% of rank-1/mAP.

The results of the cloth-changing evaluation setting presented in Table 1 indicate that the performance of the short-term re-id methods [32,33,29] roughly degrades to their one-third. This performance drop denotes that [32,33,29] methods heavily rely on the color and texture of the clothes to re-identify people. It is also interesting that a resnet-50 model could extract more useful long-term information from head-shots (22.9%/9.8% of rank-1/mAP) rather than full-body images (18.1%/8.1% of rank-1/mAP), whereas the short-term model [29] fails in the head-shot long-term re-id setting, by achieving 24.2%/11.3% of rank-1/mAP from the full-body images and obtaining 11.7%/5.9% of rank-1/mAP from the head patches. In the cloth-changing context, our method obtains 31.4%/13.6% of rank-1/mAP, while these figures are dropped 26.2%/12.4% and 28.1%/13.2% for [1] and [18] methods, respectively. After the re-ranking process, our framework achieves 41.1% of rank-1 and 19.5% of mAP, which overtakes [17] by 1.6%/3.3% of rank-1/mAP.

Fig. 6 shows t-SNE [42] visualization of long-term representations provided with our proposed method for several persons from the LTCC test set that are wearing various clothing outfits. The representations related to consecutive frames of the same person with the same clothes are not close to each other, indicating that our method does not rely on the appearance similarity to re-identify people.

4.3.2. PRCC dataset

As previously mentioned, the PRCC dataset was collected using three cameras, namely *A*, *B*, and *C*, such that the individuals' clothes in cameras *A* and *B* are the same, while in the camera *C*, subjects wear different outfits. Following the evaluation protocol in [13], we select one image of each person from camera *A* and build a one-shot gallery, while samples captured by the other two cameras are considered to be as the queries for two different settings: evaluation on the cloth-changing and cloth-consistent settings.

Table 2 shows the performance of several baselines versus our method on the PRCC dataset. The baselines could be roughly divided to four groups: 1) methods based on the hand-crafted features [25,37, 38,26], 2) plain deep residual networks [28,30,31], 3) short-term person re-id techniques [32,39–41], and 4) long-term re-id method [13,18,17, 19].

In general, Table 2 indicates that methods based on the hand-crafted features obtain the lowest recognition results, with the rank-1 accuracy less than 24% and 55% in the cloth-changing and standard settings, respectively, whereas the second group of methods could achieve a rank-1/mAP approximately between 24%/35% to 33%/44% in the cloth-changing scenario and between 70%/78% to 85%/90% in the standard setting. Interestingly, the short-term re-id techniques could improve the rank-1 results up to 86.9%, when the inquiry person wears consistent clothing outfits in the gallery. When the query person appears with different clothing styles, our method outperforms [13] by improving the rank-1/mAP from 34.4%/- to 37.2%/47.6%. The performance of our method is further improved to 42.7%/52.2% of rank-1/mAP after the re-ranking process, presenting the second-best results after [17] with a rank-1/mAP of 54.5%/-. Moreover, when people wear identical clothes in the query and gallery sets, our method still achieves competitive results in comparison with the baselines such that the LSD framework achieves 93.6%/95.8% and 97.9%/98.7% of rank-1/mAP, respectively before and after re-ranking, while these numbers are 64.2%/- for [13] and 86.0%/- for [18].

[19] is a recent work that relies on an offline image generator model that *changes* the clothing of persons via simple pixel sampling, and during the training process, these generated images are used to help the

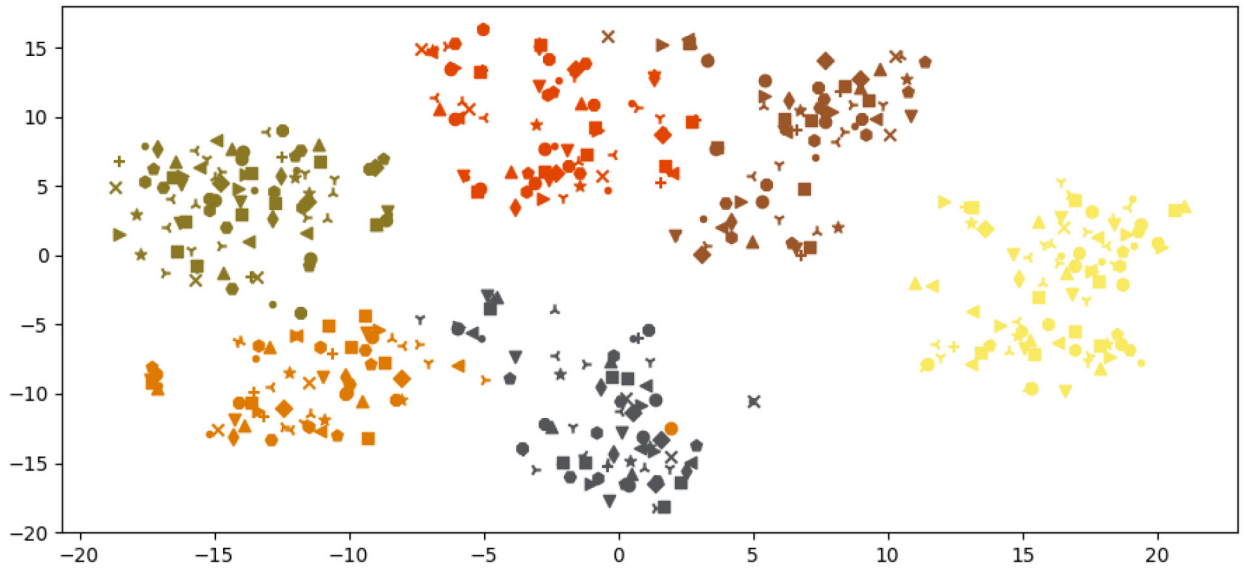

Fig. 6. Visualization of the long-term representations, according to t-SNE [42], for six IDs with varying clothes (LTCC test set). The data related to each person are presented in a different color, and variety in outfits is denoted by different markers. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

network learn identity based features rather than clothing based features. [19] uses a different evaluation setting, resulting to a high performance of an existing state of the art (PCB [39]), with 41.8%/38.7% of rank-1/mAP, whereas [13,18,17] are reported 22.9%/61.2% of rank-1/mAP for PCB [39]. Based on these differences of scale, we excluded [19] from the comparison Table 2.

4.3.3. NKUP dataset

Table 3 shows the experiments performed on the NKUP dataset. The holistic residual CNN models achieve competitive results for both standard and cloth-changing settings, respectively, with rank-1/mAP between 16.7%/11.7% to 14.6%/11.3%. As a short-term re-id framework, MGN [43] can achieve the second-best results (18.8%/15.0% of rank-1/mAP) for standard setting, while when we equip our method with the re-ranking technique [34], we achieve the highest performance for standard and cloth-changing settings with rank-1/mAP of 19.7%/15.6% and 16.4%/12.3%, respectively.

4.3.4. Discussion

As indicated in Tables 1 and 2 the proposed method has outperform several existing long-term re-id methods, while it can provide reliable results for short-term re-id task. Our interpretation of the superior performance of our method in both tasks is that, holistic CNNs can provide discriminative representation based on the identity (rather than clothes and background) when we use an aggregation loss function, in which we learn the ID labels using a cross-entropy loss term and penalize the learning of the ID-unrelated features by a similarity loss term. In fact, learning the identity cues by an aggregation loss function *implicitly* prevents the model from predicting the identity of people based on their clothes and background. Whereas, architectural based design may *explicitly* limit the model, which results into better long-term re-id but may degrade the short-term re-id accuracy.

As observed in Table 3, the results on the NKUP dataset have fallen behind the results obtained on the LTCC and PRCC datasets. We identify three factors that cause this degrading performance: 1) in our

Table 2
Results for two settings of the PRCC data set: 1) when the query person appears with different clothes in the gallery set (at left-side), 2) when the query's outfit is not changed in the gallery set (at left-side). The locally performed evaluations were repeated 10 times, and the variances from the mean values were shown by ± .

Methods	Cameras A and C (different clothes)				Cameras A and B (same clothes)			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
[25] + [26]	18.6	49.8	67.3	-	47.4	81.4	90.4	-
[37] + [38] + [25]	23.7	62.0	74.5	-	54.2	84.1	91.2	-
resnet-50 [28]	24.1±10.8	56.9±2.4	68.5±3.3	35.3±6.6	76.3±5.0	94.0±1.6	97.4±0.6	82.6±3.8
se-resnext101 [30]	27.7±1.7	57.6±6.6	70.3±3.5	37.8±2.1	69.1±8.9	94.4±4.0	97.6±2.2	78.4±5.1
senet [30]	27.2±4.7	54.5±4.9	66.9±1.7	36.6±3.2	76.7±4.6	96.0±1.7	97.9±0.6	83.9±2.6
resnet-ibn-a [31]	32.9±6.7	67.2±4.7	81.6±3.7	44.1±5.0	84.8±3.6	98.3±1.5	99.5±0.4	89.8±2.0
HACNN [32]	21.8	59.5	67.5	-	82.5	98.1	99.0	-
PCB [39]	22.9	61.2	78.3	-	86.9	98.8	99.6	-
DCN [40]	26.0	71.7	85.3	-	61.9	92.1	97.7	-
STN [41]	27.5	69.5	83.2	-	59.2	91.4	96.1	-
Yang *et al.* [13]	34.4	77.3	88.1	-	64.2	92.6	96.7	-
Jin *et al.* [18]	37.6	82.3	**93.7**	-	86.0	98.8	99.7	-
Hong *et al.* [17]	**54.5**	**86.4**	-	-	**98.8**	**100.0**	-	-
Ours (LSD)	37.2±6.7	68.7±2.0	80.5±4.1	47.6±3.4	93.6±1.7	99.5±0.6	99.8±0.1	95.8±1.1
Ours + re-ranking	42.7±4.2	71.2±3.5	81.5±2.4	**52.2**±2.2	97.9±0.4	99.8±0.0	**99.9**±0.0	**98.7**±0.1

The best and second-best results are in Bold and Underline styles, respectively.

Table 3
Results on the face-covered NKUP dataset. In the standard-setting evaluation, samples of the query person with the same camera are disregarded from the gallery, and in the cloth-changing setting, the photos of the query person with identical clothes and the same camera are discarded from the gallery. The results of the methods marked with * are taken from [24].

Methods	Standard setting					Cloth-changing setting				
	R-1	R-5	R-10	R-50	mAP	R-1	R-5	R-10	R-50	mAP
resnet-50 [28]	17.0	29.7	36.7	55.8	11.3	15.0	29.3	35.4	57.9	10.5
se-resnext [30]	16.7	25.5	31.2	46.4	11.7	15.7	24.6	29.6	45.4	11.3
senet [30]	18.2	25.2	30.6	56.7	11.1	15.7	23.2	29.3	58.2	9.3
resnet50-ibn-a [31]	17.0	**31.5**	**37.6**	**61.5**	10.7	14.6	**30.7**	37.1	**62.5**	10.3
PCB [39]*	16.9	25.6	30.1	-	12.4	-	-	-	-	-
MGN [43]*	18.8	28.8	33.0	-	15.0	-	-	-	-	-
Ours (LSD)	16.4	27.9	34.8	53.3	10.2	13.9	23.9	31.8	50.4	7.8
Ours + re-ranking [34]	**19.7**	24.8	29.1	45.2	**15.6**	**16.4**	21.8	26.1	42.1	**12.3**

The best and second-best results are in Bold and Underline styles, respectively.

framework, we consider the face as a reliable source of identity (a long term feature), but the faces of the individuals are masked in the NKUP dataset, 2) the resolution of the samples is extremely low, with 20% of the data having less than 800 number of pixels (see Fig. 5) and therefore the quality of the identity-based cues (e.g., body curves and hair-style) is reduced, 3) the NKUP dataset includes 8 outdoor cameras, causing more variety in illumination, body pose, and subject distance from the camera. These factors directly degrade the performance of our pre-processing step such that the holistic body key-points detector, i.e., Poseflow, cannot detect the body key-points of about 10% of samples in the NKUP dataset, making it impossible for the pre-processing module to extract the short-term features. Our ablation studies presented in Table 4 confirm this statement, indicating an exponential performance drop from 35.2%/13.7% of rank-1/mAP for input resolution of 256×128 to 8.4%/4.6% of rank-1/mAP for input resolution of 32×16, when trained and evaluated on the LTCC dataset. Moreover, when the resolution of data is highly degraded, from a certain level, the identity-based characteristics are diminished such that even human operators fail to re-identify the correct person.

5. Ablation studies

We performed several experiments with different backbones and input image sizes to evaluate the performance of the proposed LSD model in various conditions and find the limits of our method. The experiments in this section were carried out on the LTCC dataset, and the LSD model was trained for 50 epochs, and results were reported after the re-ranking process. The other settings remained as same as the previous experiments.

Left section of Table 4 shows the experiment results of the LSD for five different image resolutions from 32×16 to 512×256 and indicates that the improvement of the rank-1 accuracy saturates when the size of the images is increased from 256×128 to 512×256. In contrast, the mAP increases sharply in cloth-changing settings, from 13.7% to 17.4%. The reason behind the variation of accuracy is that, when we reduce the size of the images, some critical information (details probably) are lost permanently, whereas when we resize the images to 512×256, no extra detail are induced from data, probably because of the limits imposed by the image-quality of the captured data from far distances by the surveillance cameras. Furthermore, we trained and evaluated our model with several different feature extraction backbones. As shown in the right section of Table 4, the se-resnet models achieve better results than plain resnet methods. The proposed framework achieves better results when implemented based on the resnet50-ibn-a, with 57.8%/30.0% and 23.7%/11.5% of rank-1/mAP for the standard and cloth-changing settings, respectively. Moreover, these numbers improve to 58.6%/29.1% and 27.8%/11.7%, when the senet154 model is used as the backbone feature extractor.

Table 4
The performance of the proposed LSD model with different residual backbones and input resolutions, when trained for 50 epochs on the LTCC data set. When architecture is changing, the input resolution is fixed to 256×128, and when input resolution is changing, the senet154 architecture is used. SS and CCS stand for Standard Setting and Cloth-Changing Setting, respectively.

Architecture	SS		CCS		Input resolution	SS		CCS	
	R-1	mAP	R-1	mAP		R-1	mAP	R-1	mAP
resnet50	52.3	26.0	20.4	10.0	32×16	21.5	9.8	8.4	4.6
resnet101	47.9	24.9	17.1	10.0	64×32	43.2	23.1	15.3	8.8
resnet152	51.7	26.2	18.9	10.1	128×64	62.5	35.6	24.7	12.7
se-resnet101	56.2	29.6	22.4	11.6	256×128	70.0	39.5	35.2	13.7
se-resnet152	55.0	28.7	21.4	10.2	512×256	69.8	41.4	35.7	17.4
se-resnext101	55.8	27.9	23.0	11.4					
resnet50-ibn-a	57.8	30.0	23.7	11.5					
senet154	58.6	29.1	27.8	11.7					

The best and second-best results are in Bold and Underline styles, respectively.

6. Conclusions

Long-term person re-id aims at retrieving a query ID from a gallery, where elements are expected to appear with different clothing, hair-styles, or other accessories. This is an extremely ambitious setting, where the majority of the existing re-id methods still have poor performance. Hence, our primary motivation in this paper was to find alternate feature representations that are naturally insensitive to short-term re-id features. In this context, manually annotating large amounts of long-term instances for feeding supervised classification frameworks would be an insurmountable task, not only due to the lack of available data but also to the amount of human resources required for the task. Based on these observations, we proposed a Long-term/short-term feature decoupler model, where the most innovative point is to naturally learn long-term representations while ignoring the typically varying short-term cues (clothing style, body shape, and background). To this end, we proposed an image transformation pipeline over the ID-related regions (the head and the body shape) and created a model (STE-CNN) that identifies the most relevant short-term features. These representations were then separated from the long-term representation via the cosine similarity loss function. The experimental results on the state-of-the-art cloth-changing benchmarks confirmed the effectiveness of the proposed method by consistently advancing the performance with respect to the existing techniques.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT. This research was also supported by 'FCT - Fundação para a Ciência e Tecnologia' through the research grant 'UI/BD/150765/2020'.

References

[1] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Long-term cloth-changing person re-identification, Proc. ACCV, 2020 , pp. 1-1.
[2] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, Proc. IEEE ICCV, Springer 2016, pp. 17–35.
[3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, Proc. IEEE ICCV 2015, pp. 1116–1124.
[4] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimedia 22 (10) (2020) 2597–2609 10.1109/TMM.2019.2958756.
[5] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE TPAMI (2021) https://doi.org/10.1109/TPAMI.2021.3054775 1-1.
[6] B. Lavi, I. Ullah, M. Fatan, A. Rocha, Survey on Reliable Deep Learning-Based Person Re-Identification Models: Are We There Yet? 2020 arXiv preprint arXiv:2005.00355.
[7] M.O. Almasawa, L.A. Elrefaei, K. Moria, A survey on deep learning-based person re-identification systems, IEEE Access 7 (2019) 175228–175247.
[8] E. Yaghoubi, A. Kumar, H. Proença, Sss-pr: a short survey of surveys in person re-identification, Pattern Recognit. Lett. 143 (2021) 50–57.
[9] Z. Yu, Y. Zhao, B. Hong, Z. Jin, J. Huang, D. Cai, X. He, X.-S. Hua, Apparel-Invariant Feature Learning for Apparel-Changed Person Re-Identification, 2020 arXiv preprint arXiv:2008.06181.
[10] J. Dietlmeier, J. Antony, K. McGuinness, N.E. O'Connor, How important are faces for person re-identification?. Proc. ICPR 2021, pp. 6912–6919, https://doi.org/10.1109/ICPR48806.2021.9412340.
[11] F. Wan, Y. Wu, X. Qian, Y. Chen, Y. Fu, When person re-identification meets changing clothes, Proc. CVPRW 2020, pp. 830–831.
[12] Y.-J. Li, Z. Luo, X. Weng, K.M. Kitani, Learning Shape Representations For Clothing Variations in Person Re-Identification, 2020 arXiv preprint arXiv:2003.07340.
[13] Q. Yang, A. Wu, W.-S. Zheng, Person re-identification by contour sketch under moderate clothing change, IEEE TPAMI 43 (6) (2021) 2029–2046, https://doi.org/10.1109/TPAMI.2019.2960509.
[14] P. Zhang, Q. Wu, J. Xu, J. Zhang, Long-term person re-identification using true motion from videos, Proc. WACV, IEEE 2018, pp. 494–502.
[15] J. Xue, Z. Meng, K. Katipally, H. Wang, K. van Zon, Clothing change aware person identification, Proc. CVPRW 2018, pp. 2112–2120.
[16] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, Proc. IEEE ICCV 2018, pp. 650–667.
[17] P. Hong, T. Wu, A. Wu, X. Han, W.-S. Zheng, Fine-grained shape-appearance mutual learning for cloth-changing person re-identification, Proc. CVPR 2021, pp. 10513–10522.
[18] X. Jin, T. He, K. Zheng, Z. Yin, X. Shen, Z. Huang, R. Feng, J. Huang, X.-S. Hua, Z. Chen, Cloth-Changing Person Re-Identification from a Single Image with Gait Prediction and Regularization, 2021 arXiv preprint arXiv:2103.15537.
[19] X. Shu, G. Li, X. Wang, W. Ruan, Q. Tian, Semantic-guided pixel sampling for cloth-changing person re-identification, IEEE Signal Process. Lett. 28 (2021) 1365–1369, https://doi.org/10.1109/LSP.2021.3091924.
[20] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, Proc. IEEE ICCV 2017, pp. 2961–2969.
[21] Y. Xiu, J. Li, H. Wang, Y. Fang, C. Lu, Pose flow: efficient online pose tracking, Proc. BMVC 2018, pp. 1–12.
[22] Z. Yi, Q. Tang, S. Azizi, D. Jang, Z. Xu, Contextual residual aggregation for ultra high-resolution image inpainting, Proc. CVPR 2020, pp. 7508–7517.
[23] K. Ma, Z. Shu, X. Bai, J. Wang, D. Samaras, Docunet: document image unwarping via a stacked u-net, Proc. CVPR 2018, pp. 4700–4709.
[24] K. Wang, Z. Ma, S. Chen, J. Yang, K. Zhou, T. Li, A benchmark for clothes variation in person re-identification, Int. J. Intell. Syst. 35 (12) (2020) 1881–1898, https://doi.org/10.1002/int.22276.
[25] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, Proc. CVPR 2015, pp. 2197–2206.
[26] A. Kittur, E.H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, Proc SIGCHI Conf Hum Factor Comput Syst 2008, pp. 453–456.
[27] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, Proc. CVPR 2016, pp. 1239–1248.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. CVPR 2016, pp. 770–778.
[29] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, Proc. CVPRW 2019, pp. 1487–1495.
[30] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, Proc. CVPR 2018, pp. 7132–7141.
[31] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: enhancing learning and generalization capacities via ibn-net, Proc. ECCV 2018, pp. 464–479.
[32] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, Proc. CVPR 2018, pp. 2285–2294.
[33] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Leader-based multi-scale attention deep architecture for person re-identification, IEEE TPAMI 42 (2) (2019) 371–385.
[34] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, Proc. CVPR 2017, pp. 1318–1327.
[35] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE TPAMI 40 (6) (2017) 1452–1464.
[36] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2017 arXiv preprint arXiv:1412.6980.
[37] T. Ojala, M. Pietikä inen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognit. 29 (1) (1996) 51–59.
[38] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proc. CVPR, Vol. 1, IEEE 2005, pp. 886–893.
[39] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), Proc. ECCV 2018, pp. 480–496.
[40] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, Proc. ICCV 2017, pp. 764–773.
[41] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, Proc. NIPS - Volume 2, NIPS'15, MIT Press, Cambridge, MA, USA 2015, pp. 2017–2025.
[42] L. van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.
[43] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, Proc. 26th ACM-MM, ACM 2018, pp. 274–282, https://doi.org/10.1145/3240508.3240552.