Image and Vision Computing 117 (2022) 104335

Contents lists available at ScienceDirect

Image and Vision Computing

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

IRANet: Identity-relevance aware representation for cloth-changing person re-identification

Wei Shi [a], Hong Liu [a,*], Mengyuan Liu [b]

[a] Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Beijing 100871, China
[b] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China

ARTICLE INFO

Article history:
Received 19 July 2021
Received in revised form 27 October 2021
Accepted 1 November 2021
Available online 06 November 2021

Keywords:
Cloth-changing person re-identification
Identity-relevance
Feature representation

ABSTRACT

Existing person re-identification methods mainly focus on searching the target person across disjoint camera views in a short period of time. With this setting, these methods rely on the assumption that both query and gallery images of the same person have the same clothing. To tackle the challenges of clothing changes over a long duration, this paper proposes an identity-relevance aware neural network (IRANet) for cloth-changing person re-identification. Specifically, a human head detection module is designed to localize the human head part with the help of the human parsing estimation. The detected human head part contains abundant identity information, including facial features and head type. Then, raw person images in conjunction with detected head areas are respectively transformed into feature representation with the feed-forward network. The learned features of raw person images contain more attributes of global context, meanwhile the learned features of head areas contain more identity-relevance attributes. Finally, a head-guided attention module is employed to guide the global features learned by raw person images to focus more on the identity-relevance head areas. The proposed method achieves mAP accuracy of 25.4% on the Celeb-reID-light dataset, 19.0% on the Celeb-reID dataset, and 53.0% (Cloth-changing setting) on the PRCC dataset, which shows the superiority of our approach for the cloth-changing person re-identification task.

1. Introduction

Person re-identification, a key technique in multiple object tracking task [1,2], has a wide range of applications in intelligent surveillance, target association, and criminal investigation. Typical person re-identification methods are devoted to finding the target query person from a large number of candidate gallery images in a short period of time. Since typical person re-identification methods [3–5] are based on the assumption that both query and gallery images of the same person have the same clothing, they can be called short-term person re-identification (ST-ReID). Though great progress has been achieved, the ST-ReID methods cannot be well applied to more real situation where the target person may change clothing, such as a criminal taking off his jacket while escaping.

Cloth-changing person re-identification, short for CC-ReID, has been a hot topic recently, due to its more generalized setting than ST-ReID. Under this setting, numerous CC-ReID methods are proposed to overcome the cloth-changing challenges. The cloth-changing setting means that the appearance information of person will be not reliable. To this end, existing CC-ReID methods try to mine the identity information of the human body, e.g., facial information, shape characteristics, and gait features. According to the different ways of mining identity information, these CC-ReID methods can be roughly divided into two categories, including implicit identity feature learning and explicit identity feature learning. The implicit identity feature learning based CC-ReID methods are often dedicated to implicitly decoupling the clothes and shape structure. Specially, a large number of cloth-changing images which are generated with generative adversarial network, are introduced to the identity feature learning process. With these generated cloth-changing images, the clothing agnostic model [6,7] is proposed to learn identity-aware features by improving the intra-class compactness and inter-class separation. By contrast, implicit identity learning is less explanatory than explicit identity learning, and needs more complicated models to construct the relationship between cloth-changing samples and raw samples in the feature space.

Explicit identity feature learning based methods [8–11] aim to find the key body regions that contain the detailed identity information of persons. In [8], a learnable polar transformation is introduced to select discriminant curve patterns from the contour sketch of the human body. In [9], gait recognition as an auxiliary task is employed to drive

* Corresponding author.
E-mail addresses: pkusw@pku.edu.cn (W. Shi), hongliu@pku.edu.cn (H. Liu), nkliuyifang@gmail.com (M. Liu).

https://doi.org/10.1016/j.imavis.2021.104335
0262-8856/

the person re-identification model to learn cloth-agnostic representation by leveraging personal unique gait information. With the success of face recognition, both Wan et al. [10] and Xue et al. [11] jointly extracted identity information from the face and body of human, and fused them into an integrated embedding as discriminative identity representation. As is known to all, compared with shape and gait information, facial information is a more robust biometric cue for identification. However, the clear front faces of persons can not be easily captured from monitoring videos due to the large pose variations. Existing face detection methods are also mainly utilized to detect the front faces of persons.

In this work, an identity-relevance aware neural network (IRANet) by mining the head information rather than front face information is proposed for addressing CC-ReID. First, a human head detection module is designed to localize the human head without being affected by the head pose variations, like the side and half side of the human face. With the help of the human parsing estimation, the head area can be correspondingly cropped from the raw person image. Then, the raw person images and detected head areas are jointly mapped to deep feature space from global and local views. The learned features of the raw person images contain more attributes of global context, while the features of the detected head areas contain more identity-relevance attributes. Finally, a head-guided attention module is employed to guide the global features learned by raw person images to focus more on the identity-relevance head areas. Extensive experiments are conducted on the cloth-changing datasets, i.e., Cereb-reID-light, Cereb-reID and PRCC, demonstrating the advantage of our approach.

Generally, our contributions are four-fold:

- An identity-relevance aware neural network (IRANet) is proposed to address the cloth-changing issue in person re-identification task. A more reliable cue for cloth-changing person re-identification is introduced, by mining the identity-relevance information from head area.
- A human head detection module is designed to determine the position of the head and alleviate the situation where the face detection algorithm fails to detect non-front faces.
- To guide the identity feature learning with the human head, the head area and raw person image are represented in the high-level semantic feature space, respectively. A head-guided attention module is further used to highlight the head embedding in feature space.
- Extensive experiments are conducted on three large-scale cloth-changing person re-identification datasets, i.e., Cereb-reID-light, Cereb-reID and PRCC. Experimental results have shown that our method achieves higher matching rates than the competing methods.

2. Related work

2.1. Short-term person re-identification

Most short-term person re-identification (ST-ReID) approaches focus on designing discriminative pedestrian descriptors and robust distance metrics [12]. Traditionally, many effective hand-crafted pedestrian descriptors [13,14] and distance metrics [15,16] are developed to improve the performance of ST-ReID task. With the rapid advances of deep learning, a vast of learning based ST-ReID methods are investigated. Specifically, some researchers are dedicated to developing the deep metric learning [17,4,18] by comparing different identities online in the deep feature space. In addition, most ST-ReID works [5,19,20] focus on developing more discriminative deep identity embedding from global and local views.

Whether it is a method based on metric learning or a method based on representation learning, the core is to mine the most critical identity cues. To this end, Zhou et al. [21] proposed the cross-correlated attention network to explore the discriminative regions of the input person image. Zheng et al. [22] introduced a generative adversarial network to generate unlabeled samples, which regularized the supervised model to learn more identity-relevance features. Although these methods developed more discriminative cues for ST-ReID, they did not consider the influence of interference factors, such as occlusion and similar appearances. Considering the influence of the occlusion factor, Wang et al. [23] adaptively adjusted the importance of different parts of the human body for identity representation with the help of pose estimation. Xu et al. [24] were dedicated to addressing a more interesting task, the Black ReID task, which needed identifying the target person wearing black clothes from many candidate people wearing black clothes. To alleviate the influence of similar appearances, they exploited a head-shoulder descriptor to adaptively solve the situation of pedestrians wearing black clothes. Though these two methods alleviate the uncertainty of appearances, they either only pay attention to the situation where the appearance is partially obscured, or they need to introduce additional clothing annotations. Compared with these methods, this work aims to address the more challenging CC-ReID problem than ST-ReID task. By mining the identity-relevance cue from the human head, the proposed IRANet is more effective to cope with the clothing variations in the long-term scenarios.

2.2. Cloth-changing person re-identification

Recently, much attention has been paid to the challenging cloth-changing person re-identification task. To tackle the cloth-changing challenge, the typical solution is to eliminate the interference of clothes [25] or mine the identity-relevance cues, like facial information [10,11, 26], shape representation [6–8,27] and gait information [9]. Huang et al. [25] collected a large-scale CC-ReID dataset with 1,052 identities, and proposed the vector-neuron capsules to perceive cloth changes of the same person. However, this work relies on the deeper DenseNet-121 model [28] and the carefully designed multi-granularity features, which are not flexible enough to generalize to unseen scenarios. By contrast, it is more straightforward to mine the identity-relevance cues. However, the facial information [10,11] relies on the face detection methods, which can not be well utilized to detect non-front faces. The shape representations [6–8] are less explanatory and need more complicated models to construct the relationship between cloth-changing samples and raw samples in the feature space. The gait information [9] extracted from the videos will be more accurate than that extracted from a single image, while this work is devoted to addressing the image-based CC-ReID problem. Compared with these different identity cues mentioned above, this work proposes to mine the identity-relevance information from the human head area, which contains abundant facial information and head shape. Moreover, the proposed human head detection module in this work can be used to detect non-front faces caused by the camera view variations.

3. Methodology

3.1. Problem statement

The cloth-changing person re-identification (CC-ReID) task aims to retrieve the target person (Query) in a set of pedestrian candidates (Gallery), when the target person may change clothes. Let Q denote the query person, and let $G = \{G_i | i \in [1, b, ..., N]\}$ denote the gallery set. In gallery set G, G_i represents the i-th gallery image, and N is the total number of gallery images. By comparing Q with each gallery image in G, a ranked list $S = \{S_i | i \in [1, 2, ..., N]\}$ representing similarity scores can be further obtained. The corresponding identity label of each item in S can be used to define whether the retrieval result is the target person or not. This work is dedicated to developing the identity-relevance aware representation for the comparisons among the query person and gallery set in the high-level feature space.

3.2. Overview of the proposed IRANet

To address the cloth-changing challenges, this paper proposes an identity-relevance aware neural network (IRANet) for long-term person re-identification scenes by mining the identity-sensitive features. Fig. 1 shows the overall architecture of the proposed IRANet. As depicted in Fig. 1, the proposed IRANet is composed of three key modules, the human head detection (HHD) module, the feature embedding learning module, and the head-guided attention (HGA) module. For each input person image, the HHD module is first utilized to localize the human head area in the input raw person image. Specifically, a dense human parsing estimation module is used to determine the location of the human head without considering the head orientation. The reason why we localize the human head rather than the front face is that person images captured by surveillance cameras undergo large pose variations and camera view variations. According to the estimated location of the human head, the head area can be further obtained by cropping from the raw person image.

The human head area contains fine-grained identity cues, while the raw person image contains more global identity cues, like the shape representation. To this end, the raw person image and detected head area are fed into the body embedding branch and head embedding branch, respectively. To extract the identity features with high-level semantics, both body embedding branch and head embedding branch contain the global stream and local stream. Besides, the body embedding resulted from the body embedding branch also contains the head features. To enhance the importance of the human head in CC-ReID task, the HGA module is used to learn a fused global feature, which is guided by the head area. By joint supervising the body feature, head feature and fused feature with the Multi-Similarity (MS) loss L_{MS} and identification loss L_{id}, the identity-relevance aware feature embedding for CC-ReID can be learned in this work.

3.3. The human head detection module

The HHD module is designed to localize the head area in the raw person image against the different head variations. The details of the HHD module are illustrated in Fig. 2. Let $I_{train} = \left\{ I_{train}^k | k \in [1, 2, ..., M] \right\}$ with C identities denote the training set. The symbol M represents the total number of training images. For each training image I_{train}^k, it will be fed into the dense human parsing estimation component to estimate fine-grained body parts. Specifically, with the help of the DensePose method [29] used in the dense human parsing estimation component, the human body in I_{train}^k can be partitioned into 24 body parts on the human surface. These 24 body parts are composed of the head, torso, arms, legs, hands, and feet. In this work, both the left head and right head are collectively called head part for simplicity. Different from [29], we only utilize the head part for the subsequent identity feature learning rather than 24 body parts. The dense human parsing estimation is formulated as:

$$c^*(x,y) = argmax_c P(c|I_{train}^k(x,y)), \tag{1}$$

where (x, y) is the Cartesian coordinate of the pixel in I_{train}^k, $P(c|I_{train}^k(x,y))$ is the probability of pixel $I_{train}^k(x,y)$ belonging to the c-th body part, and $c^*(x,y)$ is the predicted body part of $I_{train}^k(x,y)$. The positions of pixels in the head part c_h can be obtained by:

$$X_h = \{x | c^*(x,y) = c_h\}, Y_h = \{y | c^*(x,y) = c_h\}, \tag{2}$$

where X_h and Y_h denote the all coordinates of the head part in I_{train}^k in the spatial domain. The location P_h of the head part utilized in this work can be further defined as:

$$\begin{aligned} X_h^{min}, X_h^{max} &= min(X_h), max(X_h), \\ Y_h^{min}, Y_h^{max} &= min(Y_h), max(Y_h), \\ P_h &= \left\{ (x,y) | X_h^{min} \leqslant x \leqslant X_h^{max}, Y_h^{min} \leqslant y \leqslant Y_h^{max} \right\}, \end{aligned} \tag{3}$$

where $min(\cdot)$ and $max(\cdot)$ denote the maximum and minimum operations, respectively. With the defined location P_h of the head part, the head area can be obtained with the appearance mapping, which is formulated as:

$$I_{head}^k = \left\{ I_{train}^k(x,y) | (x,y) \in P_h \right\}. \tag{4}$$

Here, the head area I_{head}^k is a rectangular area containing the foreground of the human head and the surrounding background, like the hair. The reason why we do not use only the foreground of the human head as I_{head}^k is that the subsequent embedding branches need the regular input, like the square and rectangle in the spatial domain. Moreover, the preserved surrounding background is useful for person re-identification, while removing the background may affect the structured information and smoothness of an image. This phenomenon can

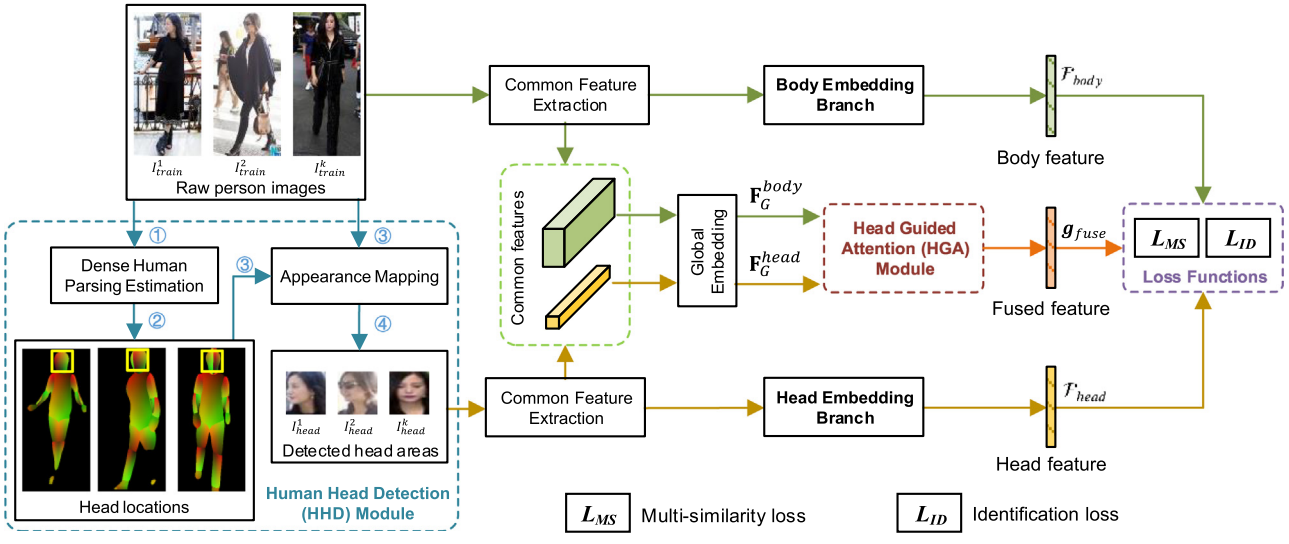

Fig. 1. Architecture of the proposed identity-relevance aware neural network. (**Best viewed in color**)

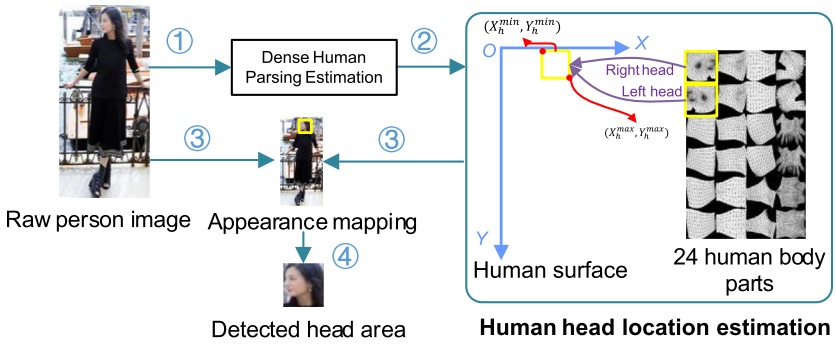

Fig. 2. Illustration of the human head detection module. The serial number denotes the execution order.

also be found in [30,31]. With the designed HHD module, the human head with identity-relevance cues is localized in the spatial domain.

3.4. Learning body and head embedding

To extract the high-level semantic feature representation from the human body and head, both the raw person image and the detected head area are fed into the Convolution Neural Network (CNN) model. In this work, an effective ST-ReID model, LightMBN [19], is employed as the backbone network. The LightMBN utilizes the truncated OSNet [32] as the common feature extraction part to extract the common features $\mathbf{F}$. Then, the common features are shared and fed into three distinct branches, including global branch, local branch and channel-based branch. In the initial stages of these three branches, the remaining layers of OSNet except for the truncated OSNet used, are employed to extract features for each branch. Specifically, the feature $\mathbf{F}_G$ extracted by the initial stage in the global branch is represented as:

$$\mathbf{F}_G = GlobalEmbedding(\mathbf{F}). \tag{5}$$

Based on $\mathbf{F}_G$, two 512-dimensional feature vectors $\mathbf{g}$ and $\mathbf{g}_{drop}$ are obtained from the global branch. The feature $\mathbf{g}$ is obtained by applying 2D average pooling on $\mathbf{F}_G$, while $\mathbf{g}_{drop}$ is obtained by applying the DropBlock [33] and 2D max pooling on $\mathbf{F}_G$. The feature $\mathbf{g}$ contains more global property, and the feature $\mathbf{g}_{drop}$ helps to improve the robustness of learned features owing to the DropBlock. Similarly, the features $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_g$ representing the upper body, lower body and global body, are extracted in the local branch following [19], respectively. To improve the representation ability of the channels in the CNN model, two features $\mathbf{c}_1$ and $\mathbf{c}_2$ are obtained in the channel-based branch in LightMBN. In this work, the backbone model mentioned above is used for learning the human body feature F^k_{body} from I^k_{train}. The feature F^k_{body} is a set of features learned by three distinct branches after passing each individual branch and BNNeck [34]. The structure of F^k_{body} can be denoted as:

$$F^k_{body} = \left\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}, \hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \hat{\mathbf{p}}_g, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\right\}, \tag{6}$$

where $\hat{}$ denotes the results passing BNNeck. All features in F^k_{body} have the same dimension. For the head area I^k_{head}, we adopt a similar model to the body embedding branch, but the weights are learned independently. Specially, the channel-based branch is removed in head embedding branch to avoid the overfitting problem. The obtained head feature F^k_{head} can be denoted as:

$$F^k_{head} = \left\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}, \hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \hat{\mathbf{p}}_g\right\}, \tag{7}$$

where $\hat{}$ also denotes the results passing BNNeck, and all features in F^k_{head} have the same dimension as features in F^k_{body}. Both F^k_{body} and F^k_{head} are further jointly optimized in the training phase. F^k_{body} denotes the k-th item in F_{body}, while the F^k_{head} denotes the k-th item in F_{head}.

3.5. Head-guided attention

Although both human body feature F^k_{body} and head feature F^k_{head} contain the identity information, the highlights of these two features are different due to the different inputs. In fact, the head area I^k_{head} is a part of the raw person image I^k_{train} in the spatial domain. For CC-ReID, the head area is more critical than the other parts of the human body, since the clothes of humans may change a lot. Moreover, the image I^k_{train} in addition to the head area also contains some identity information, like body shape. To highlight the importance of the human head, a head-guided attention (HGA) module is employed by modeling the relations between body embedding and head embedding. The detailed architecture of the HGA module is depicted in Fig. 3. Specifically, with Formula (5), the global outputs of I^k_{train} and I^k_{head} in the global branch can be obtained, respectively. Let $\mathbf{F}^{body}_G$ denote the global output of I^k_{train}, and let $\mathbf{F}^{head}_G$ denote the global output of I^k_{head}. Here, the features $\mathbf{F}^{body}_G$ and $\mathbf{F}^{head}_G$ learned by Formula (5) are used as the inputs of the

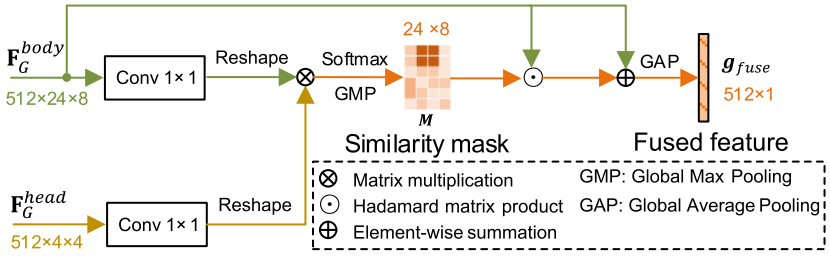

Fig. 3. Architecture of the head-guided attention (HGA) module. The term "Conv 1 × 1" denotes the convolution layer with the kernel size of 1 × 1.

HGA module, since both $\mathbf{F}_G^{body}$ and $\mathbf{F}_G^{head}$ are two dimensional in the spatial domain. Following [35], both $\mathbf{F}_G^{body}$ and $\mathbf{F}_G^{head}$ are transformed with a convolution layer with the kernel size of 1×1 for the dimension reduction, and reshaped. Then, the similarity mask M between $\mathbf{F}_G^{body}$ and $\mathbf{F}_G^{head}$ is calculated by the matrix multiplication, a Softmax layer, and a Global Max Pooling (GMP) layer, which is formulated as:

$$M = GMP(Softmax([C_1(\mathbf{F}_G^{body})]^\top C_2(\mathbf{F}_G^{head}))), \tag{8}$$

where both C_1 and C_2 represent the convolution layer with the kernel size of 1×1, and they have different weights. The learned similarity mask M can reflect the position of head area in $\mathbf{F}_G^{body}$. The fused output $\mathbf{F}_G^{fuse}$ where the head area is highlighted, can be obtained by:

$$\mathbf{F}_G^{fuse} = M \circ \mathbf{F}_G^{body} + \mathbf{F}_G^{body}, \tag{9}$$

where "∘" denotes the channel-wise Hadamard matrix product operation, and "+" represents the residual learning scheme. With the Global Average Pooling (GAP) layer, the output $\mathbf{F}_G^{fuse}$ can be further summarized as $\mathbf{g}_{fuse}$.

Algorithm 1 The algorithm flowchart of the IRANet model

Require: Training data I_{train}, initialized parameters $\mathbf{W} \leftarrow \mathbf{W}_0$, $\mathbf{W}_B \leftarrow \mathbf{W}_{B0}$,$\mathbf{W}_H \leftarrow \mathbf{W}_{H0}$, $\mathbf{W}_A \leftarrow \mathbf{W}_{A0}$ of the common feature extraction part, body embedding branch, head embedding branch, HGA module in the IRANet, $\alpha \leftarrow 0.5$, iteration number $l \leftarrow 0$ and optimizer setting.

Ensure: The parameters $\mathbf{W}$, $\mathbf{W}_B$, $\mathbf{W}_H$, $\mathbf{W}_A$.

Perform dense human parsing estimation by Formula (1).
Determine the human head position P_h by Formula (2) and (3).
Obtain the human head area I_{head} by Formula (4).
while not converge **do**
  Extract common features $\mathbf{F}_{body}$ from I_{train} by the $\mathbf{W}$.
  Learn the human body feature $\mathcal{F}_{body}$ from $\mathbf{F}_{body}$ by $\mathbf{W}_B$.
  Learn the global embedding $\mathbf{F}_G^{body}$ of $\mathbf{F}_{body}$ by Formula (5).
  Extract common features $\mathbf{F}_{head}$ from I_{head} by $\mathbf{W}$.
  Learn the human head feature $\mathcal{F}_{head}$ from $\mathbf{F}_{head}$ by $\mathbf{W}_H$.
  Learn the global embedding $\mathbf{F}_G^{head}$ of $\mathbf{F}_{head}$ by Formula (5).
  Calculate the similarity mask by Formula (8) and $\mathbf{W}_A$.
  Extract the fused output $\mathbf{F}_G^{fuse}$ by Formula (9).
  Summarize $\mathbf{F}_G^{fuse}$ as $\mathbf{g}_{fuse}$ with the GAP layer.
  Compute L_{ID} by Formula (10) for $\mathcal{F}_{body}$, $\mathcal{F}_{head}$ and $\mathbf{g}_{fuse}$.
  Compute the MS loss [34] L_{MS} for $\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}\}$ in $\mathcal{F}_{body}$, $\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}, \hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2\}$ in $\mathcal{F}_{head}$ and $\mathbf{g}_{fuse}$.
  Compute the overall loss L by Formula (11).
  Compute the gradients of loss to $\mathbf{W}$, $\mathbf{W}_B$, $\mathbf{W}_H$, $\mathbf{W}_A$.
  Update the parameters $\mathbf{W}$, $\mathbf{W}_B$, $\mathbf{W}_H$, $\mathbf{W}_A$ by the optimizer.
  Update the optimizer setting.
return $\mathbf{W}$, $\mathbf{W}_B$, $\mathbf{W}_H$, and $\mathbf{W}_A$

3.6. Network optimization

To make the features F_{body}^k, F_{head}^k and $\mathbf{g}_{fuse}$ learn the identity-relevance properties, several loss functions are used for training. Following [19], the Multi-Similarity (MS) loss L_{MS} [36] is used to supervise the global features $\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}\}$ in F_{body}^k, while the identification loss is used to supervise all features in F_{body}^k. The identification loss is defined as:

$$L_{ID} = -log(\frac{e^{\mathbf{w}^\top \mathbf{f}}}{\sum_{j\in[1,C]} e^{\mathbf{w}^\top \mathbf{f}_j}}), \tag{10}$$

where $\mathbf{f}$ denotes all features in F_{body}^k, and $\mathbf{w}$ denotes the corresponding classifiers. By contrast, the MS loss considers three types of similarities for a person pair: a self-similarity and two relative similarities. The MS loss is more suitable to train the deep metric learning based model. The identification loss which contains a fully-connected layer as the classifier, a Softmax layer, and a cross-entropy layer, is widely used to train the person re-identification model [7,5,37].

In the head embedding branch, the identification loss is also used to supervise all features in F_{head}^k. To further improve the representation ability of the human head area, the MS loss is utilized to supervise the features $\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}, \hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2\}$ containing both global and local features. For the fused identity feature $\mathbf{g}_{fuse}$ obtained by the HGA module, it is also fed into the MS loss and identification loss for the identity feature learning. The overall loss function is defined as:

$$L = \alpha L_{MS} + (1-\alpha) L_{ID}, \tag{11}$$

where α is an item that is used to balance two loss functions. To make the whole learning process of the proposed IRANet more clear, the algorithm flowchart of the IRANet is shown in Algorithm 1.

In the inference phase, we concatenate all global and local features $\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}, \hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \hat{\mathbf{p}}_g\}$ in F_{body}^k as the body embedding, and we concatenate all global and local features $\{\hat{\mathbf{g}}, \hat{\mathbf{g}}_{drop}, \hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \hat{\mathbf{p}}_g\}$ in F_{head}^k as the head embedding. If the human head can be detected, both the body embedding and head embedding will be used to calculate the distances between query Q and gallery G. If the human head cannot be detected, such as being affected by occlusion, then only the body embedding is used to calculate the distances between query Q and gallery G. This type of distance calculation method is more flexible to deal with the miss-detection and occlusion.

4. Experiments and discussions

4.1. Datasets

The extensive experiments are conducted on three large-scale CC-ReID datasets, Celeb-reID-light [25], Celeb-reID [25], and PRCC [8]. The examples of these three datasets are shown in Fig. 4, and the details of these three datasets are described as follows.

Fig. 4. Examples of Celeb-reID-light, Celeb-reID and PRCC datasets. Here, two identities in each dataset are shown, namely Person A and Person B.

Celeb-reID-light dataset: The person images in this dataset are crawled from the websites. Following [25], the Celeb-reID-light dataset is partitioned into two splits. The training split contains 9,021 images with 490 identities. In the test split, 887 images are treated as query images, and 934 images are treated as gallery set. The Celeb-reID-light dataset is challenging since all images of each person are in different clothes in this dataset.

Celeb-reID dataset: This dataset is larger than the Celeb-reID-light dataset, and it contains 34,036 images with 1,052 identities. Following [25], the Celeb-reID dataset is split into two splits for training and testing. The training split contains 20,208 images with 632 identities. In the test split, there are 2,972 query images and 11,006 gallery images. More than 70% of the images of each identity in the Celeb-reID dataset show different clothes on average.

PRCC dataset: The PRCC dataset contains 33,698 images with 221 identities, which are collected in real scenarios. The images in this dataset are captured from three different camera views. The same person under Camera A and Camera B wears the same clothes, while the same person under Camera A and Camera C wears different clothes. Following [8], the PRCC dataset is randomly split into two splits. The training split contains 150 identities, and the test split contains 71 identities.

4.2. Implementation details

The implementation in this work is based on PyTorch with one NVIDIA GeForce RTX 2080 Ti GPU. All input person images are resized to 384×128×3, and the detected head areas I_{head} are resized to 64×64×3. Following [19], the OSNet used in this work is also initialized by the ImageNet [41] pre-trained weights. The training batch size is set to 20, and each mini-batch contains 20 person images with 5 identities. The same data augmentation methods and optimizer settings as [19] are used in this work for fairness. All features in F^k_{body}, F^k_{head} and $\mathbf{g}_{fuse}$ contain 512 dimensions. The parameter α is set to 0.5 following [19]. We directly set $\alpha = 0.5$ to ensure the fairness of the comparison. In the inference stage, the cosine distance is calculated for the comparisons between query and gallery. The Cumulative Matching Characteristics (CMC) [42] and mean Average Precision (mAP) [43] are used to evaluate the proposed CC-ReID method. For the CMC metric, the key Rank 1 and Rank 5 results are reported in the experimental results following almost all existing CC-ReID methods [6–11].

4.3. Comparison with state-of-the-art methods

4.3.1. Celeb-reID-light dataset

Table 1 shows the comparisons between the proposed IRANet and the state-of-arts on Celeb-reID-light dataset. Both the representative ST-ReID methods (IDE [38]+DenseNet [28], MLFN [39], HACNN [40], MGN [20], LightMBN [19]) and CC-ReID methods (ReIDCaps [25], ReIDCaps+ [25], AFD-Net [7], CASE-Net [6]) are shown in Table 1. It can be seen that most ST-ReID methods behave worse than the CC-ReID methods. This phenomenon is understandable, since most ST-ReID methods rely on the appearance information, like clothes. In the Celeb-reID-light dataset, all instances of each person are in different clothes. Comparing with other CC-ReID methods, the proposed method can achieve better performance. Specifically, the proposed IRANet achieves Rank 1 accuracy of 11.1% higher than CASE-Net [6], 24.2% higher than AFD-Net [7] and 12.7% higher than ReIDCaps+ [25]. These improvements are attributed to the importance of head area for the cloth-changing scenarios. By contrast, the CASE-Net and AFD-Net methods focus on mining the human shape representation, but it is difficult to mine from a single person image.

Table 1
Comparisons between the proposed method and the state-of-arts on the Celeb-reID-light dataset. (Best results presented in bold font).

Method	Source	Rank 1 (%)	Rank 5 (%)	mAP (%)
IDE [38]+DenseNet [28]	–	10.5	24.8	5.3
MLFN [39]	CVPR 2018	10.6	31.0	6.3
HACNN [40]	CVPR 2018	16.2	42.8	11.5
MGN [20]	MM 2018	21.5	47.4	13.9
ReIDCaps [25]	TCSVT 2019	20.3	48.2	11.2
ReIDCaps+ [25]	TCSVT 2019	33.5	63.3	19.0
AFD-Net [7]	IJCAI 2021	22.0	51.0	11.3
LightMBN [19]	ICIP 2021	32.2	63.3	18.2
CASE-Net [6]	WACV 2021	35.1	66.7	20.4
IRANet (Ours)	–	**46.2**	**72.7**	**25.4**

4.3.2. Celeb-reID dataset

Table 2 shows the comparisons between the proposed method and the state-of-arts on challenging Celeb-reID dataset. Both the representative ST-ReID methods (IDE [38]+DenseNet [28], Zheng et al. [44], MLFN [39], PCB [5], HACNN [40], MGN [20], LightMBN [19]) and CC-ReID methods (ReIDCaps [25], ReIDCaps+ [25], Qian et al. [45], AFD-Net [7], CASE-Net [6]) are shown in Table 2. In the Celeb-reID dataset, more than 70% instances of each person change clothes on average, which reflects the uncertainty of the appearance information. The LightMBN model is a light-weight model by fusing global, local and channel-based feature, simultaneously, which achieves state-of-the-art performance on several ST-ReID datasets. The CC-ReID methods listed in Table 2 show the more superior results than the ST-ReID methods, which is attributed to the identity-relevance feature learning. The proposed IRANet method achieves 1.1% Rank 1 accuracy, 2.4% Rank 5 accuracy and 3.2% mAP accuracy higher than the ReIDCaps+ method, which reflects mining head information explicitly is more effective than learning whether pedestrians change clothes. Moreover, comparing with the usage of the human body shape information studied in Qian et al. [45], AFD-Net [7], and CASE-Net [6], the proposed method also achieves the better performance, which validates the importance of the head area for CC-ReID. It is more straightforward to learn identity-sensitive property from the head part, rather than barely extract the implicit shape representation. It can be found that the Rank 1 results of all methods on the Celeb-reID-light dataset are worse than the results on Celeb-reID dataset due to the challenging setting of Celeb-reID-light dataset. Most methods listed in Table 1 achieve higher mAP accuracy, since the Celeb-reID-light dataset is a light version of the Celeb-reID dataset. Moreover, the proposed method obtains higher improvements with other references on the Celeb-reID-light dataset than on the Celeb-reID dataset, which can better illustrate the effectiveness of the proposed method in the cloth-changing scene.

4.3.3. PRCC dataset

The comparisons with other representative ST-ReID methods (Zheng et al. [44], HACNN [40], MGN [20], PCB [5], HPM [46], LightMBN [19]) and CC-ReID methods (SPT [8], GI-ReID [9], CASE-Net [6], AFD-Net

Table 2
Comparisons between the proposed method and the state-of-arts on the Celeb-reID dataset. (Best results presented in bold font).

Method	Source	Rank 1 (%)	Rank 5 (%)	mAP (%)
IDE [38]+DenseNet [28]	–	42.9	56.4	5.9
Zheng et al. [44]	TOMM 2017	36.3	54.5	7.8
MLFN [39]	CVPR 2018	41.4	54.7	6.0
PCB [5]	ECCV 2018	37.1	57.0	8.2
HACNN [40]	CVPR 2018	47.6	63.3	9.5
MGN [20]	MM 2018	49.0	64.9	10.8
ReIDCaps [25]	TCSVT 2019	51.2	65.4	9.8
ReIDCaps+ [25]	TCSVT 2019	63.0	76.3	15.8
Qian et al. [45]	ACCV 2020	50.9	66.3	9.8
AFD-Net [7]	IJCAI 2021	52.1	66.1	10.6
LightMBN [19]	ICIP 2021	59.2	74.5	15.2
CASE-Net [6]	WACV 2021	**66.4**	78.1	18.2
IRANet (Ours)	–	64.1	**78.7**	**19.0**

Table 3
Comparisons between our proposed method and the state-of-arts on PRCC dataset. "Cloth-changing" and "Same clothes" denote the settings for CC-ReID and ST-ReID, respectively. (Best results presented in bold font).

Method	Source	Cloth-changing		Same clothes	
		Rank 1 (%)	mAP (%)	Rank 1 (%)	mAP (%)
Zheng et al. [44]	TOMM 2017	19.1	–	76.3	–
HACNN [40]	CVPR 2018	21.8	–	82.5	–
MGN [20]	MM 2018	33.8	35.9	99.5	98.4
PCB [5]	ECCV 2018	41.8	38.7	99.8	97.0
SPT [8]	TPAMI 2019	34.4	–	64.2	–
HPM [46]	AAAI 2019	40.4	37.2	99.4	96.9
GI-ReID [9]	arXiv 2021	37.6	–	86.0	–
CASE-Net [6]	WACV 2021	39.5	–	71.2	–
AFD-Net [7]	IJCAI 2021	42.8	-	95.7	-
LightMBN [19]	ICIP 2021	50.0	50.7	**100.0**	**99.4**
IRANet (Ours)	–	**54.9**	**53.0**	99.7	97.8

[7]) on the PRCC dataset are shown in Table 3. Under the "Cloth-changing" setting, the images captured from Camera C are treated as queries, while the images captured from Camera B are treated as queries under the "Same clothes" setting. The images captured from Camera A are used as a gallery set for two settings. The ST-ReID methods are indeed not tested on the PRCC dataset, due to the different scenarios. Therefore, the results collected from the existing CC-ReID work [6,26] are shown for fairness. Specially, the LightMBN is treated as the baseline in our work, and we rerun this method on the PRCC dataset. Under the "Cloth-changing" setting, the proposed method exceeds other compared methods, which shows the effectiveness of the proposed IRANet. Under the "Same clothes" setting, although the IRANet does not outperform the LightMBN method, it still achieves competitive results.

Specially, the gap between the IRANet and LightMBN under the "Same clothes" setting is because the same attention is paid to both the human body and head. This phenomenon is very common in existing CC-ReID works. Specifically, LightMBN is one of the most state-of-the-art ST-ReID methods, which performs well under the "Same clothes" setting. For LightMBN, it highlights the whole body features, containing the appearance representation, e.g., clothing information, as identity feature for re-identification. However, both our method and other existing CC-ReID methods are specially designed to address the cloth-changing problem. To this end, more attention is paid to the limited identity-relevance regions, such as head area, human body shape, and gait information, rather than the clothing information that is suitable for the "Same clothes" setting. The challenging setting of the CC-ReID task makes the mining of the identity features difficult. Even so, the proposed IRANet can still achieve competitive performance under the "Same clothes" setting. To balance the performance of the proposed method under the "Cloth-changing" and "Same clothes" settings, we pay the same attention to both head features and body features to avoid complicated hyper-parameter selection.

4.4. Ablation studies

The ablation studies on key components in the IRANet under the cloth-changing setting are shown in Table 4. In this work, the LightMBN method is used as the baseline, and the results of the improvements based on the baseline are shown for fairness. The method "IRANet w/o Body&HGA" denotes the proposed IRANet with only head embedding used, while "IRANet w/o Body" denotes the proposed IRANet with head embedding and HGA module. The method "IRANet w/o Head&HGA" denotes the proposed IRANet with only body embedding used, while "IRANet w/o Head" denotes the proposed IRANet with body embedding and HGA module. By comparing methods "IRANet w/o Body&HGA", "IRANet w/o Body", and baseline, it can be seen that only using the head area is still not enough to distinguish identities under the cloth-changing setting. A similar phenomenon about the influence of body embedding can also be found by comparing the baseline, "IRANet w/o Head&HGA", and "IRANet w/o Head". Moreover, it can be seen that the method "IRANet w/o HGA" achieves 7.9% mAP accuracy on the Celeb-reID-light dataset, 6.2% mAP accuracy on the Celeb-reID dataset, and 16.3% mAP accuracy on the PRCC dataset higher than the method "IRANet w/o Body&HGA". This result reflects that the fusion of body and head embedding is better than the situation where only head embedding is used. The comparisons between "IRANet w/o Head&HGA" and "IRANet w/o HGA" can also be used to prove that the fusion of body and head embeddings is better than the situation where only body embedding is used. The fusion of body embedding and head embedding can help to mine the human shape representation and head information, like facial features and head type. The comparisons among "IRANet w/o Body", "IRANet w/o Head" and "IRANet" can also be used to verify the effectiveness of the fusion of body and head embeddings when the HGA module is used. By comparing "IRANet" with "IRANet w/o HGA", it can be found that the HGA module can further improve the performance of fused identity features, since the head area can guide the feature learning of the body embedding branch.

Fig. 5 shows the visualization of the learned features on the challenging Celeb-reID-light dataset via t-SNE [47]. In Fig. 5, there are 10 different identities that are used to extract features by LightMBN and IRANet, each of which is shown in a unique color. To ensure the fairness of evaluation, only body features are used as identity features. From the comparison between lightMBN and IRANet, it can be seen that the features learned by LightMBN are relatively scattered, while the features learned by our IRANet are more compact and easy to distinguish. This phenomenon is attributed to the learned identity-relevance features guided by human head area in our work.

4.5. Feature visualization

To qualitatively analyze the proposed IRANet, the visualization of activation maps learned by the proposed method is shown in Fig. 6. Here, the showed activation maps are the outputs of the global embedding in the IRANet, namely $\mathbf{F}_G^{body}$. The activation maps can reflect what characteristics the learned model pays more attention to. From the comparisons among the three datasets, it can be seen that the highlights of learned activation maps on the Celeb-reID-light and Celeb-reID datasets are more centralized, while the results on the PRCC dataset are more dispersive. This phenomenon is mainly caused by the quality of the images, since the images with higher resolution are much easier to find the

Table 4
Ablation results of key components on three datasets under the cloth-changing setting. (Best results presented in bold font).

Method	Celeb-reID-light			Celeb-reID			PRCC	
	Rank 1 (%)	Rank 5 (%)	mAP (%)	Rank 1 (%)	Rank 5 (%)	mAP (%)	Rank 1 (%)	mAP (%)
Baseline	32.2	63.3	18.2	59.2	74.5	15.2	50.0	50.7
IRANet w/o Body&HGA	28.3	61.7	17.0	49.9	67.8	12.2	37.9	35.4
IRANet w/o Body	33.2	63.8	18.6	52.0	69.0	12.6	38.8	35.0
IRANet w/o Head&HGA	32.1	66.5	19.7	55.4	72.5	14.1	49.9	50.8
IRANet w/o Head	35.4	65.4	19.5	57.4	74.0	14.9	50.3	51.7
IRANet w/o HGA	41.8	70.7	24.9	62.3	78.5	18.4	51.3	51.7
IRANet (Ours)	**46.2**	**72.7**	**25.4**	**64.1**	**78.7**	**19.0**	**54.9**	**53.0**

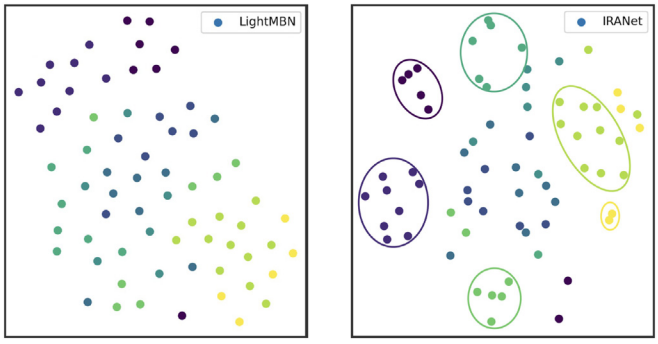

Fig. 5. Visualization of the learned features on the Celeb-reID-light dataset via t-SNE.

detailed identity cues, like fine-grained head information. More specifically, it can be found that high attention is paid to the human head area. This phenomenon further represents that the human head area is critical for the CC-ReID task.

From the results in the first row in Fig. 6, it can be observed that the human faces are clear enough, so both the baseline method and IRANet pay more attention to the face area. Moreover, our method aims to mine the identity cues from the human head area. It can be seen that our method highlights the whole head area including the face area. From the results in the second and third rows in Fig. 6, it can be seen that when the face is occluded or blurred, our method can not only focus on the head area of the pedestrian, but also consider the shape characteristics at the same time. However, the baseline method is easily affected in these situations. The images on the PRCC dataset are relatively obscure, and the learned activation maps have the larger responses on the head, arms, hands, legs and feet of pedestrians. Although the instances on the PRCC dataset may change the clothes, the shape representation can still be taken into account by focusing the arms, hands, legs and feet simultaneously. Despite this, the human head area is also crucial for the CC-ReID task by observing the results on the PRCC dataset.

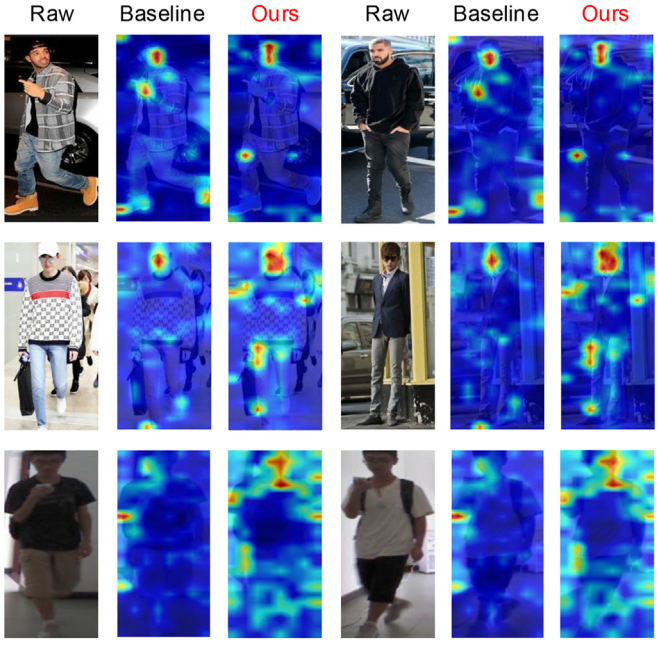

Fig. 6. Visualization of raw person images and the activation maps learned by LightMBN and our IRANet. The raw person images sampled from the Celeb-reID-light, Celeb-reID, and PRCC datasets, respectively. Each row shows two cloth-changing samples of each dataset.

5. Conclusion

This paper presents the IRANet model for the CC-ReID task by mining the identity-relevance information from the human head area. The designed human head detection module can localize the human head rather than only detect the visible facial region by the dense human parsing estimation. The more discriminative identity features can be learned with the guidance of the head embedding in the high-level semantic space. Moreover, the fusion of the body embedding and head embedding helps relieve the head occlusion problem. Extensive experiments conducted on three challenging CC-ReID datasets show the effectiveness of the proposed IRANet. In this work, the same attention is paid to the human body and head embedding which can avoid the complicated hyper-parameter selection. However, the more flexible online fusion method needs further studying to better balance two different identity features.

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.

Acknowledgements

This work is supported by National Key R&D Program of China (No. 2020AAA0108904), and Science and Technology Plan of Shenzhen (No. JCYJ20190808182209321).

References

[1] B. Munjal, A.R. Aftab, S. Amin, M.D. Brandlmaier, F. Tombari, F. Galasso, Joint detection and tracking in videos with identification features, Image Vis. Comput. 100 (2020) 103932.
[2] W. Shi, H. Liu, M. Liu, Identity-sensitive loss guided and instance feature boosted deep embedding for person search, Neurocomputing. 415 (2020) 1–14.
[3] H. Ye, H. Liu, F. Meng, X. Li, Bi-directional exponential angular triplet loss for RGB-infrared person re-identification, IEEE Trans. Image Process. 30 (2020) 1583–1595.
[4] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), Proceedings of the European Conference on Computer Vision 2018, pp. 480–496.
[6] Y.-J. Li, X. Weng, K.M. Kitani, Learning shape representations for person re-identification under clothing change, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2021, pp. 2432–2441.
[7] W. Xu, H. Liu, W. Shi, Z. Miao, Z. Lu, F. Chen, Adversarial feature disentanglement for long-term person re-identification, Proceedings of the International Joint Conference on Artificial Intelligence, 2021.
[8] Q. Yang, A. Wu, W.-S. Zheng, Person re-identification by contour sketch under moderate clothing change, IEEE Trans. Pattern Anal. Mach. Intell 43 (6) (2021) 2029–2046.
[9] 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 https://arxiv.org/pdf/2103.15537.pdfarXiv preprint arXiv:2103.15537.
[10] F. Wan, Y. Wu, X. Qian, Y. Chen, Y. Fu, When person re-identification meets changing clothes, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2020, pp. 830–831.
[11] J. Xue, Z. Meng, K. Katipally, H. Wang, K. van Zon, Clothing change aware person identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2018, pp. 2112–2120.
[12] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell (2021) In press.
[13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2010, pp. 2360–2367.
[14] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2015, pp. 2197–2206.
[15] D. Chen, Z. Yuan, G. Hua, N. Zheng, J. Wang, Similarity learning on an explicit polynomial kernel feature map for person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2015, pp. 1565–1573.

[16] B.J. Prosser, W.S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking, Proceedings of the British Machine Vision Conference, Vol. 2, 2010 1–6.
[17] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recogn. 48 (10) (2015) 2993–3003.
[18] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, Proceedings of the European Conference on Computer Vision 2016, pp. 791–808.
[19] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, G. Rigoll, Lightweight multi-branch network for person re-identification, Proceedings of the IEEE International Conference on Image Processing, 2021.
[20] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, Proceedings of the ACM International Conference on Multimedia 2018, pp. 274–282.
[21] J. Zhou, S.K. Roy, P. Fang, M. Harandi, L. Petersson, Cross-correlated attention networks for person re-identification, Image Vis. Comput. 100 (2020) 103931.
[22] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2017, pp. 3754–3762.
[23] H. Wang, X. Chen, C. Liu, Pose-guided part matching network via shrinking and reweighting for occluded person re-identification, Image Vis. Comput. 111 (2021) 104186.
[24] B. Xu, L. He, X. Liao, W. Liu, Z. Sun, T. Mei, Black re-id: a head-shoulder descriptor for the challenging problem of person re-identification, Proceedings of the ACM International Conference on Multimedia 2020, pp. 673–681.
[25] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, Z. Zhang, Beyond scalar neuron: adopting vector-neuron capsules for long-term person re-identification, IEEE Trans. Circuits Sys. Vedio Tech. 30 (10) (2019) 3459–3471.
[26] X. Shu, G. Li, X. Wang, W. Ruan, Q. Tian, Semantic-guided pixel sampling for cloth-changing person re-identification, IEEE Sig. Proc Lett. 28 (2021) 1365–1369.
[27] P. Hong, T. Wu, A. Wu, X. Han, W.-S. Zheng, Fine-grained shape-appearance mutual learning for cloth-changing person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 10513–10522.
[28] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2017, pp. 4700–4708.
[29] R. Alp Güler, N. Neverova, I. Kokkinos, Densepose: dense human pose estimation in the wild, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 7297–7306.
[30] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 1179–1188.
[31] D. Chen, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person search via a mask-guided two-stream CNN model, Proceedings of the European Conference on Computer Vision 2018, pp. 734–750.
[32] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 3702–3712.
[33] R. Quispe, H. Pedrini, Top-db-net: top dropblock for activation enhancement in person re-identification, Proceedings of the IEEE International Conference on Pattern Recognition 2021, pp. 2980–2987.
[34] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[35] S. Zhao, C. Gao, J. Zhang, H. Cheng, C. Han, X. Jiang, X. Guo, W.-S. Zheng, N. Sang, X. Sun, Do not disturb me: person re-identification under the interference of other pedestrians, Proceedings of the European Conference on Computer Vision, Springer 2020, pp. 647–663.
[36] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 5022–5030.
[37] R. Quispe, H. Pedrini, Improved person re-identification based on saliency and semantic parsing with deep neural network models, Image Vis. Comput 92 (2019) 103809.
[38] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2017, pp. 1367–1376.
[39] X. Chang, T.M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 2109–2118.
[40] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 2285–2294.
[41] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Proceedings of the Conference on Neural Information Processing Systems, Vol. 25, 2012 1097–1105.
[42] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, Proceeding of IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Vol. 3, , Citeseer, 2007 1–7.
[43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 1116–1124.
[44] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person reidentification, ACM Trans. Multimedia Comput. Commun. Appl. 14 (1) (2017) 1–20.
[45] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Long-term cloth-changing person re-identification, Proceedings of the Asian Conference on Computer Vision, 2020.
[46] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019 8295–8302.
[47] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res 9 (11) (2008).