This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

DCR-ReID: Deep Component Reconstruction for Cloth-Changing Person Re-Identification

Zhenyu Cui, Jiahuan Zhou, *Member, IEEE*, Yuxin Peng, *Senior Member, IEEE*,
Shiliang Zhang, *Senior Member, IEEE*, and Yaowei Wang, *Member, IEEE*

***Abstract*—Person re-identification (Re-ID) plays an important role in many areas such as robotics, multimedia and forensics. However, it becomes difficult when considering long-term scenarios, due to changing clothes irregularly for people. Therefore, cloth-changing person re-identification (CC-ReID) has attracted more attention recently. CC-ReID aims to identify the same person but with different clothes. Its main challenge is how to disentangle clothes-irrelevant features, such as face, shape, body, etc. Most existing methods force the model to learn clothes-irrelevant features by changing the colour of clothes or reconstructing people dressed in different colours. However, due to the lack of the ground truth for supervision, these methods inevitably introduce noises which spoil the discriminativeness of features and lead to uncontrollable disentanglement. In this paper, we propose a novel disentanglement framework, called Deep Component Reconstruction Re-ID (DCR-ReID), which can disentangle the clothes-irrelevant features and the clothes-relevant features in a controllable manner. Specifically, we propose a Component Reconstruction Disentanglement (CRD) module to disentangle the clothes-irrelevant features and the clothes-relevant features based on the reconstruction of human component regions. In addition, we propose a Deep Assembled Disentanglement (DAD) module, which further improves the discriminativeness of these disentangled features. Extensive experiments on three real-world benchmark CC-ReID datasets, LTCC, PRCC, and CCVID, are conducted to demonstrate the effectiveness of the proposed DCR-ReID. Empirical studies show that our DCR-ReID achieves the state-of-the-art performance against the other CC-ReID methods. The source code of this paper is available at https://github.com/PKU-ICST-MIPL/DCR-ReID_TCSVT2023.**

***Index Terms*—Person re-identification, Cloth-changing re-identification, Disentanglement, Component reconstruction.**

I. Introduction

PERSON re-identification (Re-ID) aims at matching the same person across different times and locations. With the rapid development of deep learning, deep neural network-based Re-ID methods have significantly advanced the progress of this challenging and important task. These methods mine discriminative features of people by extracting global or local features [1] [2] to obtain deep representations of pedestrians that can distinguish identities.

However, most of the existing methods [3] [4] [5] can only handle the short-term Re-ID scenario as shown in Fig. 1(a). More specifically, the short-term Re-ID usually assumes that people hardly change their clothes within a short duration. Therefore, the salient colour and texture information of clothes can be readily used in discriminative features extraction. However, such information is usually unfavourable in the long-term Re-ID scenario, as shown in Fig. 1(b), because people may change clothes irregularly. More than that, such information has a negative impact when different people wear similar clothes. Overcoming changes of clothes is one of the most critical challenges in the long-term Re-ID when considering how to extract clothes-irrelevant discriminative features. Therefore, some recent works have noticed such an important but challenging Re-ID setting, Cloth-Changing Re-ID (CC-ReID) [8] [9] [10]. Thus, in this paper, we specifically focus on facilitating CC-ReID by proposing a novel method.

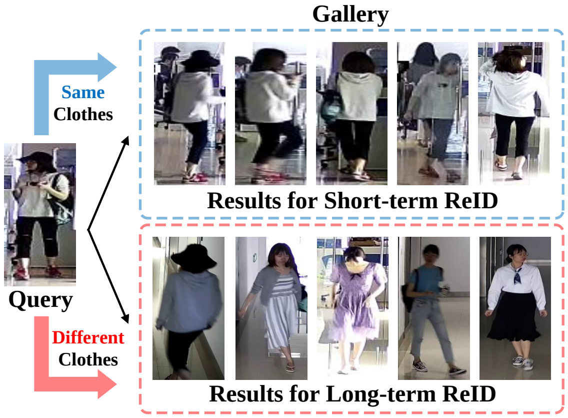


Fig. 1. The difference between the short-term Re-ID and the long-term Re-ID. The short-term Re-ID assumes that people hardly change their clothes within a short duration, while long-term Re-ID does not form such prior.

Existing solutions mainly aim to address CC-ReID based on two kinds of approaches: clothes-irrelevant features (fusion-based) [11] [12] and the disentanglement of clothes-irrelevant features (disentanglement-based) [7] [10]. Usually, disentanglement-based Re-ID methods outperform the fusion-based ones since the fusion-based method cannot comprehensively enumerate and fuse all discriminative clothes-irrelevant features in an ad-hoc manner, whereas the disentanglement-

This work was supported by the grants from the National Key R&D Program of China (2021YFF0901502) and the National Natural Science Foundation of China (61925201, 62132001).

Zhenyu Cui and Jiahuan Zhou are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China.
Yuxin Peng is with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China, and the Peng Cheng Laboratory, Shenzhen 518055, China.
Shiliang Zhang is with the School of Computer Science, Peking University, Beijing 100871, China.
Yaowei Wang is with the Peng Cheng Laboratory, Shenzhen 518055, China.
Corresponding author: Yuxin Peng (e-mail: pengyuxin@pku.edu.cn).

Copyright © 2023 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

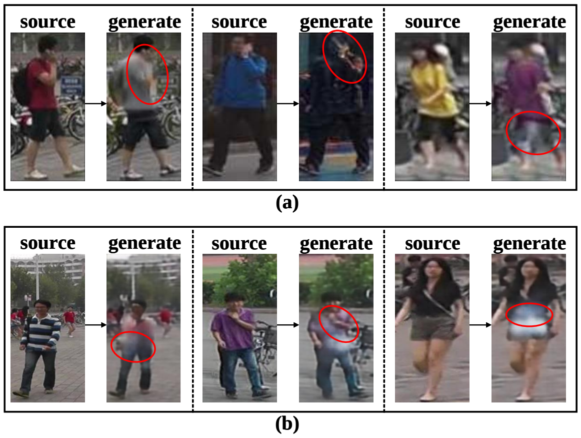


Fig. 2. **Bad Cases** generated by existing CC-ReID methods based on data-driven disentanglement. (a)[6] and (b)[7] failed on spoiling the most discriminative clothes-irrelevant information such as arms, faces, legs, and figures.

based method overcomes the problem by separating such features from the original image. Therefore, it preserves more identity-related information for identification. More specifically, the existing disentangle-based CC-ReID methods can be mainly categorized into two branches: data-driven disentanglement methods [7] [13] [14] [15] and feature-driven disentanglement methods [10] [16] [17] [18]. Data-driven disentanglement methods force the model to learn cloth-irrelevant features by randomly changing the pixel of clothing regions, thereby reducing its dependence on the colour and the texture information of clothes. To this end, GAN-based [7] [15] and some ad-hoc pixel resampling-based [13] methods are widely adopted. Although these methods can handle the aforementioned challenges in CC-ReID to some extent, their performance is also limited by the derived negative effects. More specifically, this is no guarantee that the generated images will not spoil the clothes-irrelevant features. As shown in Fig. 2, the most discriminative clothes-irrelevant information such as arms, faces, legs, and figures are spoiled in the generated image. Therefore, feature-driven disentanglement addresses this problem by introducing additional clothes-irrelevant branches with mutual learning [16] [19]. These methods hope that the features extracted from RGB images can be close to that from extra images, e.g., grey images, contour images, etc., thereby overcoming the changes of clothes' colours. However, this is a trade-off because it also ignores discriminative clothes-irrelevant features with colours, such as the colour of skin, bags, and shoes.

To mitigate the aforementioned limitations of existing CC-ReID methods, reconstruction learning [18] [17] is used to constrain the clothes-irrelevant features to some extent. These methods achieve disentanglement by reorganizing clothes-irrelevant features and clothes-relevant features of different identities extracted within the same batch and mapping them to the RGB domain. However, due to the lack of the ground truth, the reconstruction result cannot be directly supervised and controlled. These methods can only be supervised indirectly using additional discriminators [17] [6] [18], regularization terms [20] [6] [21] [22] [18], or mapping back to the source image [17] [23] [6] [18]. Therefore, these methods are uncontrollable, because the clothes-irrelevant features extracted by these methods cannot be restricted and reconstructed independently. As a result, the clothes-irrelevant features are inevitably interfered by the clothes-relevant features and thus lead to limited performance and reconstruction effects. In this paper, we aim to tackle the above issue.

We argue that if the reconstruction is controllable, it can facilitate the disentanglement of the clothes-irrelevant features, thereby improving the accuracy of CC-ReID. To this end, we rethink the purpose of the reconstruction. In CC-ReID, the image is reconstructed to express the clothes-irrelevant features and the clothes-relevant features extracted by the model and supervised at the pixel level. Therefore, we need to ensure that these features correspond to the human component regions, as well as the discriminativeness of these features. More specifically, these component regions include the clothing region and the non-clothing region. Therefore, such spatial correspondence can be modelled by the human component region. In addition, as mentioned above, changing the pixel of the image will spoil the discriminative clothes-irrelevant information. Therefore, we need to improve the discriminativeness based on the extracted deep features, which can be learned from the clothes classifier.

From this point, in this paper, we propose a novel deep component reconstruction-based method for CC-ReID, called DCR-ReID. The core idea of DCR-ReID is to disentangle the clothes-irrelevant features and the clothes-relevant features into specific features, and directly remove the clothes-relevant features for inference to achieve controllable disentanglement. Specifically, DCR-ReID disentangles the deep features into specific feature segments by the proposed Component Reconstruction Disentangle (CRD) module. Based on the disentangled features, we further propose a Deep Assembled Disentangle (DAD) module, which improves the discriminativeness of the disentangled features. In summary, the main contributions of this paper are as follows:

1) We propose a novel method for CC-ReID, called DCR-ReID, which can achieve controllable disentanglement for the clothes-irrelevant features and the clothes-relevant features. For inference, we directly remove the clothes-relevant features to achieve controllable disentanglement.

2) A Component Reconstruction Disentanglement (CRD) module is proposed to disentangle the clothes-irrelevant features and the clothes-relevant features based on human component regions, which facilitates that the disentangled features correspond to the human component regions.

3) A Deep Assembled Disentanglement (DAD) module is proposed to improve the discriminativeness of the disentangled features, avoiding the spoiling of the most discriminative clothes-irrelevant information.

4) Extensive experiments on three real-world benchmark Re-ID datasets demonstrate that our proposed DCR-ReID shows better results against the state-of-the-art CC-ReID methods.

The rest of this paper is organized as follows: Section

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

II gives a brief review of related work about person re-identification. Section III presents the procedure of the proposed DCR-ReID. Section IV shows the details, results, and analysis of the experiment. Section V concludes the paper.

II. Related Work

A. Short-Term Re-ID

Short-Term Re-ID assumes that people hardly change their clothes within a short duration. Therefore, such methods can use the colour and texture information of clothes for discriminative features extraction. Specifically, when handling the image data, global and local features (e.g., pose and body) [1] [24] [2] [25] [26] [27] are usually considered as discriminative features. Global features extract the global discriminative information of people. Zheng et al. [28] proposed treating each person as a separate class and utilize a multi-class loss function to improve the discriminativeness of the global features. However, global features fail to distinguish people with tiny differences, such as hair, shoes, etc. Therefore, local features are used to extract local information as discriminative features. Sun et al. [29] proposed vertically decomposing the deep convolutional features into multiple parts, and then used each part to learn identity independently to improve local features. When handling the video data, temporal features [30] [31] such as gait and motion, are extracted to enhance the discriminativeness of features. Omar et al. [30] proposed a gait recognition-based method for Re-ID. It fuses the estimated angle of the gait for gait prediction to improve the discriminative features. Besides extracting the above visual features, multi-modal features (e.g., viewpoints and attributes) [32] [33] [34] are also used to improve the discriminative features. Su et al. [32] proposed a multi-stage attribute enhanced Re-ID method. It firstly sets the semantic ground truth by the one-hot encoding of the pre-defined attributes, and then fuses the predicted attributes to improve the discriminative features.

Although the above methods can handle the short-term Re-ID scenario, they cannot address the long-term Re-ID, because these methods rely too much on salient information about clothes. Therefore, it is necessary to study long-term Re-ID.

B. Long-Term Re-ID

Although long-term Re-ID is more important and realistic in practice, it is also much more challenging. Because the same person may wear completely different clothes as well as different persons may wear similar clothes, the distraction of cloth information severely limits the performance of existing Re-ID methods. Adapted from general re-identification, some works are dedicated to solving CC-ReID [35] [36] using well-designed network structures and regularization terms. However, these methods are inevitably affected by clothes-relevant features, such as clothing color, style, etc. Therefore, several disentangle-based long-term Re-ID methods have been proposed aiming to disentangle clothes-irrelevant features from the discriminative features. Specifically, these methods can be divided into two categories: the data-driven disentanglement method [7] [13] [14] [15] [37] and the feature-driven disentanglement method [10] [16] [17] [38] [18].

The data-driven disentanglement method aims at making the CC-ReID model less dependent on clothing colours and textures by simulating the same person wearing different clothes. Shu et al. [13] proposed a data augmentation method by randomizing clothes' colours. It uses a body parser to collect the clothing regions and reallocate its pixel values in a mini-batch to augment the training data. Jia et al. [14] proposed a semantic-aware patching strategy for data augmentation. It randomly assembles the clothing patches to simulate the appearances of the same person wearing different clothes. However, the above prior knowledge-based data-driven disentanglement methods always spoil the shape information, which is entangled with the appearance of clothes, thus limiting the further development of CC-ReID. Recently, owing to the superior reconstruction ability of GAN [39], several GAN-based methods [7] [15] are proposed to simulate images of the same person wearing different clothes. However, due to the lack of ground truth in the target domain of the synthetic image and the inability to express fine-grained information, such as the hair and the shoes, these methods usually generate unreasonable clothes and spoil the discriminative information for feature extraction. Therefore, these methods are not only complex in structure but also have limited performance.

In addition, the feature-driven disentanglement CC-ReID methods focus on learning clothes-irrelevant features by various regularization terms for deep feature learning. Hong et al. [16] proposed a two-stream CC-ReID network, which uses pose and shape features to avoid the influence of clothes' colours. It uses interactive mutual learning to force the appearance stream to learn structural information such as pose and shape from the shape stream. Chen et al. [19] proposed to mutually learn colour and contour images to mine reliable shape-aware features. It leveraged contour feature learning as regularization and excavated more efficient shape-aware feature representations from colour images by maximizing the mutual information between colour appearance features and contour shape features. Yaghoubi et al. [38] proposed a long-term feature representation learning method to handle CC-ReID. The reconstruction learning is adopted to obtain a short-term representation without the most relevant biometric information. The difference between the long-term representation and the short-term representation is maximized, so that the obtained feature representation is clothes-irrelevant. CASE-Net [17] learns identity representations that only depend on body shape through adversarial learning and feature decomposition. It uses multiple instances of the same identity to assign colours to each other, and uses an adversarial loss function to evaluate the quality of image restoration to disentangle clothing appearance features. Gu et al. [10] proposed a two-stage implicit disentanglement method. It trains a clothes classifier to learn clothes-relevant features in the first stage, and then forces the model to learn clothes-irrelevant features in the second stage by optimizing a clothes adversarial loss function.

To sum up, data-driven disentanglement-based methods can transform different clothes in an ad-hoc manner, but may spoil the most discriminative clothes-irrelevant information as well as generate unreasonable clothes. For example, as

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

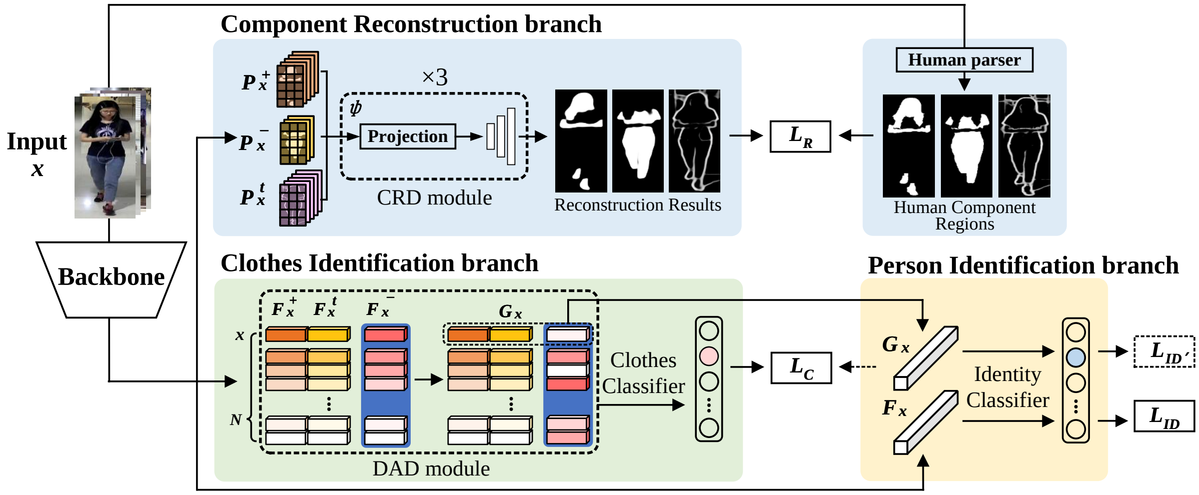


Fig. 3. The architecture of the proposed DCR-ReID. DCR-ReID consists of three branches: a person identification (PI) branch, a component reconstruction (CR) branch, and a clothes identification (CI) branch. Among them, the PI branch is the main branch to distinguish instances with different identities. The CR branch uses CRD to disentangle the clothes-irrelevant features P_x^+ and the clothes-relevant features P_x^- based on the human component region, and the contour features P_x^t. The CI branch uses DAD to further improve the discriminativeness of the disentangled features based on the assembled feature G_x.

shown in Fig. 2, the generated image cannot restore the details of clothes, so the shape of a person cannot be accurately judged. Feature-driven disentanglement-based methods force the model to learn clothes-irrelevant features, such as shape, body, etc. However, these features cannot be enumerated comprehensively, resulting in incomplete learning for discriminative features, thus limiting the performance. Although CAL[10] proposed to use adversarial loss function to learn clothes-irrelevant features, there is no intuitive evidence of what the model actually learned.

C. Reconstruction in Re-ID

In recent years, reconstruction has received extensive attention for its ability to model semantics and their context in visual representations [40] [41] [42]. In Re-ID systems, reconstruction learning is usually used for occluded person Re-ID [20] [23] [21] [22] [43], visible-infrared person Re-ID [44] [45], as well as CC-ReID [17] [18].

In CC-ReID, reconstruction learning is mainly used to disentangle the clothes-irrelevant features and the clothes-relevant features. In order to jointly optimize reconstruction learning and ReID, Zheng et al. [6] proposed an end-to-end joint learning framework. It decomposes each instance into appearance codes and structural codes, and then generates reconstruction images by combining codes from different instances. The reconstructed image are fed back to the generator and the discriminator in an online manner. Li et al. [17] proposes a reconstruction method based on structural identity and colour features. It uses the grey image as structural information and reconstructs the original image by fusing the colour features of different people with the same identity. Xu et al. [18] proposed an adversarial feature disentanglement network. It reconstruct intra-class instances to reduce intra-class feature variation, while generate inter-class adversarial clothes-changing instances to improve the clothes-irrelevant features.

However, due to the lack of ground-truth guidance for the generated images, the above reconstruction-based methods usually suffer from inferior reconstruction performance. The spatial information is likely to be spoiled and uncontrollable. In this paper, we proposed a novel DCR-ReID method to tackle the above issues. The reconstruction is performed directly on the human component regions in the target domain, so the reconstruction results can directly control the disentanglement. By doing this, our method can explicitly remove clothes-relevant features. The experimental results verify the effectiveness of the proposed method.

III. Our Approach

In this section, we detail our proposed DCR-ReID method. First, the problem definition of CC-ReID is illustrated. Then, the details of the proposed DCR-ReID are presented.

A. Preliminary.

Similar to the image retrieval task [46], person Re-ID aims to retrieve the same-identity images from a gallery set according to the probe. To do so, the instances in the gallery set are ranked based on the similarity distance to the probe, where the higher the instances with the same identity are ranked, the better the Re-ID performance is. Formally, let $\{g_i\}_{i=1}^{N}$ be the gallery set with N person images belonging to L different identities, and $\{q_j\}_{j=1}^{M}$ represent the probe set with M person images. Suppose that the model $\phi(\cdot;\theta)$ is parameterized by θ, the purpose is to determine the identity of q_j, such that:

$$i^* = \underset{i=1,\ldots,N}{\arg\min}\, d(\phi(q;\theta),\phi(g_i;\theta)) \tag{1}$$

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

, where $d(\cdot;\cdot)$ represents the similarity function calculated by the distance of two vectors.

In CC-ReID, the model learning becomes much more challenging since there are many people of different identities wearing similar clothes. This challenge can be formalized as:

$$d(w(q_{j'}), w(g_i)) >> d(w(q_j), w(g_i)) \tag{2}$$

, where $w(\cdot)$ represents the clothes feature of the given instance. Among them, q_j and g_i have the same identity but are different from $q_{j'}$.

Suffering from the above issues, the extracted features are not discriminative enough to distinguish these people since the intra-identity distance is usually larger than the inter-identity one.

$$||d(\phi(q_j;\theta), \phi(g_i;\theta)) - d(\phi(q_{j'};\theta), \phi(g_i;\theta))|| >> 0 \tag{3}$$

In this case, in order to improve the discriminativeness, CC-ReID is committed to optimize the following formulation:

$$\min \sum_{j=1}^{M} \sum_{j'=1}^{M} (d(\phi(q_j;\theta), \phi(g_i;\theta)) - d(\phi(q_{j'};\theta), \phi(g_i;\theta)) \tag{4}$$

B. DCR-ReID

To tackle the aforementioned critical issues in CC-ReID, we propose a novel method named DCR-ReID (Deep Component Reconstruction-ReID), which aims to disentangle the clothes-irrelevant features and the clothes-relevant features in a controllable manner based on the reconstruction of human component regions. As shown in Fig. 3, our proposed DCR-ReID consists of three branches:

- The person identification (PI) branch is the main model to distinguish instances with different identities. During training, this PI branch is trained via an identity classifier. For inference, only the clothes-irrelevant features extracted by the backbone network is kept for final evaluation.
- The component reconstruction (CR) branch performs controllable disentanglement based on the human component regions. A novel module, called CRD, is proposed to reconstruct the customized feature segments of the clothes-irrelevant features and the clothes-relevant features to the corresponding human component regions. In addition, the contour of the human is also reconstructed independently within the clothes-irrelevant feature segments.
- The clothes identification (CI) branch is a two-stage disentanglement branch to improve the discriminativeness of the clothes-irrelevant features and the clothes-relevant features. In the first stage, CI trains a clothes classifier to learn clothes-relevant features while using an adversarial loss function [10] to specifically learn clothes-irrelevant features. In the second stage, we propose a novel module, called DAD, to further improve the discriminativeness of the learned features using the assembled feature segments.

Then, we further elaborate the model design and feature modeling of the proposed DCR-ReID. Given an image for query, DCR-ReID first extracts deep features using the backbone network. Considering that the Re-ID task is to discriminate the identities, we put the extracted deep features into the PI branch, which learns identity features through an identity classifier. Consequently, such features are inevitably interfered by the clothes information. Therefore, DCR-ReID performs disentanglement learning on the extracted deep features, aiming to disentangle the clothes-irrelevant features and the clothes-relevant features, and remove the clothes-relevant features to realize the modeling of the clothes-irrelevant features. To this end, we propose the CR branch, which uses the reconstruction network to independently reconstruct the clothes-irrelevant regions and the clothes-relevant regions corresponding to the customized feature segments. Consequently, although the extracted features can be reconstructed into the corresponding human component regions, they are not discriminative enough due to the lack of the feature discriminativeness. Therefore, we propose the CI branch to further improve the discriminativeness of the learned features using the identity classifier and a clothes classifier. For inference, only the clothes-irrelevant features extracted by the backbone network is kept for final evaluation, and the clothes-relevant features are removed directly. Finally, the proposed CR branch collaborates together with the CI branch, and achieves a controllable disentanglement.

*1) **Learning person identification**:* Considering that Re-ID is mainly to match different people across the gallery, the goal of our proposed PI branch is to classify people with different identities. Therefore, we use an identity classifier to learn discriminative identity features. Specifically, let $C_{ID}(\cdot|l)$ be the identity classifier. Given an instance x_i with identity label l_i, the identification loss function $\mathcal{L}_{ID}$ can be formulated as:

$$\mathcal{L}_{ID} = -\sum_{i=1}^{N} log \left(\frac{y(x_i, l_i)}{\sum_{j=1}^{N_{ID}} y(x_i, l_j)} \right) \tag{5}$$

$$y(x_i, l) = C_{ID}(\phi(x_i;\theta)|l) \tag{6}$$

, where N and N_{ID} are the number of instances and the number of identities, and $y(x_i, l)$ represents the predicted probability of x_i for identity label l. By minimizing $\mathcal{L}_{ID}$, DCR-ReID can preserve discriminative identity features.

*2) **Learning component reconstruction**:* The core of DCR-ReID is to disentangle the clothes-irrelevant features and the clothes-relevant features in a controllable manner. In general, the clothes-relevant features correspond to the clothing regions, while the clothes-irrelevant features correspond to the non-clothing regions of the human component. Therefore, we aim to restrict these features to the corresponding human component regions. To this end, we propose the component reconstruction (CR) branch, which uses the proposed CRD to disentangle the clothes-irrelevant features and the clothes-relevant features by reconstructing the corresponding regions.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

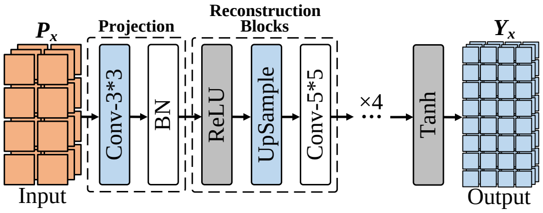


Fig. 4. The architecture of the reconstruction network. It mainly consists of a projection layer and four reconstruction blocks. Among them, the projection layer normalizes the input features, and the reconstruction blocks reconstruct the human component regions. Finally, the Tanh activation function [47] is used to accelerate the convergence of the reconstruction network.

Compared with the mainstream reconstruction of the RGB image of the component regions [17] [18], we propose reconstructing the binary image because it can represent the corresponding regions more consistently, and thus simplify and stabilize the learning process of reconstruction.

Specifically, given an instance x_i, we propose using the original convolutional features P_i to better reconstruct the human component regions in the visual space. The calculation process of P_i can be derived from the following formula:

$$P_i = E(x_i;\theta) \tag{7}$$

, where $E(\cdot;\cdot)$ represents the backbone network. Then, we decompose P_i into the clothing region features P_i^- and the non-clothing region features P_i^+. However, the contour features P_i^t are ignored, which are closely related to P_i^- and P_i^+ and are important for CC-ReID. Therefore, we explicitly decompose P_i into the three parts of feature segments to reconstruct them separately as the following formula:

$$P_i = P_i^- \oplus P_i^+ \oplus P_i^t \tag{8}$$

, where $\oplus$ represents the channel-wise concatenation. Eq. 8 separates the extracted deep features by channel-level splitting, which prevents inter-channel information mixing. It protects the disentangled features from interfering with each other and achieving controllable disentanglement. Then, we propose a reconstruction network to reconstruct P_i^-, P_i^+, and P_i^t separately. The structure of the reconstruction network is shown in Fig. 4. For convenience, we define the three reconstruction networks with the same structure as ψ^-, ψ^+, and ψ^t for P_i^-, P_i^+, and P_i^t, respectively. Therefore, we obtain the corresponding reconstruction results Y_{c-}, Y_{c+}, and Y_{ct} with the following formula:

$$Y_i^\upsilon = \psi^\upsilon(P_i^\upsilon), \upsilon \in \{\pm, t\} \tag{9}$$

To supervise the reconstruction results in a controllable manner, we obtain the corresponding ground truth image through a pre-trained human parser [48] and a pre-trained edge detector [49] without any fine-tuning. The aim of the CRD module is to restrict the disentangled clothes-irrelevant features and the clothes relevant features to the corresponding human component regions. Thus, it can be regularized by reconstructing the human component regions based on the disentangled features. That is because directly using the human parser can only predict the binary non-clothing region mask on the original image, which cannot realize the above restrictions as it cannot bring any regularization effect to the disentanglement. More specifically, we use the human parser [48] to extract the clothing and non-clothing regions for x_i, which are represented as T_i^- and T_i^+. Meanwhile, we use the edge detector [49] to obtain the ground truth map T_i^t for the contour reconstruction. [49] is a widely-used edge detector in many previous researches [8] [9] [16]. We employ it in DCR-ReID for the fair comparison. Then the total loss function of CR branch is calculated as follows:

$$\mathcal{L}_R = \frac{1}{N}\sum_{i=1}^{N}\sum_{\upsilon\in\{\pm,t\}} l_1(T_i^\upsilon, Y_i^\upsilon) \tag{10}$$

The weights of the loss functions for the three branches are the same, since the three branches are equally critical to the CRD. In summary, CR branch uses the component reconstruction to model the correspondence between the disentangled features and the human component regions, and thus achieves controllable disentanglement.

*3) **Learning clothes identification:*** As discussed above, CR branch facilitates the correspondence of the disentangled features and the human component regions. However, CR branch can hardly keep the discriminativeness of the disentangled features, due to the lack of supervision over the corresponding feature segments. Therefore, we propose the clothes identification (CI) branch to further improve the discriminativeness of the clothes-irrelevant features and the clothes-relevant features of the specific feature segments.

To this end, we design CI as a two-stage optimization process. In the first stage, CI uses a clothes classifier to learn the clothes-irrelevant features and the clothes-relevant features. In the second stage, CI uses the proposed DAD to further improve the discriminativeness of the learned features.

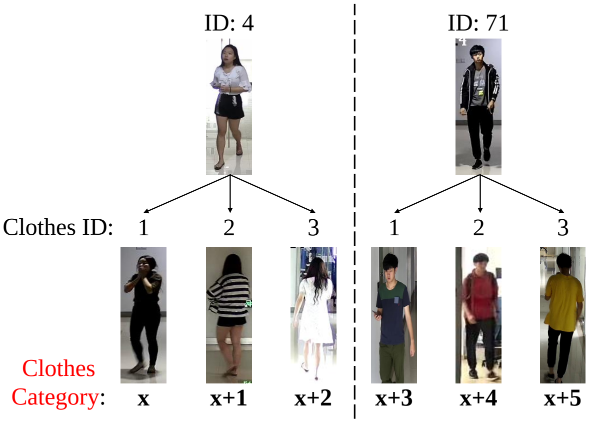


Fig. 5. The illustration of the clothes category. '**x**' represents the clothes category number. The clothes category is defined as the fine-grained category under the identity.

More specifically, first of all, let $C_P(\cdot;c)$ a clothes classifier for clothes category c. Here, c is defined as the fine-grained

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

category under the identity, as shown in Fig. 5. Therefore, the same person wears different clothes or two people wear similar clothes are considered different fine-grained categories. Because it is costly to compare every two instances one by one to determine whether they are wearing the same clothes in reality. Given an instance x_i, with clothes label c_i, in the first stage, CI uses the clothes classification loss function $\mathcal{L}_c$ to learn the clothes-relevant features as follows:

$$\mathcal{L}_c = -\sum_{i=1}^{N} log\left(\frac{u(x_i, c_i)}{\sum_{j=1}^{N_U} u(x_i, c_j)}\right) \tag{11}$$

$$u(x_i, l) = C_P(\phi(x_i; \theta)|l) \tag{12}$$

, where l represents the identity label of x_i, and $u(x_i, l)$ represents the the predicted probability of x_i for identity label l.

Next, similar to CAL [10], we introduce a clothes adversarial loss function $\mathcal{L}_{ca}$ to learn the clothes-irrelevant features, which can be formalized as follows:

$$\begin{aligned}\mathcal{L}_{ca} = &-\sum_{i=1}^{N}\sum_{c=1}^{N_U} q(c)\\ &\times log\left(\frac{u(x_i, c)}{u(x_i, c) + \sum\limits_{id(j)\neq id(i)} u(x_i, j)}\right)\end{aligned} \tag{13}$$

$$q(c) = \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{K}, c = c_i \\ \dfrac{\epsilon}{K} \qquad , c \neq c_i \textbf{ and } id(c) = id(i) \\ 0 \qquad , id(c) \neq id(i) \end{cases} \tag{14}$$

, where N_U represents the number of clothes categories, K represents the number of clothes categories under $id(i)$, and ϵ is a hyper-parameter set to 0.1. Eq. 13 is used to learn clothes-irrelevant features. However, learning clothes-irrelevant features cannot come at the expense of general performance. If the probabilities of different clothes categories are the same, the Re-ID performance of the proposed method in general scenarios and its generalizability will be weakened. As a result, we remain the corresponding clothes category has the highest weight $(1 - \epsilon + \epsilon/k)$. The essential difference between the two losses in Eq. 11 and Eq. 13 is that the former is the classification of single-positive-class for training a robust classifier to learn the clothes-relevant features, while the latter is the classification of multiple-positive-class for learning the clothes-irrelevant features, respectively.

To improve the discriminativeness of the learned clothes-irrelevant features and the clothes-relevant features, we proposed to use the DAD module in CI branch. Compared with the mainstream improvement strategy based on the GAN [7] or pixel resampling [13] in RGB images, DAD assembles clothes-relevant features from different identities in deep feature vectors, and thus avoiding the spoiling of the most discriminative clothes-irrelevant information. Then, we detail the DAD.

let F_i^v be the feature vectors calculated from P_i^v as follows:

$$F_i^v = \varphi(P_i^v), v \in \{\pm, t\} \tag{15}$$

, where $\varphi(\cdot)$ represents the batch normalization [50] with the global pooling. For the obtained feature vector F_i^-, F_i^+, and F_i^t, we randomly shuffle F_i^- and F_i^t to get F_{ai}^- and F_{ai}^t, where a_i is the new subscript. Then, we concatenate them to obtain the assembled results G_i as follows:

$$G_i = F_i^+ \oplus F_{ai}^- \oplus F_i^t \tag{16}$$

Drawing on the data-driven disentanglement methods, the utilization of G_i can be viewed as the feature-level data augmentation. The assembled G_i forces the model to predict $id(i)$ and its fine-grained categories using the clothes-irrelevant features belong to $id(i)$ and the clothes-related features belong to others. Therefore, the discriminativeness of the extracted clothes-irrelevant features will be further highlighted due to its key role in predicting $id(i)$ and its fine-grained categories, and thus its discriminativeness being improved. In addition, G_i does not contribute in the inference, and thus will not bring intra-class variation for CC-ReID. Now, the assembled results G_i have the clothes-irrelevant features belonging to x_i and the clothes-relevant features belonging to x_{ai}. For the former, the identity classifier C_{ID} is still explored to preserve the identity-sensitive discriminative information for x_i. Following Eq. 5, the same loss function $\mathcal{L}_{ID'}$ is optimized based on G_i. For the latter, let the clothes-relevant features of x_i and x_{ai} belong to different identities. Therefore, to improve the discriminativeness of F_i^+ and F_{ai}^t, G_i is used to predict all fine-grained clothes categories belonging to $id(i)$ base on the predictor of clothes C_P. Similarly, to improve the discriminativeness of F_i^-, G_i is used to suppress the prediction of all fine-grained clothes categories that do not belong to $id(i)$. To this end, we propose a novel assembled clothes loss function $\mathcal{L}_{ac}$ to achieve this, which can be formalized as follows:

$$\begin{aligned}\mathcal{L}_{ac} = &-\sum_{i=1}^{N}\sum_{c=1}^{N_U} h(c)\\ &\times log\left(\frac{C_P(G_i|c)}{C_P(G_i|c) + \sum\limits_{id(j)\neq id(i)} C_P(G_i|j)}\right)\end{aligned} \tag{17}$$

$$h(c) = \begin{cases} \dfrac{1}{K} \ , id(c) = id(i) \\ 0 \quad , id(c) \neq id(i) \end{cases} \tag{18}$$

Finally, the total loss function of CI branch is calculated as follows:

$$\mathcal{L}_C = \mathcal{L}_c + \mathcal{L}_{ca} + \alpha\mathcal{L}_{ac} + \gamma\mathcal{L}_{ID'} \tag{19}$$

, where α and γ represent two hyper-parameters for training.

In summary, CI branch uses the assembled features to further improve the discriminativeness of the clothes-irrelevant features and the clothes-relevant features, and thus further facilitates CR branch.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

*4) **Total loss function:*** Finally, the total loss function of CI branch is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_C + \mathcal{L}_R \tag{20}$$

We adopt a two-stage optimization strategy [10] for training, which is detailed in the next section. Through the joint loss function $\mathcal{L}$, DCR-ReID achieves controllable disentanglement for clothes-irrelevant features and clothes-relevant features. For inference, we directly remove the clothes-relevant features to force the model focus on the clothes-irrelevant features.

IV. Experiments

In this section, we conduct extensive experiments to validate the effectiveness of our proposed DCR-ReID method.

A. Datasets and Evaluations

We conduct the validation experiments on three real-world benchmark CC-ReID datasets: LTCC [9], PRCC [8], and CCVID [10]. The summary of these datasets could be found in Tab. I.

TABLE I
The statistics of the datasets utilized for evaluation.

Dataset	Identities	Instances	Cams	Max. cloth changes
LTCC	152	17,119	12	14
PRCC	221	33,698	3	2
CCVID	226	347,833	/	5

PRCC [8] dataset is a popular real-world CC-ReID dataset, which contains 33698 images of 221 identities. PRCC is one of the pioneers of cloth-changing Re-ID and is widely used to evaluate the state-of-the-art algorithms. However, it is limited by the diversity of scenarios (only collected by 3 cameras) and the variation of clothes (only 2 changes of clothes for each person).

LTCC [9] dataset is one of the latest CC-ReID datasets, which contains 17119 images of 152 identities. Although the number of images of LTCC dataset is lower than that of PRCC dataset, the multiplied scenarios (collected by 12 cameras) and variants of clothes (up to 14 changes of clothes for each person at most) make it one of the most challenging CC-ReID dataset in recent years.

CCVID [10] dataset is a large video CC-ReID dataset, which contains 347,833 images with 2,856 sequences from 226 identities and each identity has 2 $\sim$ 5 suits of clothes. The CCVID dataset is closer to real-world Re-ID scenarios. The experiment on the large dataset helps to validate the generalization ability of the model.

Evaluation Protocols. CMC@K (Cumulative Matching Characteristics of top K results) and mAP (mean Average Precision) are two widely used evaluation metrics in person Re-ID. We employ both of them to verify the proposed DCR-ReID. CMC@K is used to evaluate whether the top K re-identification results contain ground truth, while mAP is used to evaluate the overall accuracy of the algorithm for the re-identification results. We conduct experiments under the same setting for fair comparisons with the existing methods.

For PRCC dataset, we conduct the same comparison under Same-Clothes (SC) and Cloth-Changing (CC) settings [8], where SC means that people in the query image and the gallery image are wearing the same clothes and CC means that they are wearing different clothes. For LTCC and CCVID datasets, we compare the proposed DCR-ReID with the state-of-the-art methods under General and CC settings [9], where the General is the combination of SC and CC.

B. Implementation Details

During training, following a similar protocol in [10], we adopt a two-stage optimization strategy to disentangle the clothes-irrelevant features based on a well-trained clothes classifier. It optimizes the loss functions of $\mathcal{L}_c + \mathcal{L}_{ID} + \mathcal{L}_R$ in the first stage and optimizes the full loss function L in the second stage. We set the hyper-parameters α and γ to 0.05 and 1 in PRCC and CCVID datasets, and 0.05 and 0 in LTCC dataset, respectively. We use ResNet-50 [56] with the pre-trained weight on ImageNet [58] as the backbone. We train the network on NVIDIA GTX1080TI GPUs. For PRCC and LTCC datasets, the batch size is set to 64, each batch contains 8 instances of 8 people with different identities. For CCVID datasets, the batch size is set to 8, each batch contains 4 instances of 2 people with different identities. The model is optimized using Adam [59] optimizer. The initial learning rate is set to 3.5×10^{-4}, and the learning rate drops to 10% of the original every 20 epochs for For PRCC and LTCC datasets. For CCVID dataset, it drops every 40 epochs. Following [10], the random cropping, erasing, and flipping are employed as the data augmentation. Input images are resized to 384×192. The channel number of P_i^+, P_i^-, and P_i^t are 1024, 512, and 1024, respectively. Global pooling is the stack of max-pooling [60] and average-pooling.

During inferring, we directly remove F_i^- in PI branch to calculate the distance between the query image and the images in the gallery.

C. Determination of the Hyper-parameters

The proposed DCR-ReID introduces two hyper-parameters for training, α and γ. α is used to determine the loss weight of the assembled features for the clothes classification, while γ is used to determine the loss weight of the assembled features for the identity classification. The hyper-parameters α=0.05 with γ=1 are adopted on PRCC and CCVID datasets, while α=0.05 with γ=0 are adopted on LTCC dataset, respectively.

D. Comparison with State-of-the-art Methods

First of all, we compare the proposed DCR-ReID with multiple state-of-the-art Re-ID methods, including HACNN [51], PCB [29], IANet [52], SPT+ASE [8], GI-ReID [53], CESD [9], RCSANet [54], 3DSL [55] FSMA [16], CAL [10] on LTCC and PRCC datasets. Among them, HACNN [51], PCB [29], and IANet [52] are general Re-ID methods, while the others are CC-ReID methods. As shown in Tab. II, the proposed DCR-ReID achieves the state-of-the-art performance on LTCC and PRCC datasets. Specifically, the CMC@1/mAP performance of the proposed DCR-ReID method is 41.1%/20.4%

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON LTCC AND PRCC DATASETS. (THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE UNDERLINED.)

method	clothes label	LTCC				PRCC			
		General		CC		SC		CC	
		CMC@1	mAP	CMC@1	mAP	CMC@1	mAP	CMC@1	mAP
HACNN[51]		60.2	26.7	21.6	9.3	82.5	-	21.8	-
PCB[29]		65.1	30.6	23.5	10.0	99.8	97.0	41.8	38.7
IANet[52]		63.7	31.0	25.0	12.6	99.4	98.3	46.3	45.9
SPT+ASE[8]		-	-	-	-	64.2	-	34.4	-
GI-ReID[53]		63.2	29.4	23.7	10.4	80.0	-	33.3	-
CESD[9]	√	71.4	34.3	26.2	12.4	-	-	-	-
RCSANet[54]		-	-	-	-	100	97.2	50.2	48.6
3DSL[55]	√	-	-	31.2	14.8	-	-	51.3	-
FSMA[16]		73.2	35.4	38.5	16.2	98.8	-	54.5	-
CAL[10]	√	74.2	40.8	40.1	18.0	**100**	**99.8**	55.2	55.8
DCR-ReID	√	**76.1**	**42.3**	**41.1**	**20.4**	**100**	99.7	**57.2**	**57.4**

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON CCVID DATASETS. (THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE UNDERLINED.)

method	CCVID			
	General		CC	
	CMC@1	mAP	CMC@1	mAP
Baseline [56]	78.3	75.4	77.3	73.9
Triplet Loss [57]	81.5	78.1	81.1	77.0
CAL [10]	82.6	81.3	81.7	79.6
DCR-ReID	**84.7**	**82.7**	**83.6**	**81.4**

and 57.2%/57.4% on both the LTCC and PRCC datasets under the CC setting. Compared with the existing state-of-the-art method [10], our DCR-ReID outperforms it by 1.0% and 2.4% for CMC@1 and mAP on LTCC.

To validate the generalization ability of the proposed DCR-ReID, we compared the proposed DCR-ReID with state-of-the-art methods on a larger CC-ReID dataset, called CCVID, including: Baseline [56], which use ResNet-50 as the backbone network and the identification loss for training, Triplet Loss [57], and CAL [10]. As shown in Tab. III, the proposed DCR-ReID outperforms the state-of-the-art methods on the CCVID dataset. Specifically, DCR-ReID achieves 83.6%/81.4% and 84.7%/82.7% on the CMC@1/mAP under the General setting and the CC setting. Particularly, DCR-ReID leads the state-of-the-art method [10] by 1.9% and 1.8% on the CMC@1/mAP under the CC setting.

The above experimental results support that the proposed DCR-ReID can perform controllable disentanglement to obtain the clothes-irrelevant features and the clothes-relevant features, and can also improve the discriminativeness of these features effectively. In addition, although we remove the clothes-relevant features for inference, under the General setting on LTCC, DCR-ReID also achieves improvement (1.9% for CMC@1 and 1.5% for mAP on LTCC). The reason is that the clothes-irrelevant features extracted by DCR-ReID is more robust than the clothes-relevant features. Therefore, DCR-ReID is outstanding under the both settings. Moreover, the quantitative results in Tab. II are also verified by the visualization results in Fig. 6. The visualization results show that the proposed DCR-ReID can re-identify the person wearing different clothes more accurately when the style and colour of the clothes change drastically. The above results illustrates that the clothing-relevant feature disentangled by DCR-ReID can help re-identify people wearing the same or different clothes better.

E. Ablation Study

Hyper-parameter sensitivity experiments. The proposed DCR-ReID method introduces two vital hyper-parameters for training, α and γ, which are directly related to the performance of DCR-ReID. Therefore, first of all, to determine the optimal combination of α and γ, we design the hyper-parameter sensitivity experiments on LTCC and PRCC datasets. First, we initially tried the settings of α and γ under 0.01, 0.1 and 1 for the preliminary attempt. The results are shown in Fig. 7. Next, we fix γ with 1.0 and 0.0 to further explore the weight of α on PRCC and LTCC datasets, respectively. The further experimental results are shown in Tab. IV. This demonstrates that in scenarios with limited variants of clothes such as PRCC, the limited fine-grained clothes categories are difficult to facilitate the association with identity categories, and thereby weaken the identity discriminativeness of the assembled features. Therefore, in scenarios with limited variants of clothes, the identity discriminativeness needs to be promoted by setting γ=1, while in scenarios with rich variants of clothes, γ should be smaller to facilitate the model to focus on the disentanglement. In summary, the hyper-parameters α=0.05 with γ=1 and α=0.05 with γ=0 are adopted on PRCC and LTCC datasets, respectively.

To verify the uneven divided channel number of P_i^+, P_i^-, and P_i^t that are 1024, 512, and 1024, we conduct experiments to compare the effects of even division and uneven division setting on PRCC and LTCC datasets. For even division, the channel numbers of P_i^+, P_i^-, and P_i^t are all 1024. We use a convolutional layer with a kernel size of 1×1 to achieve the transformation of the channel numbers. As shown in Fig. 10(a) and 10(b), uneven division achieves the state-of-the-art performance on both LTCC and PRCC datasets from CMC@1

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

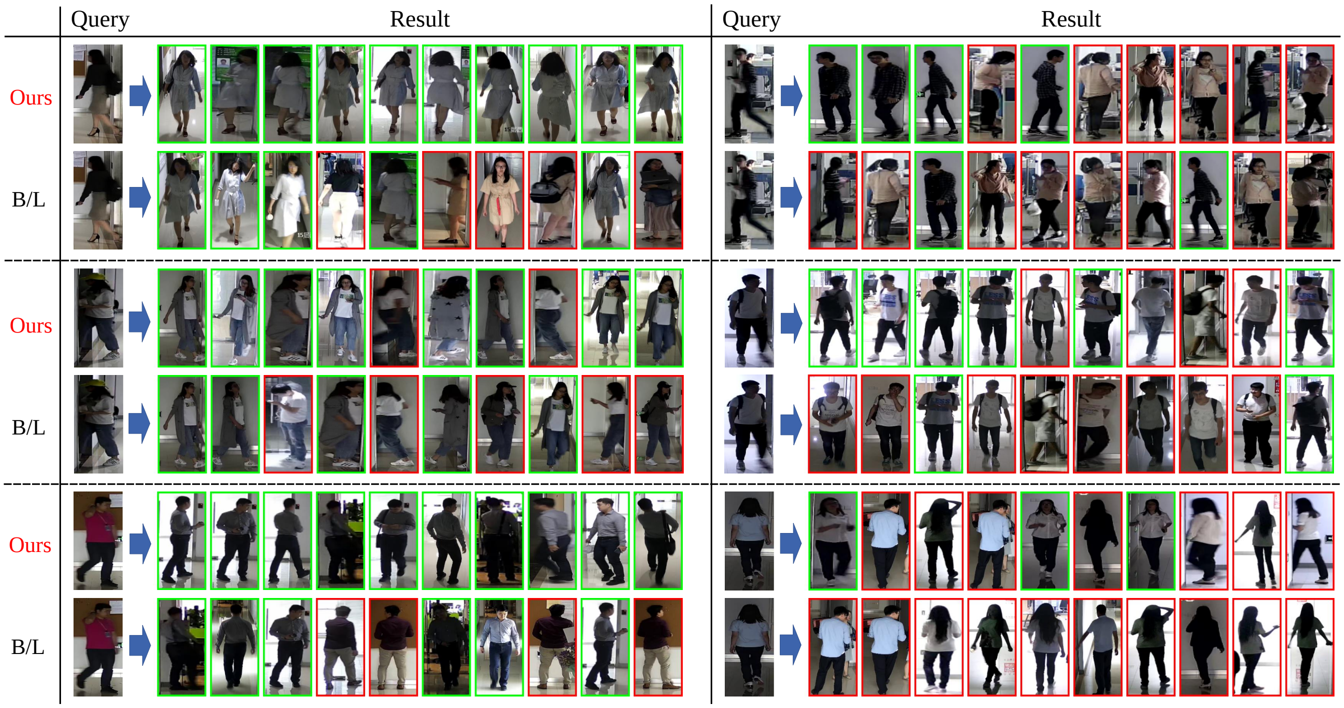


Fig. 6. Visualization of the comparison results of our method (Ours) and the baseline algorithm (B/L) [10]. The visualization results show that the proposed DCR-ReID can re-identify more correct instances in the gallery set compared with the baseline algorithm.

to CMC@20. This demonstrates that the disentangled features (P_i^+, P_i^-, and P_i^t) have uneven contributions to CC-ReID. Specifically, the clothes-relevant feature is harmful to CC-ReID and thus its contribution is low and has fewer channel numbers. However, the clothes-irrelevant feature and the contour feature are equally significant to CC-ReID. Therefore, the channel number of P_i^+, P_i^-, and P_i^t are set to 1024, 512, and 1024, respectively.

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED DCR-REID WITH FIXED γ ON LTCC AND PRCC DATASETS. (THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE UNDERLINED.)

hyper-parameters		LTCC(CC)	
γ	α	CMC@1	mAP
0	0.01	40.1	19.4
0	0.02	40.3	20.1
0	**0.05**	**41.1**	**20.4**
0	0.1	40.8	20.0
hyper-parameters		PRCC(CC)	
γ	α	CMC@1	mAP
1	0.01	56.5	57.3
1	0.02	57.1	57.3
1	**0.05**	57.2	**57.4**
1	0.1	**57.4**	57.0

Effectiveness of DAD and CRD. Recall our proposed DCR-ReID method, there are two important modules, DAD and CRD. To verify the effectiveness of the proposed DAD and CRD. We conduct ablation studies on both LTCC and PRCC datasets. The baseline method is [10]. As shown in Tab. V, when we only use DAD for disentanglement, the Re-ID accuracy is improved in PRCC dataset under the CC setting, where CMC@1/mAP improved from 53.4% and 54.0% to 55.7% and 55.8%. However, it is lower than the state-of-the-art baseline algorithm on LTCC under the General setting, where CMC@1/mAP reduced from 75.1% and 41.0% to 72.4% and 39.9%, as well as under the CC setting. When we only use CRD, the accuracy improves significantly on both LTCC and PRCC datasets. This illustrates that CRD can effectively disentangle clothes-irrelevant features, while using DAD alone will reduce the accuracy when dealing with more complex scenarios and variants of clothes. It happens because DAD cannot ensure that the clothes-relevant features correspond to the region of clothes. Thus, when removing them for inference, it may bring mistakes by removing discriminative clothes-irrelevant features. Moreover, using DAD and CRD independently does not bring about an improvement for CMC@1 on LTCC, where CMC@1 is reduced by 2.7% and 0.5%, respectively. This is because DAD and CRD can each cover parts of the correct results on rank-1 respectively. However, when both CRD and DAD are used, the accuracy is further improved, which demonstrates that the DAD can benefit the CRD to further improving the overall performance. When the feature segments and the spatial component regions are consistent, CRD can further promote the disentanglement of clothes-irrelevant features and achieve the state-of-the-art performance.

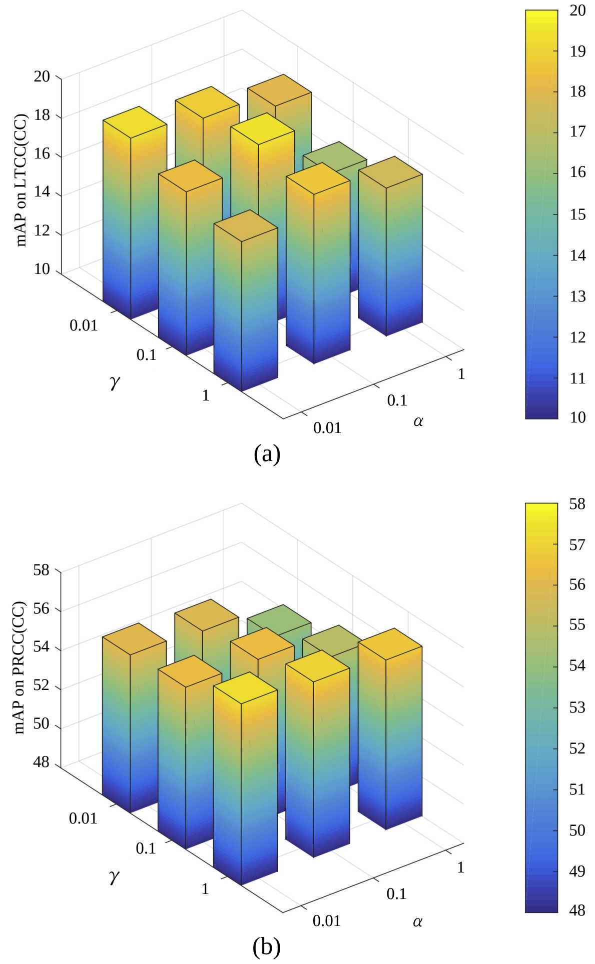
Influence of the clothes-relevant features. As mentioned before, when the clothes-relevant features are not correctly corresponded to the regions of clothes, it may severely deteriorate the final performance when removing them for inference. To verify whether DCR-ReID can achieve this correspondence,

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

TABLE V
ABLATION STUDIES OF DAD AND CRD ON LTCC AND PRCC DATASETS. (THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE UNDERLINED.)

Baseline	DAD	CRD	LTCC				PRCC			
			General		CC		SC		CC	
			CMC@1	mAP	CMC@1	mAP	CMC@1	mAP	CMC@1	mAP
√			75.1	41.0	40.1	18.0	**100**	99.7	53.4	54.0
√	√		72.4	39.9	38.0	18.1	**100**	**99.8**	55.7	55.8
√		√	74.6	42.1	**41.1**	20.0	**100**	**99.8**	56.1	56.3
√	√	√	**76.1**	**42.3**	**41.1**	**20.4**	**100**	99.7	**57.2**	**57.4**



(a)

(b)

Fig. 7. Performance comparison of the proposed DCR-ReID with different hyper-parameters combinations of α and γ. (a) is the performance on LTCC. (b) is the performance on PRCC.

we further experiment and visualize the results on LTCC and PRCC datasets for validation.

As shown in Tab. VI, when we remove the cloths-relevant features, DCR-ReID achieves better Re-ID accuracy on LTCC (the CMC@1/mAP performance of the proposed DCR-ReID method improved by 1.6%/0.3% on both LTCC). However, on PRCC, the CMC@1/mAP performance of w/ and w/o CR is similar. The reason is that the cloth changes are limited on PRCC (only 2 clothes for each person), thus removing the clothes-relevant features are not critical. This illustrates that removing the clothes-relevant features is more suitable for scenarios with variations clothes in CC-ReID. Nonetheless, for scenarios with less changes of clothes, removing the clothes-relevant features does not impair the performance as we analyzed in Tab. II. In general, removing clothes-relevant features can better adapt to the Re-ID under different clothing changes.

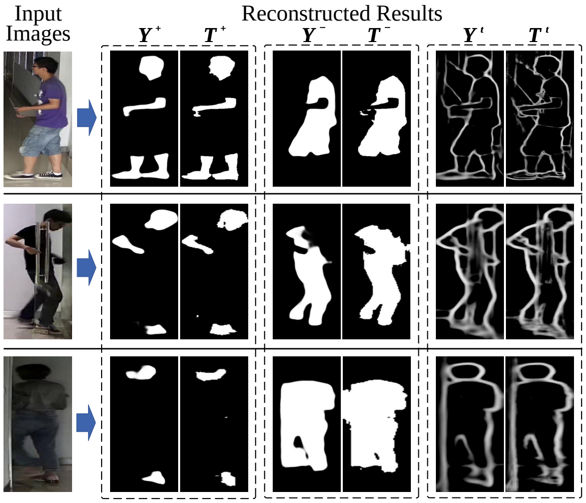


Fig. 8. Visualization of the reconstruction results output by CRD. Y^+, Y^-, and Y^t represents the reconstruction results of the clothes-irrelevant features, the clothes-relevant features, and the contour features, respectively. T^+, T^-, and T^t are their corresponding ground truth. The visualization results show that the extracted clothes-irrelevant features and the clothes-relevant features are correspond to the corresponding human component regions.

As shown in Fig. 8, we visualize the reconstruction results output by CRD. The visualization results show that the clothes-relevant feature corresponds to the region of clothes, while the clothes-irrelevant feature corresponds to the remaining region of people. In addition, the contour features are also completely reconstructed. In summary, DCR-ReID can exactly disentangle clothes-irrelevant features and maintain the correspondence in inference. In the following sections, we further investigate the contribution of different reconstruction branches in the proposed DCR-ReID model.

Influence of the reconstructing branches. To illustrate the contribution of the reconstruction branches, we conduct various experiments on LTCC by gradually removing some of the reconstruction branches to evaluate the influence. As shown in Tab. VII, the baseline method is DCR-ReID without CRD (”Ours w/o I,R,C”). When we only use the cloths-irrelevant reconstruction branch (”Ours w/o R,C”), it improves mAP

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

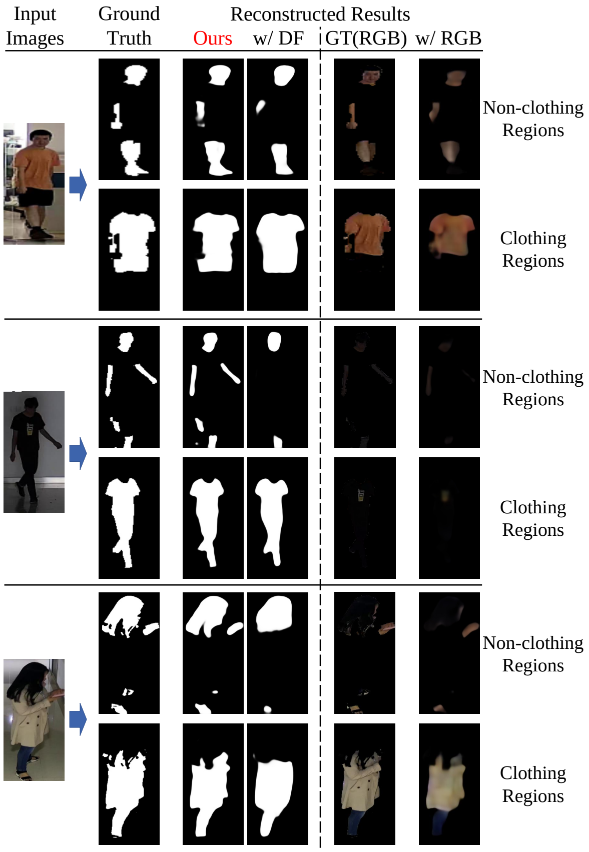
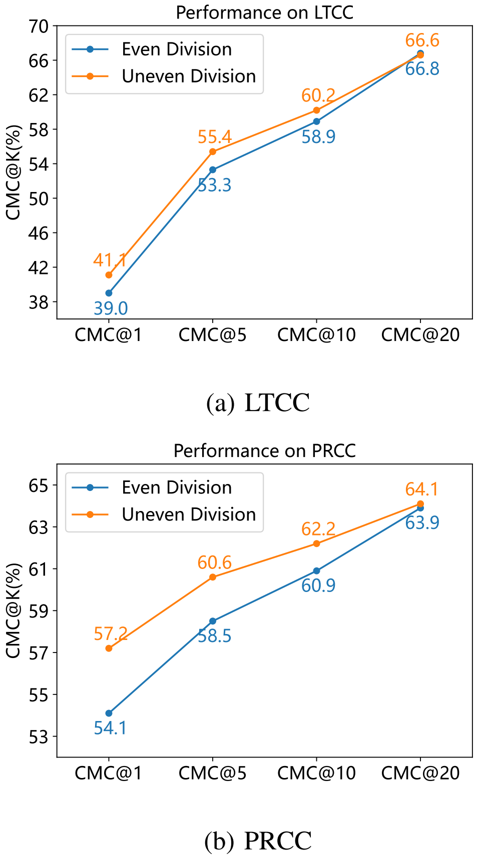


Fig. 9. Visualization of the comparison results of our reconstruction (Ours) with the reconstruction using the deep features (w/ DF) and the construction using the RGB images (w/ RGB) output by CRD.

TABLE VI
COMPARISON WITH USING CLOTHES-RELEVANT FEATURES (CR) IN INFERENCE ON LTCC AND PRCC DATASETS. (THE BEST RESULTS ARE BOLDED.)

Dataset / Evaluation	LTCC(CC)		PRCC(CC)	
	w/o CR	w/ CR	w/o CR	w/ CR
mAP	**20.4**	20.1	**57.4**	**57.4**
CMC@1	**41.1**	39.5	57.2	**57.3**
CMC@5	**54.3**	53.8	**60.6**	60.5
CMC@10	**59.4**	59.2	**62.2**	62.1



(a) LTCC

(b) PRCC

Fig. 10. Impact of even/uneven division. (a) is the result on LTCC. (b) is the result on PRCC.

under the CC setting by 0.5%. When we further added clothes-relevant reconstruction and contour reconstruction (”Ours w/o C” and (”Ours”)), the mAP was improved by 1.3% and 2.3%, respectively. The above experimental results illustrate that simultaneously disentangling the clothes-irrelevant features and the clothes-relevant features can better utilize the complementary relationship between the two features to achieve the improvement. Therefore, learning to disentangle both of the features is more effective than learning a mixed clothes-irrelevant features. In addition, using the contour reconstruction can further improve the accuracy. This is because the clothes-relevant features and the clothes-irrelevant features can only model the local component regions, lacking the perception of the global shape. Therefore, fusing the contour reconstruction brings robust global shape information and thereby further improve the accuracy. In summary, the above experiments verify the effectiveness of the three branches for disentanglement.

TABLE VII
COMPARISON WITH DIFFERENT RECONSTRUCTION BRANCHES. (THE BEST RESULTS ARE BOLDED.)

method	LTCC			
	General		CC	
	CMC@1	mAP	CMC@1	mAP
Ours	**76.1**(+3.7)	**42.3**(+2.4)	**41.1**(+3.1)	**20.4**(+2.3)
Ours w/o C	74.8(+2.4)	41.0(+1.1)	40.1(+2.1)	19.4(+1.3)
Ours w/o R,C	75.1(+2.7)	40.6(+0.7)	38.3(+0.3)	18.6(+0.5)
Ours w/o I,R,C	72.4	39.9	38.0	18.1

TABLE VIII
COMPARISON WITH DIFFERENT RECONSTRUCTION METHODS. (THE BEST RESULTS ARE BOLDED.)

method	LTCC			
	General		CC	
	CMC@1	mAP	CMC@1	mAP
Ours	**76.1**	**42.3**	**41.1**	**20.4**
w/ RGB	75.3(-0.8)	40.9(-1.4)	40.1(-1.0)	19.2(-1.2)
w/ DF	74.6(-1.5)	39.2(-3.1)	39.3(-1.8)	17.6(-2.8)

Influence of difference reconstruction targets. As mentioned before, the proposed DCR-ReID uses convolutional features to reconstruct the response map for disentanglement, instead of directly using the deep feature vectors extracted by

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

the backbone network. Therefore, it is necessary to compare the two kinds of disentanglement to further illustrate the innovation of DCR-ReID. Thus, we further experiment and visualize the results on LTCC dataset for validation.

As shown in Tab. VIII, when we use the deep features extracted by the backbone network for reconstruction (”w/ DF”), the accuracy dropped by 3.1% and 2.8% for mAP under General and CC settings, respectively. When we use the RGB image as the ground truth (”w/ RGB”), the accuracy dropped by 2.1% and 1.9% for mAP. We further conduct visualization experiments for comparison, as shown in Fig. 9. The results show that when we use the deep features for reconstruction, the reconstructed images do not accurately correspond to the clothing and non-clothing regions. Moreover, when we use the RGB image as the ground truth, although the reconstructed images can reveal the colour information of the clothes, it does not reflect discriminative features well and the reconstructed regions are seriously incomplete. It reveals that the reconstruction using convolutional features can achieve better accuracy than the deep features-based reconstruction. The reason is that the convolutional features can preserve more information on spatial than the deep features, thus it is more suitable to be used for reconstruction. In addition, reconstruction RGB images lead to poorer performance because the extracted high-level deep features usually contain less low-level detail information which is important for RGB image reconstruction. Fortunately, the proposed CRD and DAD can readily tackle this issue since it is simplified as the collaboration of CRD and DAD, which only reconstructs the component regions of the clothes-irrelevant features and the clothes-relevant features in CRD, and then improves the discriminativeness of these features in DAD. CRD and DAD promote each other and finally achieve the state-of-the-art result.

V. Conclusion

In this paper, to tackle the challenge of changes of clothes in the long-term Re-ID, we propose a novel method, called Deep Component Reconstruction (DCR-ReID) for CC-ReID. Different from existing reconstruction-based methods, we propose a Component Reconstruction Disentanglement (CRD) module to disentangle the clothes-irrelevant features and the clothes-relevant features in a controllable manner. We also propose a Deep Assembled Disentanglement (DAD) module to further improve the discriminativeness of the disentangled features in CRD, avoiding the spoiling of the most discriminative clothes-irrelevant information. For inference, we directly remove the clothes-relevant features for controllable disentanglement. Extensive experiments on three real-world benchmark CC-ReID datasets demonstrate the effectiveness of the proposed DCR-ReID.

In the future, DCR-ReID still has some room for improvement. First, the reconstructed contour map in CR branch will inevitably brings tiny noises, such as contours of the background. It is necessary to extract more accurate contour maps of the person for better reconstruction of the contour features. Second, Existing Re-ID methods mainly utilize images to extract discriminative features. However, multi-modal data contains more complementary discriminative information. Therefore, we intend to explore effective methods to fuse these data, and hope to improve the ability to disentangle and extract discriminative features in more complex scenarios.

References

[1] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, “Joint learning of single-image and cross-image representations for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1288–1296.

[2] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, “Part-aligned bilinear representations for person re-identification,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 402–419.

[3] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[4] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *European conference on computer vision*. Springer, 2016, pp. 17–35.

[5] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[6] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2138–2147.

[7] Z. Yu, Y. Zhao, B. Hong, Z. Jin, J. Huang, D. Cai, X. He, and X.-S. Hua, “Apparel-invariant feature learning for person re-identification,” *IEEE Transactions on Multimedia*, 2021.

[8] Q. Yang, A. Wu, and W.-S. Zheng, “Person re-identification by contour sketch under moderate clothing change,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 2029–2046, 2019.

[9] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, “Long-term cloth-changing person re-identification,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[10] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen, “Clothes-changing person re-identification with rgb modality only,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1060–1069.

[11] F. Wan, Y. Wu, X. Qian, Y. Chen, and Y. Fu, “When person re-identification meets changing clothes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 830–831.

[12] J. Xue, Z. Meng, K. Katipally, H. Wang, and K. van Zon, “Clothing change aware person identification,” in *Proceedings of the IEEE Conference on Computer Vision*

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

and Pattern Recognition Workshops, 2018, pp. 2112–2120.

[13] X. Shu, G. Li, X. Wang, W. Ruan, and Q. Tian, "Semantic-guided pixel sampling for cloth-changing person re-identification," *IEEE Signal Processing Letters*, vol. 28, pp. 1365–1369, 2021.

[14] X. Jia, X. Zhong, M. Ye, W. Liu, W. Huang, and S. Zhao, "Patching your clothes: Semantic-aware learning for cloth-changed person re-identification," in *International Conference on Multimedia Modeling*. Springer, 2022, pp. 121–133.

[15] L. WANG, Y. ZHANG, T. LU, W. FANG, and Y. WANG, "Multi feature fusion attention learning for clothing-changing person re-identification," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, p. 2021EAL2097, 2022.

[16] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng, "Fine-grained shape-appearance mutual learning for cloth-changing person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 513–10 522.

[17] Y.-J. Li, Z. Luo, X. Weng, and K. M. Kitani, "Learning shape representations for clothing variations in person re-identification," *arXiv preprint arXiv:2003.07340*, 2020.

[18] W. Xu, H. Liu, W. Shi, Z. Miao, Z. Lu, and F. Chen, "Adversarial feature disentanglement for long-term person re-identification," in *IJCAI*, 2021, pp. 1201–1207.

[19] J. Chen, W.-S. Zheng, Q. Yang, J. Meng, R. Hong, and Q. Tian, "Deep shape-aware person re-identification for overcoming moderate clothing changes," *IEEE Transactions on Multimedia*, 2021.

[20] M. Tu, K. Zhu, H. Guo, Q. Miao, C. Zhao, G. Zhu, H. Qiao, G. Huang, M. Tang, and J. Wang, "Multi-granularity mutual learning network for object re-identification," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[21] S. Yu, D. Chen, R. Zhao, H. Chen, and Y. Qiao, "Neighbourhood-guided feature reconstruction for occluded person re-identification," *arXiv preprint arXiv:2105.07345*, 2021.

[22] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7073–7082.

[23] K. Kansal and A. V. Subramanyam, "Hdrnet: Person re-identification using hybrid sampling in deep reconstruction network," *IEEE Access*, vol. 7, pp. 40 856–40 865, 2019.

[24] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.

[25] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3219–3228.

[26] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European conference on computer vision*. Springer, 2016, pp. 135–153.

[27] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907.

[28] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1367–1376.

[29] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.

[30] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and A. Bouridane, "Gait recognition for person re-identification," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3653–3672, 2021.

[31] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1366–1377, 2018.

[32] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 475–491.

[33] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2109–2118.

[34] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5771–5780.

[35] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, and Z. Zhang, "Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3459–3471, 2019.

[36] X. Shu, X. Wang, X. Zang, S. Zhang, Y. Chen, G. Li, and Q. Tian, "Large-scale spatio-temporal person re-identification: Algorithms and benchmark," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[37] G.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z.-G. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 144–12 151.

[38] E. Yaghoubi, D. Borza, B. Degardin, and H. Proença, "You look so different! haven't i seen you a long time ago?" *Image and Vision Computing*, vol. 115, p. 104288, 2021.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

"Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[41] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[42] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *arXiv preprint arXiv:2203.12602*, 2022.

[43] X. Zhang, Y. Yan, J.-H. Xue, Y. Hua, and H. Wang, "Semantic-aware occlusion-robust network for occluded person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2764–2778, 2020.

[44] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 257–10 266.

[45] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1418–1430, 2021.

[46] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (Csur)*, vol. 40, no. 2, pp. 1–60, 2008.

[47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[48] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[49] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1939–1946, 2019.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[51] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[52] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9317–9326.

[53] X. Jin, T. He, K. Zheng, Z. Yin, X. Shen, Z. Huang, R. Feng, J. Huang, Z. Chen, and X.-S. Hua, "Cloth-changing person re-identification from a single image with gait prediction and regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 278–14 287.

[54] Y. Huang, Q. Wu, J. Xu, Y. Zhong, and Z. Zhang, "Clothing status awareness for long-term person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 895–11 904.

[55] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, and W.-S. Zheng, "Learning 3d shape feature for texture-insensitive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8146–8155.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

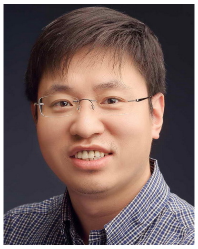
Zhenyu Cui received the B.S. degree in computer science and technology from China University of Petroleum (East China), Tsingdao, China, in 2018 and the M.S. degree in computer science from University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include computer vision and deep learning.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3241988

Jiahuan Zhou received his B.E. (2013) from Tsinghua University, the Ph.D. degree (2018) in the Department of Electrical Engineering & Computer Science, Northwestern University. During summer 2018, he was a research intern with Microsoft Research, Redmond, Washington. From 2019 to 2022, he was a Postdoctoral Fellow and Research Assistant Professor in Northwestern University. Currently, he is a Tenure-Track Assistant Professor with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include computer vision, deep learning, and machine learning. He has authored more than 15 papers in international journals and conferences including IEEE T-PAMI, IEEE TIP, CVPR, ICCV, ECCV and so on. He serves as an area chair for CVPR'2023, ICME'2020,2021,2023, ICPR'2022, a regular reviewer member for a number of journals and conferences, e.g., T-PAMI, IJCV, TIP, CVPR, ICCV, ECCV, NeurIPS, ICML, and so on.

Yuxin Peng (Senior Member, IEEE) received the Ph.D. degree in computer applied technology from Peking University, Beijing, China, in 2003. He is currently the Boya Distinguished Professor with the Wangxuan Institute of Computer Technology, Peking University. He has authored over 170 papers, including more than 80 papers in the top-tier journals and conference proceedings. He has submitted 48 patent applications and been granted 37 of them. His current research interests mainly include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision. He led his team to win the First Place in video semantic search evaluation of TRECVID ten times in the recent years. He won the First Prize of the Beijing Technological Invention Award in 2016 (ranking first) and the First Prize of the Scientific and Technological Progress Award of Chinese Institute of Electronics in 2020 (ranking first). He was a recipient of the National Science Fund for Distinguished Young Scholars of China in 2019, and the best paper award at MMM 2019 and NCIG 2018. He serves as the associate editor of IEEE TMM, TCSVT, etc.



Shiliang Zhang (Senior Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences. He was a Post-Doctoral Scientist with NEC Laboratories America and a Post-Doctoral Research Fellow with The University of Texas at San Antonio. He is currently an Associate Professor with Tenure with the Department of Computer Science, School of Electronic Engineering and Computer Science, Peking University. His research interests include large-scale image retrieval and computer vision. He was a recipient of the Outstanding Doctoral Dissertation Awards from the Chinese Academy of Sciences and Chinese Computer Federation, the President Scholarship from the Chinese Academy of Sciences, the NEC Laboratories America Spot Recognition Award, the Nvidia Pioneering Research Award, and the Microsoft Research Fellowship.

Yaowei Wang (Member, IEEE) received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences in 2005. He served as a Distinguished Professor with the National Engineering Laboratory for Video Technology Shenzhen (NELVT), Peking University Shenzhen Graduate School, in 2019. He is currently a professor with the Peng Cheng Laboratory, Shenzhen, China. He is the author or coauthor of more than 120 technical articles in international journals and conferences, including TOMM, ACM MM, IEEE TIP, CVPR, ICCV, IJCAI, and AAAI. His current research interests include multimedia content analysis and understanding, machine learning and computer vision. He servers as the chair of the IEEE Digital Retina Systems Working Group and a member of IEEE, CIE, CCF, CSIG. He was the recipient of the second prize of the National Technology Invention in 2017 and the first prize of the CIE Technology Invention in 2015.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Peking University. Downloaded on May 09,2023 at 09:16:42 UTC from IEEE Xplore. Restrictions apply.