Color-Unrelated Head-Shoulder Networks for Fine-Grained Person Re-identification

BOQIANG XU, University of Chinese Academy of Sciences, China
JIAN LIANG, CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences, China
LINGXIAO HE, AI Research of JD, China
JINLIN WU, MAIS, Institute of Automation, Chinese Academy of Sciences; Centre for Artificial Intelligence and Robotics, HKISI, CAS, China
CHAO FAN, Chengdu Discaray Technology Co., Ltd., China
ZHENAN SUN, CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences, China

Person **re-identification (re-id)** attempts to match pedestrian images with the same identity across non-overlapping cameras. Existing methods usually study person re-id by learning discriminative features based on the clothing attributes (e.g., color, texture). However, the clothing appearance is not sufficient to distinguish different persons especially when they are in similar clothes, which is known as the **fine-grained (FG)** person re-id problem. By contrast, this paper proposes to exploit the color-unrelated feature along with the head-shoulder feature for FG person re-id. Specifically, a **color-unrelated head-shoulder network (CUHS)** is developed, which is featured in three aspects: (1) It consists of a lightweight head-shoulder segmentation layer for localizing the head-shoulder region and learning the corresponding feature. (2) It exploits **instance normalization (IN)** for learning color-unrelated features. (3) As IN inevitably reduces inter-class differences, we propose to explore richer visual cues for IN by an attention exploration mechanism to ensure high discrimination. We evaluate our model on the FG-reID, Market1501, and DukeMTMC-reID datasets, and the results show that CUHS surpasses previous methods on both the FG and conventional person re-id problems.

CCS Concepts: • **Computing methodologies** → **Object identification**; **Image representations;**

Additional Key Words and Phrases: Person re-identification, fine-grained matching, visual surveillance

ACM Reference format:
Boqiang Xu, Jian Liang, Lingxiao He, Jinlin Wu, Chao Fan, and Zhenan Sun. 2023. Color-Unrelated Head-Shoulder Networks for Fine-Grained Person Re-identification. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6, Article 210 (July 2023), 21 pages.
https://doi.org/10.1145/3599730

This work is supported by the National Natural Science Foundation of China (Grant No. 434 U1836217), National Natural Science Foundation of China (Grant No. 62276256) and Beijing Nova 435 Program under Grant Z211100002121108.
Authors' addresses: B. Xu, University of Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Haidian District, Beijing, China; email: boqiang.xu@cripac.ia.ac.cn; J. Liang (corresponding author), CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Haidian District, Beijing, China; L. He, AI Research of JD, No. 95 Zhongguancun East Road, Haidian District, Beijing, China; email: helingxiao3@jd.com; J. Wu, MAIS, Institute of Automation, Chinese Academy of Sciences; Centre for Artificial Intelligence and Robotics, HKISI, CAS, No. 95 Zhongguancun East Road, Haidian District, Beijing, China; C. Fan, Chengdu Discaray Technology Co., Ltd., No. 95 Zhongguancun East Road, Haidian District, Beijing, China; Z. Sun, CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Haidian District, Beijing, China.

1551-6857/2023/07-ART210 $15.00
https://doi.org/10.1145/3599730

1 INTRODUCTION

Person **re-identification (re-id)** attempts to match pedestrian images with the same id across non-overlapping cameras. Most of the existing re-id methods [3, 6, 13, 18, 20, 25, 59, 61, 64, 77] assume that pedestrians wear various clothes and appearance features are sufficient for person re-id. However, this can be problematic when people wear similar clothes. For example, in some monitoring scenarios such as factories, schools, and banks, people always wear uniforms or similar clothes. As shown in Figure 1, in these cases pedestrians are difficult to be distinguished based on the clothing appearance. Conducting person re-id when people dress in similar clothes is called the ***fine-grained person re-identification (FG person re-id)*** [67].

Recently, various methods have been developed to utilize local features [50, 52], pose information [71, 74], or attention maps [5, 72] for the conventional person re-id problem. All these methods mainly rely on the clothing attributes (e.g., style, color, textures) for feature matching. However, people wear similar clothes, making these re-id methods ineffective on FG person re-id. Although [67] proposed a new network to solve the video-based FG person re-id problem by exploiting dynamic pose features in the video sequences, these features are not available for the image-based FG person re-id problem.

In this paper, different from previous approaches, we propose to leverage the color-unrelated representation and head-shoulder information to enhance the discrimination of re-id features. As shown in Figure 2, the head-shoulder region provides a wealth of clues for re-id, such as the appearance, gender, and haircut, which illustrates the potential of the head-shoulder information to solve the FG person re-id problem. In addition, since color-related features (i.e., clothing appearance) are not sufficient for retrieval on the FG person re-id, we tend to exploit the color-unrelated information for a supplement. **Instance Normalization (IN)** [8] normalizes features with the statistics of individual instances, and the instance-specific contrast information could be filtered out from the content. Inspired by these merits, we tend to leverage IN for color-unrelated features extraction. Specifically, as shown in Figure 3, we design a three-stream network named **color-unrelated head-shoulder network (CUHS)**, which consists of a head-shoulder stream, a BN stream, and an IN stream. The head-shoulder stream consists of a lightweight head-shoulder segmentation layer for localizing the head-shoulder region and a **head-shoulder attention network (HAN)** for learning the corresponding feature. The BN stream leverages **Batch Normalization (BN)** [17] to learn the color-related representation. The IN stream exploits IN [8] to learn the color-unrelated representation. However, IN inevitably results in the loss of some discriminative features [21]. To alleviate this problem, we further propose to explore richer visual cues for IN by an attention exploration mechanism to ensure the high discriminative ability.

Since there does not exist a benchmark for studying the image-based FG person re-id problem, we introduce a new dataset called FG-reID, which contains 36,282 images of 2,784 identities. The FG-reID dataset is divided into a black group and a white group by the different colored clothes, and each group contains images of 655 and 586 identities, respectively. We conducted the experiments on the FG-reID, Market-1501 [75] and DukeMTMC-reID [39] datasets. The results demonstrate that our approach is effective on both fine-grained and conventional re-id problems.

The main contributions of this paper can be summarized as follows:

- We introduce the first dataset, namely FG-reID, for studying the image-based FG person re-id problem. FG-reID dataset is divided into a black group and a white group according to the different colored clothes and contains 36,282 images of 2,784 identities in total.
- We propose the color-unrelated head-shoulder network (CUHS) for FG person re-id problem. CUHS leverages the head-shoulder information and color-unrelated feature for a more discriminative representation.

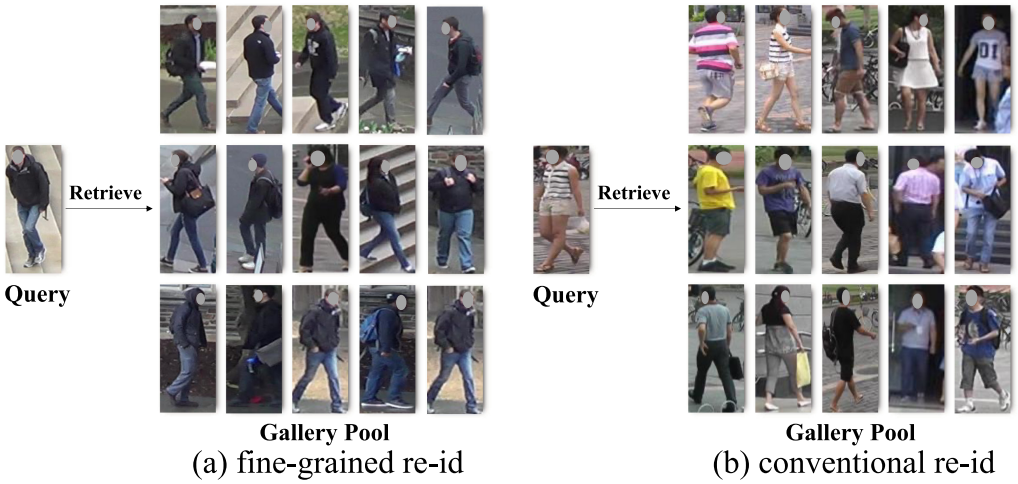

Fig. 1. The illustration of the fine-grained person re-identification (FG person re-id) problem. The greatest challenge on FG person re-id is that people in similar clothes are difficult to be distinguished by the clothing appearance.

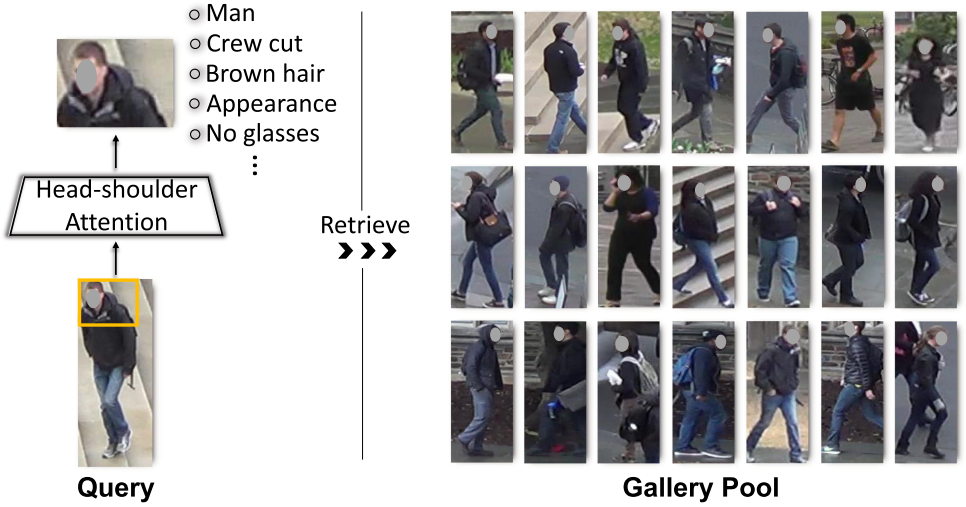

Fig. 2. Head-shoulder region obtains valuable cues for fine-grained person re-identification, such as gender, haircut, appearance, and glasses.

- We propose an attention exploration mechanism for alleviating the reduction of the inter-class differences in the IN processing.
- CUHS surpasses previous methods on both fine-grained and conventional person re-id problems.

This paper is built upon our preliminary work [60] with the following improvements. Firstly, besides the head-shoulder information utilized in [60], we further propose to exploit the color-unrelated features to assist fine-grained person re-id. Secondly, we consider instance normalization and attention mechanism and develop an auxiliary stream to fully exploit color-unrelated features. Thirdly, our method does not need additional labels as in [60], which requires labels for

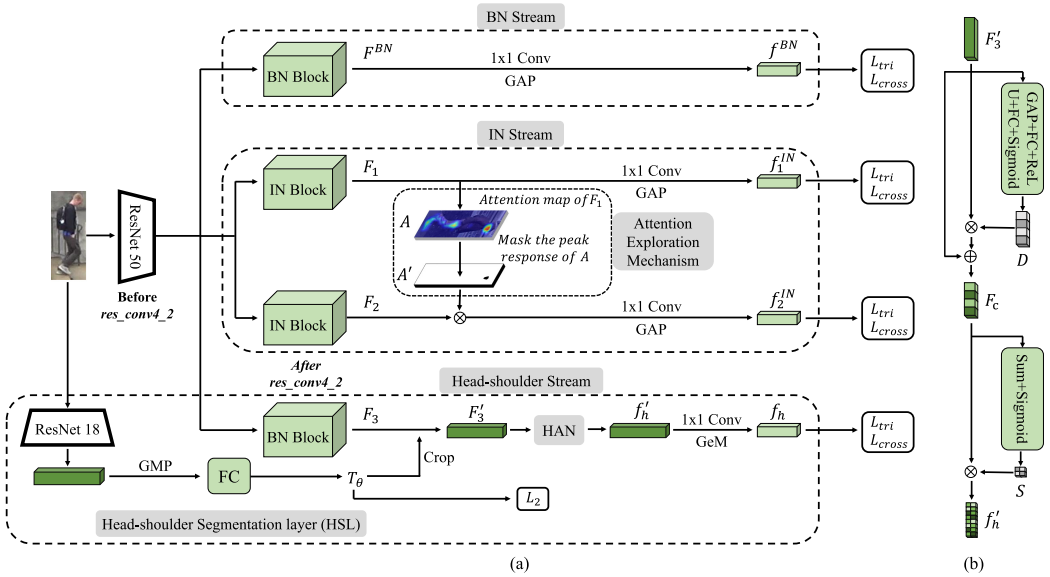

Fig. 3. Overview of the color-unrelated head-shoulder network (CUHS). (a) Our model is composed of three streams: the BN stream extracts the color-related feature with batch normalization; the IN stream extracts the color-unrelated feature with instance normalization and is designed as a two-branched structure. The two branches are able to complement each other for exploring richer visual cues by the attention exploration mechanism; the head-shoulder stream consists of a head-shoulder segmentation layer (HSL) for localizing the head-shoulder region and a head-shoulder attention network (HAN) for extracting the corresponding feature. (b) The head-shoulder attention network (HAN) structure. Here, $\oplus$, $\otimes$, GeM, GAP, GMP, L_{tri}, L_{cross}, L_2 indicate element-wise addition, element-wise multiplication, generalized mean pooling, global average pooling, global max pooling, triplet loss, cross-entropy loss, and L2 loss, respectively, and T_θ is the coordinates of the head-shoulder region bounding box.

subjects who wear black and white clothes for training. Fourthly, we simplify the model in [60] and the number of parameters is considerably reduced. Fifthly, as DukeMTMC-reID [39] has been withdrawn by its authors, we remove DukeMTMC-reID data from Black-reID [60] and add data from CUHK03 [28] to build the FG-reID dataset. Thanks to these improvements, the experiments prove that our model outperforms [60] with a more lightweight architecture.

2 RELATED WORK

2.1 Person Re-identification

Person re-identification (re-id) aims at matching pedestrian images with the same id across non-overlapping cameras. Re-id is usually treated as a classification task by many re-id methods, which attempts to divide the people with the same identity label into one category. Many metric learning and handcrafted features based methods [9, 24, 25, 29, 73] have been proposed. Recently, re-id performance has made great improvements [11, 14, 27, 31, 40, 51, 54, 66] due to the developments of CNNs.

We briefly divide re-id methods into four categories. The first class is part-based re-id [50, 52]. Part-based methods work on learning more discriminative local features by horizontally slicing a person image or feature map into several parts and extracting local features individually. Sun et al. [50] proposed **Part-based Convolutional Baseline (PCB)** for learning part-based features. PCB slices feature maps into several horizontal grids, learning local features individually, and

finally concatenating these part features for inference. Wang et al. [52] designed a multi-branch network, which consists of two local branches for part-based features learning and one global branch for global representation. However, part-based methods are influenced by pose variation, spatial misalignment, and occlusions, as they may lead to feature misalignment during feature matching.

The second class is pose-based re-id [30, 34, 37, 46, 71]. Pose-based methods attempt to learn semantic representations by pose information for feature alignment or generating person images with corresponding poses. Zhao et al. [71] proposed the Spindle Net, which extracts local features by pose information. Su et al. [46] designed a two-stream network for learning both the local and global features and fused these two types of features by a proposed Feature Weighting Net. [30, 34, 37] achieve pose normalization by generating person images of corresponding poses to overcome the pose variations in the re-id. However, pose-based approaches are sensitive to the accuracy of the pose estimation and the utilization of the pose estimators causes the network to become bigger and slower.

The third class is attention-based re-id [65, 72]. Attention-based methods are robust to the background clutter by paying attention to a region of interest and could be trained under less supervisory labels. Zhao et al. [72] proposed an attention-based model to divide the human body into several semantic regions, accordingly learn the representations over the regions, and compute the similarities between the corresponding regions of a pair of images as the matching score. Yang et al. [65] proposed an attention augmentation model, namely **Class Activation Maps Augmentation (CAMA)**, to expand the response regions of the attention map for exploring richer visual cues. CAMA consists of several branches which output various attention maps for a complement.

The fourth class is data-driven methods. Data-driven methods attempt to exploit GAN [37] or construct data generation systems [48, 55, 68] for making full use of synthesized images. Zhang et al. [68] construct a novel pipeline to generate a new synthesized dataset named UnrealPerson, which consists of 3,000 IDs and 120,000 instances. Qian et al. [37] proposed **pose-normalization GAN (PN-GAN)** for generating synthesizing person images conditional on the pose and hereafter extracting features robust to pose variations.

Our model is a pose-based method. Compared to the other pose-based methods, we simplify the pose estimators by a proposed head-shoulder segmentation layer for head-shoulder localization.

2.2 Fine-grained Person Re-identification

Most of the re-id methods assume that people are wearing clothes of different styles such that they can be distinguished by their appearance features. However, this assumption is problematic when people wear similar clothes. Yin et al. [67] proposed the study of the FG person re-id, which refers to conducting re-id when people dress in very similar clothes. [67] proposed a network to exploit dynamic pose features in the video sequences for video-based FG person re-id. [67] also built the first video-based FG person re-id dataset named **fine-grained person re-identification (FGPR)**, which contains 385 identities. However, dynamic pose features are not available on the image-based FG person re-id problem. Yan et al. [63] proposed to learn fine-grained features by a pairwise loss function that enforces a bounded penalization on the images of large differences and an exponential penalization on the images of small differences. This paper focuses on image-based FG person re-id. In our work, we introduce an image-based FG person re-id dataset and design a network to exploit head-shoulder information and color-unrelated features for dealing with the FG person re-id problem.

3 METHOD

We show the structure of the proposed color-unrelated head-shoulder network in Figure 3. CUHS is composed of the Batch Normalization (BN) stream, Instance Normalization (IN) stream,

and head-shoulder stream. The BN stream and IN stream work on extracting color-related and color-unrelated representation, respectively. The head-shoulder stream focuses on localizing the head-shoulder region and extracting the corresponding representation. The model is trained end-to-end using L2 loss, triplet loss, and cross-entropy loss. During inference, we attempt to retrieve pedestrians by calculating the Euclidean distance between features.

3.1 Backbone

We use the ResNet50 [10] fashion network as the backbone, which consists of a convolutional layer (res_conv1) and four residual blocks (res_conv2 - res_conv5). As shown in Figure 3, we retain the original ResNet50 [10] before res_conv4_2 as the common base model and divide the parts after res_conv4_1 into three independent streams.

3.2 Batch Normalization Stream

We extract the color-related feature through the BN stream. Let $X \in \mathbb{R}^{N\times C\times H\times W}$ denotes a feature map extracted from an input image, where N, C, H, W respectively indicate the batch size, the number of channels, the height, and the width. The BN layer normalizes features by:

$$BN(X) = \gamma^{bn} \cdot \frac{X - \mu^{bn}}{\sqrt{{\sigma^{bn}}^2 + \epsilon}} + \beta^{bn}, \tag{1}$$

where $\epsilon > 0$ is a small constant to avoid divided-by-zero, $\gamma^{bn} \in \mathbb{R}^C$ and $\beta^{bn} \in \mathbb{R}^C$ are affine parameters. $\mu^{bn} \in \mathbb{R}^C$ and $\sigma^{bn} \in \mathbb{R}^C$ are respectively mean value and standard deviation calculated with respect to a mini-batch and each channel:

$$\mu^{bn} = \frac{\sum_n \sum_{h,w} X}{N \cdot H \cdot W} \quad \text{and} \quad \sigma^{bn} = \sqrt{\frac{\sum_n \sum_{h,w} (X - \mu^{bn})^2}{N \cdot H \cdot W}}. \tag{2}$$

where μ^{bn} and σ^{bn} are updated by the moving average operation [17] at training time and fixed during inference. We retain the original ResNet50 [10] after res_conv4_2 as the BN block for extracting color-related features. Specifically, a feature map $F^{BN} \in \mathbb{R}^{H\times W\times C}$ is firstly extracted by the BN block, in which H, W, C denote the height, width, and channel number, respectively. Then, F^{BN} is processed successively by the **Global Average Pooling (GAP)** and a 1×1 convolution layer for channel reduction, producing the color-related representation $f^{BN} \in \mathbb{R}^{c\times 1\times 1}$.

3.3 Instance Normalization Stream

IN layers [8] normalize features by:

$$IN(X) = \gamma^{in} \cdot \frac{X - \mu^{in}}{\sqrt{{\sigma^{in}}^2 + \epsilon}} + \beta^{in}. \tag{3}$$

Different from BN, mean value μ^{in} and standard deviation σ^{in} here are calculated with respect to each sample and each channel:

$$\mu^{in} = \frac{\sum_{h,w} X}{H \cdot W} \quad \text{and} \quad \sigma^{in} = \sqrt{\frac{\sum_{h,w} (X - \mu^{in})^2}{H \cdot W}}. \tag{4}$$

After being processed by IN layers [8], the instance-specific contrast information (i.e., color and style information) could be filtered out from the content. Therefore, we construct the IN stream for extracting color-unrelated features. We replace BN layers in the BN block with IN layers to construct the IN block. However, IN inevitably results in the loss of some discriminative features [21].

To alleviate this problem, we design a two-branched structure. The two branches are able to complement each other for exploring richer visual cues.

In the first branch, firstly, a feature map $F_1 \in \mathbb{R}^{H\times W\times C}$ is extracted by the IN block. Then, F_1 is processed by the GAP and a 1×1 convolution layer to produce the color-unrelated representation f_1^{IN} of size $c\times 1\times 1$.

In the second branch, firstly, a feature map F_2 is extracted by the IN block. Then, we propose an **Attention Exploration Mechanism** for exploring more visual cues. As shown in Figure 3 (a), we calculate the spatial attention map of the F_1 by:

$$A = \sum_{n=1}^{C} F_1^n, \tag{5}$$

where F_1^n denotes the n-th channel of feature F_1 and A is the spatial attention map of F_1. After that, we set the top-k values in the spatial attention map A to be zero and others to be one to generate a mask A'. A' would mask the spatial peak response of the first IN branch and encourage the second IN branch to explore other interesting regions. We apply A' to the F_2 as follows to make the second branch work on exploring more visual cues:

$$F_2^{'} = F_2 \odot A', \tag{6}$$

where $\odot$ denotes the element-wise multiplication. Then, the representation $F_2^{'}$ is processed by the GAP and a 1×1 convolution layer to produce another color-unrelated representation f_2^{IN} of size $c\times 1\times 1$.

3.4 Head-Shoulder Stream

Head-shoulder Feature Construction. The head-shoulder region is loosely defined and a slight offset of the head-shoulder region localization would have little effect on the re-id performance. Therefore, different from other pose-based methods [30, 34, 37, 46, 71] which adopt off-the-shelf pose estimators for localizing body parts (e.g., arms, legs) and extracting features from them, we localize the head-shoulder region by a proposed lightweight head-shoulder segmentation layer. Furthermore, to increase the quality of the head-shoulder representation, a head-shoulder attention network is designed for feature enhancement.

As shown in Figure 3, we propose the head-shoulder stream for localizing the head-shoulder region and extracting corresponding features. The BN block in the head-shoulder stream is the same as that in the BN stream. Inspired by the ***Spatial Transformer Network*** **(STN)** [19], the HSL is designed for applying an affine transformation (i.e., scaling, translation, and rotation) to the feature map. As what we need is a bounding box to represent the head-shoulder region, we just let the HSL scale and translate, equivalent to give HSL the ability to crop on the input person image. The process of the HSL can be formulated as follows:

$$\begin{pmatrix} x_s \\ y_s \end{pmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \begin{pmatrix} x_t \\ y_t \\ 1 \end{pmatrix}, \tag{7}$$

where x_t,y_t and x_s,y_s are the targets and source coordinates, respectively, t_x,t_y and s_x,s_y are translation and scaling parameters, respectively. As shown in Figure 3, the affine transformation parameters T_θ (e.g., t_x,t_y,s_x,s_y) are produced by the fully connected layer of size $C\times 4$, and then the head-shoulder region bounding box is generated based on the T_θ.

Specifically, a feature map $F_3 \in \mathbb{R}^{H\times W\times C}$ is firstly extracted by the BN block, then the head-shoulder bounding box is produced by the HSL and leveraged to crop the head-shoulder

representation $F_3^{'} \in \mathbb{R}^{h \times w \times C}$ from F_3. After that, $F_3^{'}$ is successively processed by the head-shoulder attention network (HAN), **Generalized Mean Pooling (GeM)** and a 1×1 convolution layer to produce the head-shoulder feature $f_h \in \mathbb{R}^{c \times 1 \times 1}$.

Head-shoulder Attention Network. The structure of the head-shoulder attention network (HAN) is shown in Figure 3(b). Since different channels and spatial locations of the feature map represent different patterns and semantics, respectively. HAN works on both spatial and channel dimensions to enhance the head-shoulder features.

For the channel attention, the feature map $F_3^{'} \in \mathbb{R}^{h \times w \times C}$ firstly passes through a gating mechanism which is composed of a GAP, a fully connected layer $F_r \in \mathbb{R}^{C \times \frac{C}{r}}$ for dimension reduction, a RELU activation, another dimension incrementation fully connected layer $F_i \in \mathbb{R}^{\frac{C}{r} \times C}$ and a sigmoid activation σ, where r denotes the reduction ratio. Then, the channel attention is conducted with a shortcut connection as follows:

$$F_c = F_3^{'} + F_3^{'} \cdot D, \tag{8}$$

$$D = \sigma(F_i ReLU(F_r GAP(X))), \tag{9}$$

where $\cdot$ is element-wise multiplication and F_c is the enhanced feature map in the channel dimension.

For the spatial attention, we enhance the feature map by strengthening the peak responses in the spatial location, which is formulated as:

$$f_h^{'} = F_c \cdot S, \tag{10}$$

$$S = \sigma\left(\sum_{n=1}^{C} F_c^n\right), \tag{11}$$

where $f_h^{'}$ is the enhanced head-shoulder representation, and F_c^n denotes the n-th channel of feature F_c.

3.5 Model Training

During the model training, we employ cross-entropy loss and triplet loss for classification and metric learning, respectively. We treat the re-id task as a multi-class classification problem. The cross-entropy loss is defined as follows:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N} \log \frac{e^{W_{yi}^T f_i}}{\sum_{k=1}^{C} e^{W_k^T f_i}}, \tag{12}$$

where N denotes the batch size, C is the categories number in the training set, and W_k denotes the weight for the k-th class.

We use the batch-hard triplet loss [15], which is applied to the hardest examples in a mini-batch for a more discriminative representation learning. The triplet loss is formulated as follows:

$$\begin{aligned} \mathcal{L}_{triplet} = \sum_{i=1}^{k} \sum_{a=1}^{M} [\alpha + \max_{p=1\ldots M} ||\mathbf{f}_a^i - \mathbf{f}_p^i||_2 - \\ \min_{\substack{n=1\ldots M \\ j=1\ldots K \\ j \neq i}} ||\mathbf{f}_a^i - \mathbf{f}_n^j||_2]_+, \end{aligned} \tag{13}$$

where f_a^i, f_p^i, f_n^i are anchor features, positive features, and negative features, respectively. α is a hyper-parameter to control the margin between the negative and positive pairs in the feature

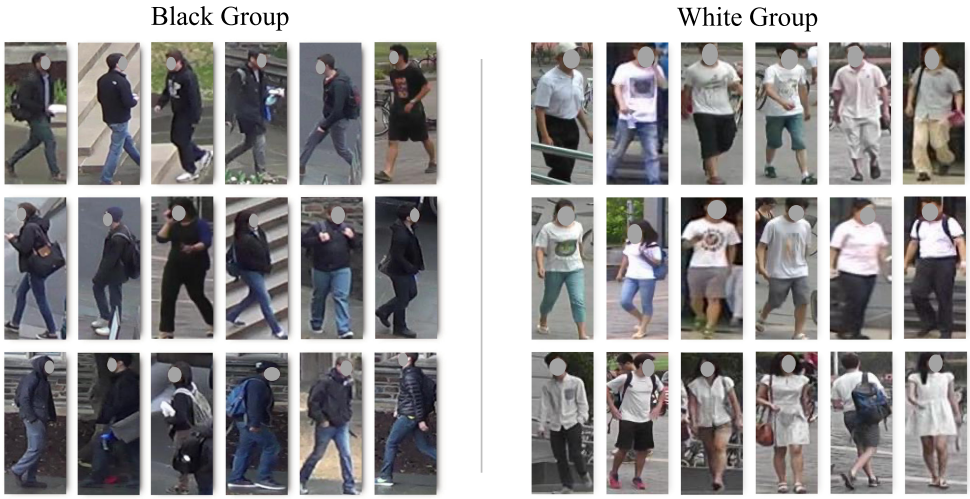

Fig. 4. Some examples of the FG-reID dataset. The FG-reID dataset consists of a black group and a white group.

space. In each mini-batch, we sample k identities with M images per identity to calculate the triplet loss.

The cross-entropy and triplet loss are applied to f^{BN}, f_h, f_1^{IN} and f_2^{IN} for model training and balanced by the parameters η_1 and η_2 as follows:

$$\mathcal{L} = \eta_1 \mathcal{L}_{ce} + \eta_2 \mathcal{L}_{triplet}. \tag{14}$$

In addition, we employ L2-loss to the head-shoulder segmentation layer, which is formulated as follows:

$$\mathcal{L}_2 = \frac{1}{N} \sum_{i=1} N ||c^i - r(p^i)||_2^2, \tag{15}$$

where N denotes the batch size, c^i and p^i are the ground-truth and predicted coordinates of the head-shoulder region bounding box in the i-th image, and r is the transformation function that transforms c^i and p^i to the same coordinate system.

During model training, we first train the HSL with Equation (15), then freeze HSL and train the model with Equation (14). During inference, we concat f^{BN}, f_1^{IN}, f_2^{IN}, and f_h for person re-id.

4 THE FG-REID DATASET

We introduce the FG-reID dataset for the FG person re-id problem. FG-reID is derived from the CUHK03 [28], Market-1501 [75], Occluded-REID [84], and Partial-REID [76] datasets.

4.1 Description

We put some examples of the FG-reID dataset in Figure 4 and the details of the dataset are illustrated in Table 1. The FG-reID dataset has several advantages as follows. Firstly, this is the first dataset that focuses on the image-based FG person re-id problem. The FG-reID dataset is divided into two groups by the different colored clothes. The first group is called the black group, which contains 5,756 images of 655 identities in the training set, 1,471 images of 418 identities in the query set, and 3,947 images of 837 identities in the gallery set. The second group is called the white group, which contains 10,040 images covering 586 identities in the training set, 2,756 query

Table 1. The Descriptions of the FG-reID Dataset

Dataset	Black Group		White Group	
	#id	#img	#id	#img
train	655	5,756	586	10,040
query	418	1,471	628	2,756
gallery	837	3,947	706	12,312

images of 628 identities, and 12,312 gallery images of 706 identities for the test. Secondly, the head-shoulder region bounding box is provided and people who wear black and white clothes have been labeled. Thirdly, we add some people wearing clothes of other colors to the two groups. The reason for this is that we want our model to be effective for both Fine-Grained and conventional re-id problems.

4.2 Data Collection

We pick out 170, 5, 24, and 57 subjects who wear black clothes from CUHK03 [28], Partial-REID [76], Occluded-REID [84], and Market-1501 [75], respectively for the black group. We also pick out 336 pedestrians who wear white clothes from Market-1501 [75] to build the white group. Furthermore, we add a small number of people who wear clothes of other colors to the two groups.

5 EXPERIMENT

5.1 Implementation Details

Datasets. To investigate the effectiveness of our model on both the FG and conventional re-id problems, we evaluate our method on the following three datasets. (1) The FG-reID dataset, which is established by us and contains 1,241 and 1,543 identities in training and test sets, respectively. (2) The Market-1501 [75] dataset which is captured by six cameras and in total offers 32,688 images of 1,501 identities. (3) The DukeMTMC-reID [39] dataset which contains 1,404 subjects, 16,522 images in the training set, 2,228 images in the query set, and 17,661 images in the gallery set. Furthermore, we provide the head-shoulder region bounding box for the DukeMTMC-reID and Market-1501 datasets.

Training Details. We resize images to 384×128. We set the batch size N to 64 and channel number c to 1536. We adopt data augmentation including random erasing [81] and horizontal flipping. Following other works [50, 52], the GAP and fully connected layers at the end of the original ResNet-50 are removed and we set the stride of the last convolutional layer to 1. We first train the **head-shoulder segmentation layer (HSL)** for 50 epochs and freeze it. Then, we train the whole network for another 90 epochs. We utilize adaptive gradient (Adam) [23] to optimize the network with β_1, β_2 and weight decay of 0.9, 0.999, and 5e-4, respectively. We initially set the learning rate to be 3e-4 and then divide it by 10 at 40 and 70 epochs. The parameters η_1 and η_2 in Equation (14) are set to 1 and α in Equation (13) is set to 0.3.

Evaluation Metrics. Cumulative Matching Characteristic (CMC) curves and **mean average precision (mAP)** are used as the protocols to evaluate the re-id performance. We conduct experiments in a single query setting.

5.2 Comparison with State-of-the-art Methods

Results on FG-reID Dataset. Table 2 shows the comparison of our method with previous methods (i.e., PCB [50], AlignedReID [69], HAA(ResNet50) [60], HAA(MGN) [60], Top-DB-Net [38], APNet-C [2], MGN [52], ReIDCaps [16], ISP [83], RGA [70]) on the FG-reID dataset,

Table 2. Quantitative Comparisons with State-of-the-art Methods on FG-reID Dataset

Method	Black Group		White Group		Params Size (MB)
	mAP	Rank-1	mAP	Rank-1	
AlignedReID [69]	71.2	73.8	80.5	91.3	–
PCB [50]	68.5	75.3	78.2	90.8	–
MGN [52]	76.6	78.8	85.8	94.3	262.5
ReIDCaps [16]	69.5	75.2	81.8	92.1	–
ISP [83]	67.5	74.2	80.2	91.0	–
RGA [70]	71.5	73.5	81.4	92.8	–
HAA(MGN) [60]	78.2	80.9	88.1	95.3	486.2
Top-DB-Net [38]	70.8	74.5	83.7	93.1	–
APNet-C [2]	74.6	77.3	87.2	93.8	–
Baseline	71.5	73.8	75.8	89.5	91.7
HAA(ResNet50) [60]	76.2	78.7	84.4	93.5	335.4
CUHS (ours)	**80.1**	**82.5**	**88.5**	**95.5**	342.3

The bold number denotes the best performance. All the methods use ResNet50 [10] as the backbone. Our model outperforms other methods on the FG-reID dataset.

all the methods use ResNet50 [10] as backbone. The results illustrate that CUHS outperforms other methods in dealing with the FG person re-id problem. Specifically, in the black group, HAA (ResNet50) gets 76.2% mAP and 78.7% rank-1, HAA (MGN) gets 78.2% mAP and 80.9% rank-1, and our CUHS obtains mAP and rank-1 of 80.1% and 82.5%, respectively, exceeding HAA (ResNet50) by 3.9% mAP and 3.8% rank-1, and exceeding HAA (MGN) by 1.9% mAP and 1.6% rank-1. In the white group, the results also show that CUHS gets the best result with 88.5% mAP and 95.5% rank-1, which surpasses HAA (ResNet50) by 4.1% mAP and 2.0% rank-1, and exceeding HAA (MGN) by 0.4% mAP and 0.2% rank-1. Furthermore, Table 2 shows that the parameters size of CUHS is about 342.3*MB*, which is comparable with HAA (ResNet50) and 143.9*MB* smaller than the HAA (MGN). The results prove that CUHS is effective in solving the FG person re-id problem.

Results on Market-1501, DukeMTMC-reID and MSMT17 dataset. To evaluate our method on the conventional re-id problem, we show the comparison between our method and previous methods on the Market-1501 [75], DukeMTMC-reID [39], and MSMT17 [57] datasets in Table 3. All the methods use ResNet50 [10] as the backbone and are classified into five categories. *Pose-guided methods* attempt to extracting local features with the help of the pose information and precisely align these representations during feature matching. *Part-based methods* horizontally slice feature maps or images into several grids and train them individually to improve re-id performance. *Attention-based methods* compute attention maps to increase the attention weights of the discriminative regions of interest. *Head-shoulder methods* leverage head-shoulder information to assist re-id. Note that re-ranking [80] is not adopted in all the reported results for fair comparisons.

Table 3 shows that CUHS also gets the best results on both the Market-1501, DukeMTMC-reID, and MSMT17 datasets. Specifically, on the Market-1501 dataset, HAA (ResNet50) achieves mAP of 85.6% and rank-1 of 94.2%, HAA (MGN) achieves mAP of 89.5% and rank-1 of 95.8%, but our CUHS achieves mAP and rank-1 of 90.1% and 96.1%, respectively, exceeding HAA (ResNet50) by 4.5% mAP and 1.9% rank-1, and exceeding HAA (MGN) by 0.6% mAP and 0.3% rank-1. On the DukeMTMC-reID dataset, the results also show that CUHS gets the best result with 80.7% mAP and 89.7% rank-1, which surpasses HAA (ResNet50) by 6.5% mAP and 3.3% rank-1, and surpasses HAA (MGN) by 0.3% mAP and 0.3% rank-1. On the MSMT17 dataset, CUHS outperforms other methods by at least 1.5% mAP and 0.8% Rank-1 accuracy.

Table 3. Quantitative Comparisons with State-of-the-art Methods on Market-1501, DukeMTMC-reID, and MSMT17 Dataset

	Method	Market-1501		DukeMTMC-reID		MSMT17	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Basic-CNN	Baseline	84.6	93.3	75.3	86.2	50.2	74.1
Pose-guided methods	Spindle [71]	–	76.9	–	–	–	–
	MSCAN [26]	57.5	80.8	–	–	–	–
	PDC [46]	63.4	84.1	–	–	–	–
	Pose Transfer [30]	68.9	87.7	48.1	68.6	–	–
	PN-GAN [37]	72.6	89.4	53.2	73.6	–	–
	PSE [41]	69.0	87.7	62.0	79.8	–	–
	MGCAM [45]	74.3	83.8	–	–	–	–
	MaskReID [35]	75.3	90.0	61.9	78.9	–	–
	Part-Aligned [47]	79.6	91.7	84.4	69.3	–	–
	AACN [62]	66.9	85.9	59.3	76.8	–	–
	SPReID [22]	81.3	92.5	71.0	84.4	–	–
	PIE [74]	53.9	78.7	–	–	–	–
	PGFA [32]	76.8	91.2	65.5	82.6	–	–
	HOReID [51]	84.9	94.2	75.6	86.9	–	–
	BPBreID [44]	87.0	95.1	78.3	89.6	–	–
Part-based methods	AlignedReID [69]	79.3	91.8	–	–	–	–
	PCB+RPP [50]	81.6	93.8	69.2	83.3	–	–
	MGN [52]	86.9	95.7	78.4	88.7	–	–
	Deep-Person [1]	79.6	92.3	64.8	80.9	–	–
Attention-based methods	DLPAP [72]	63.4	81.0	–	–	–	–
	HA-CNN [36]	75.7	91.2	63.8	80.5	–	–
	DuATM [43]	76.6	91.4	64.6	81.8	–	–
	OSNet [82]	84.9	94.8	73.5	88.6	52.9	78.7
	BDB [7]	86.7	95.3	76.0	89.0	–	–
	GASM [12]	84.7	95.3	74.4	88.3	52.5	79.5
	ABDNet [4]	88.3	95.6	78.6	89.0	60.8	82.3
	FED [56]	86.3	95.0	78.0	89.4	–	–
	NFormer [53]	**91.1**	94.7	**83.5**	89.4	–	–
	CARL [58]	89.2	95.8	81.4	**91.2**	–	–
Head-Shoulder methods	HAA (MGN) [60]	89.5	95.8	80.4	89.0	–	–
	HAA (ResNet50) [60]	85.6	94.2	74.2	86.4	–	–
	CUHS (ours)	90.1	**96.1**	80.7	89.7	**62.3**	**83.2**

The bold number denotes the best performance. All the methods use ResNet50 [10] as the backbone. Our method achieves the best performance on both the Market-1501, DukeMTMC-reID, and MSMT17 datasets.

Results on FG-reID V2 dataset. To further evaluate the effectiveness in solving the FG re-id problem, we build a more challenging benchmark named FG-reID V2. Specifically, we combine the training sets and galleries of CUHK03 [28], Partial-REID [76], Occluded-REID [84], and Market-1501 [75], and pick out their query subjects wearing black clothes and white clothes, respectively, to build the new query set for FG-reID V2. We compare our method with state-of-the-art methods (i.e., PCB [50], HAA(ResNet50) [60], HAA(MGN) [60], Top-DB-Net [38], APNet-C [2], MGN [52]) on FG-reID V2 dataset, as shown in Table 4. All the methods use ResNet50 [10] as the backbone. The result shows that our method achieves the best result, with mAP of 73.4%, 86.7% and rank-1 of 77.5%, 93.8% on the black group V2 and white group V2, respectively. In addition, we can find that the performance on FG-reID V2 is lower than that on FG-reID, as shown in Table 2. This illustrates that FG-reID V2 is more challenging for evaluating FG re-id performance.

Table 4. Quantitative Comparisons with State-of-the-art Methods on FG-reID V2 Dataset

Method	Black Group V2		White Group V2	
	mAP	Rank-1	mAP	Rank-1
PCB [50]	62.5	69.8	77.5	87.9
MGN [52]	68.5	73.6	81.1	90.0
Top-DB-Net [38]	64.5	72.1	81.3	91.3
APNet-C [2]	72.8	76.1	85.9	92.5
HAA(MGN) [60]	71.4	75.6	85.5	91.8
Baseline	65.2	72.6	78.3	88.2
HAA(ResNet50) [60]	69.1	74.2	80.6	90.1
CUHS (ours)	**73.4**	**77.5**	**86.7**	**93.8**

All the methods use ResNet50 [10] as the backbone. The bold number denotes the best performance. Our model outperforms other methods on the FG-reID V2 dataset.

Table 5. Ablation Study on the Proposed IN Stream and Pooling Methods on the FG-reID, Market-1501, and DukeMTMC-reID Datasets

Method	Black Group		White Group		Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	71.5	73.8	75.8	89.5	84.6	93.3	75.3	86.2
CUHS-1 IN	76.6	78.1	86.4	95.0	88.9	95.7	78.9	89.0
CUHS (GMP)	75.7	78.1	86.2	93.9	89.1	95.5	79.2	88.9
CUHS (GAP)	79.7	82.4	87.0	94.5	89.2	95.1	79.3	88.9
CUHS (ours)	**80.1**	**82.5**	**88.5**	**95.5**	**90.1**	**96.1**	**80.7**	**89.7**

CUHS-1IN indicates only utilizing 1 IN branch in the IN stream. GMP and GAP indicate global max pooling and global average pooling, respectively. Rank-1 accuracy (%) and mAP score (%) are reported.

5.3 Ablation Study

Ablation study on the proposed IN stream. To evaluate the effectiveness of the proposed IN stream, we compare CUHS with CUHS-1IN, which indicates only utilizing one IN branch in the IN stream. The results are shown in Table 5. From Table 5 we can find CUHS outperforms CUHS-1IN on all the datasets. Specifically, CUHS achieves mAP scores of 80.1%, 88.5%, 90.1%, and 80.7% on the black group, white group, Market-1501, and DukeMTMC-reID datasets, respectively, which is 3.5%, 2.1%, 1.2%, and 1.8% higher than the corresponding metrics of CUHS-1IN. The results in Table 5 have validated the effectiveness of the proposed IN stream.

Ablation study on the pooling methods. In the CUHS, GeM pooling is adopted in the head-shoulder stream. In Table 5, we compare GeM pooling with other pooling methods and evaluate the effectiveness of the GeM pooling on the FG-reID, Market-1501, and DukeMTMC-reID datasets. From Table 5 we can find that CUHS achieves the best results on all the datasets, with mAP of 80.1%, 88.5%, 90.1%, 80.7% and rank-1 of 82.5%, 95.5%, 96.1%, 89.7% on the black group, white group, Market-1501, and DukeMTMC-reID, respectively, which surpasses other pooling methods by at least 0.4%, 1.5%, 0.9%, 1.4% and 0.1%, 1.0%, 1.0%, 0.8% in mAP and rank-1, respectively. GeM pooling is formulated as follows:

$$GeM = \left[\left(\sum_{n=1}^{H \times W} x_{cn}^{\theta} \right)^{\frac{1}{\theta}} \right]_{c=1 \ldots C} \tag{16}$$

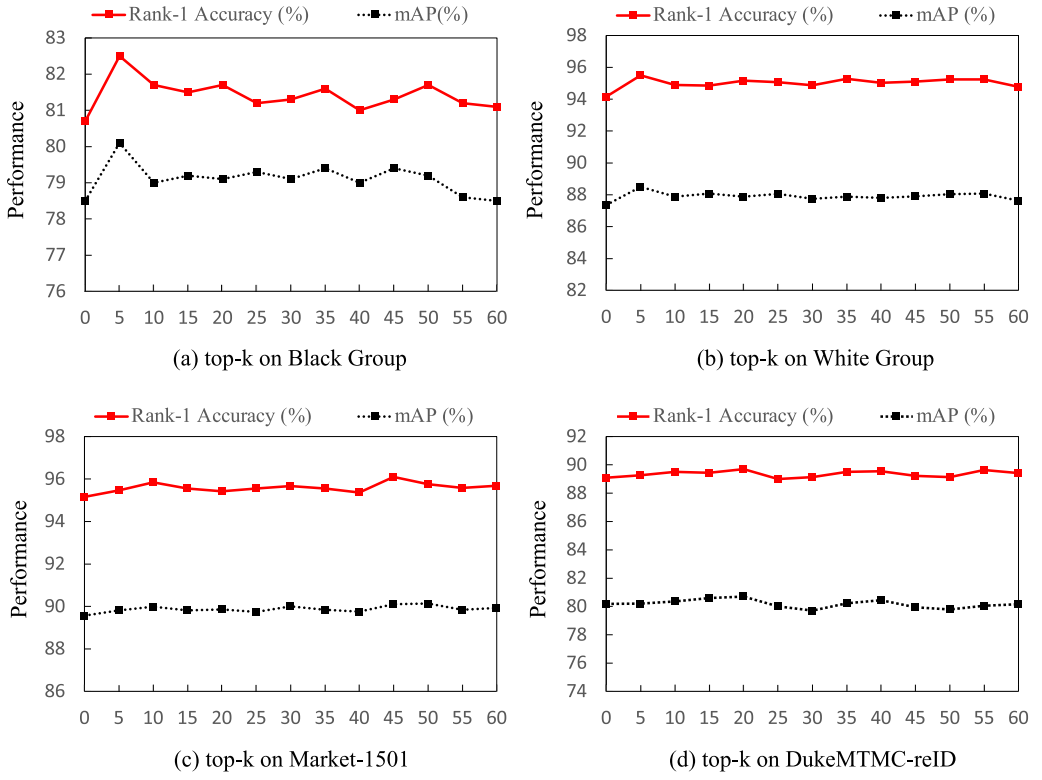

Fig. 5. Ablation study on the attention exploration mechanism on the FG-reID, Market-1501 and DukeMTMC-reID datasets. Top-k denotes that the top-k values of the first IN stream attention map are set to zero in the attention exploration mechanism. rank-1 accuracy (%) and mAP score (%) are reported.

where $n = 1...H \times W$ is a 'pixel' in the feature map, $c = 1...C$ is the channel number, x_{cu} is the corresponding tensor element, and θ is a learnable parameter. The localization degree of the feature map response increases with the increase of P. GeM pooling could be learned between the GAP($p = 1$) and GMP($p = \infty$). Therefore, the appropriate size area of the feature map could be aggregated for pooling by learning the p, which varies between the 'pixel' and 'whole image'.

Ablation study on the attention exploration mechanism. In this part, we mainly evaluate the effectiveness of the attention exploration mechanism. The results on FG-reID, Market-1501, and DukeMTMC-reID datasets under different top-k are illustrated in Figure 5. Top-k denotes that the top-k spatial peak responses of the first IN branch are masked by the attention exploration mechanism. Top-k equals 0 means that the attention exploration mechanism is not utilized. Rank-1 accuracy (%) and mAP score (%) are both reported. Firstly, in Figure 5(a), we can find that CUHS achieves the best result in the black group when top-k equals 5, which leads to performance improvements of approximately 1.8% and 1.6% for rank-1 accuracy and mAP score, respectively, compared to when top-k equals 0. Secondly, in Figure 5(b), the results show that CUHS achieves the best performance in the white group when top-k equals 5, leading to improvements of approximately 1.4% and 1.2% in rank-1 accuracy and mAP score, respectively, compared to when top-k equals 0. Thirdly, we can find from Figure 5(c) that CUHS achieves the best result on the Market-1501 when top-k equals 45, which gives performance gains of about 1.0% and 0.6% for rank-1 accuracy and mAP score respectively than top-k equals 0. Fourthly, in Figure 5(d), CUHS

Table 6. Ablation Study on the BN Stream, Head-Shoulder Stream, IN Stream on the FG-reID, Market-1501, and DukeMTMC-reID Datasets

Dataset	BN Stream	Head-Shoulder Attention Stream	IN Stream	mAP	Rank-1
Black Group	√	×	×	72.8	74.9
	√	√	×	75.4	78.5
	√	√	√	**80.1**	**82.5**
White Group	√	×	×	82.7	93.2
	√	√	×	84.6	93.8
	√	√	√	**88.5**	**95.5**
Market-1501	√	×	×	86.2	94.4
	√	√	×	87.1	95.5
	√	√	√	**90.1**	**96.1**
DukeMTMC-reID	√	×	×	77.1	88.0
	√	√	×	77.7	88.4
	√	√	√	**80.7**	**89.7**

Rank-1 accuracy (%) and mAP score (%) are reported.

Table 7. Comparison of Head-Shoulder Segmentation Layer and Fixed Head-Shoulder Region

Method	Black Group		White Group	
	mAP	Rank-1	mAP	Rank-1
Fixed head-shoulder region	76.5	78.6	85.2	92.3
CUHS (learned head-shoulder region)	**80.1**	**82.5**	**88.5**	**95.5**

The experiment is conducted on the FG-reID dataset. 'Fixed head-shoulder region' means we horizontally slice a feature map into three parts and use the first part as the head-shoulder region. Rank-1 accuracy (%) and mAP score (%) are reported.

achieves the best performance on the DukeMTMC-reID dataset when top-k equals 20, resulting in improvements of approximately 0.6% and 0.3% in rank-1 accuracy and mAP score, respectively, compared to when top-k equals 0. The results in Figure 5 have demonstrated the effectiveness of the attention exploration mechanism, especially for the FG person re-id problem.

Ablation study on the proposed streams. In this part, we investigate the effectiveness of our proposed streams, i.e., BN stream, head-shoulder stream and IN stream. The experiments are conducted on the FG-reID, Market-1501, and DukeMTMC-reID datasets, and the results are shown in Table 6. From the results, we can find that the head-shoulder stream gives mAP scores gains of 2.6%, 1.9%, 0.9, 0.6% on the black group, white group, Market-1501, and DukeMTMC-reID respectively. Furthermore, the IN stream gives another mAP scores and rank-1 accuracy gains of 4.7% and 4.0%, 3.9% and 1.7%, 3.0% and 0.6%, 3.0% and 1.3% respectively on the black group, white group, Market-1501, and DukeMTMC-reID. The results in Table 6 have proved the effectiveness of the proposed streams.

Ablation study on the effectiveness of head-shoulder segmentation layer. The head-shoulder segmentation layer is proposed for the head-shoulder region localization. We compare the proposed head-shoulder segmentation layer with the fixed head-shoulder region, as shown in Table 7. 'Fixed head-shoulder region' means we horizontally slice a feature map into three parts and use the first part as the head-shoulder region. The results show that the head-shoulder segmentation layer could give performance gains of 3.6% mAP and 3.3% mAP on the black group and white group respectively. Furthermore, we visualize the effectiveness of the head-shoulder

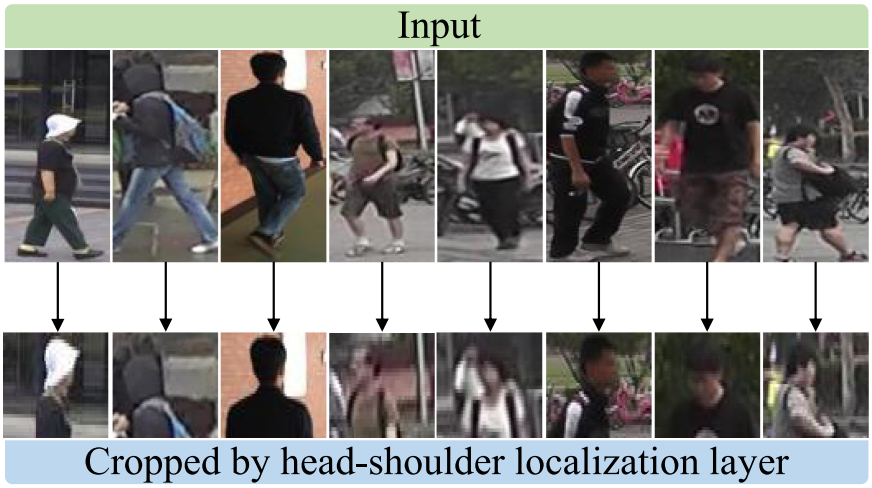

Fig. 6. Visualization of the head-shoulder segmentation layer.

Table 8. Comparison of Different Part Features

Method	Black Group		White Group	
	mAP	Rank-1	mAP	Rank-1
Torso	72.7	77.4	84.8	90.4
Leg	68.8	76.2	82.7	88.8
Head-shoulder (used in CUHS)	79.5	81.8	88.5	95.5
Fused (head-shoulder+torso+leg)	**80.1**	**82.5**	**88.7**	**95.8**

The experiment is conducted on FG-reID dataset. Rank-1 accuracy (%) and mAP score (%) are reported.

segmentation layer in Figure 6. The first row shows the original input image, and the second row shows the results processed by the head-shoulder segmentation layer. The results show that the head-shoulder segmentation layer is able to localize the head-shoulder region.

Ablation study on the effectiveness of head-shoulder features. To validate the effectiveness of head-shoulder features, we use the head-shoulder stream structure to extract torso and leg features and fuse head-shoulder, torso and leg features for comparison. We replace the head-shoulder segmentation layer with a horizontal division of features into three parts and use the second and third parts for torso and leg features extraction, respectively. As shown in Table 8, CUHS outperforms torso and leg features by at least 7.4% mAP and 11.3% mAP on the black group respectively. Also, head-shoulder features in CUHS also get similar performance to the fused features but with only a single branch. The results have validated the effectiveness of head-shoulder features for solving fine-grained person re-id problems.

Comparisons with different loss functions. We study ablation studies on the effectiveness of different loss functions, as shown in Table 9. The experiment is conducted on the FG-reID dataset. We replace $\mathcal{L}_{triplet}$ in Equation (14) with Lifted Loss [33], Contrastive Loss [79], Circle Loss [49], and Instance Loss [78], respectively. The result shows that our method with triplet loss [15] achieves the best performance, with mAP of 80.1%, 88.5% and rank-1 of 82.5%, 95.5% on the black group and white group, respectively.

Visualization analysis. Figure 7 shows the attention maps of each stream on FG-reID, Market-1501 and DukeMTMC-reID datasets calculated by CAM [42]. Row images and attention maps of the

Table 9. Ablation Study on the Effectiveness of Different Loss Functions

Method	Black Group		White Group	
	mAP	Rank-1	mAP	Rank-1
Lifted Loss [33]	73.5	78.8	81.4	93.2
Contrastive Loss [79]	75.6	80.7	81.9	92.4
Circle Loss [49]	74.3	79.5	79.6	92.7
Instance Loss [78]	78.7	81.8	88.0	95.3
CUHS (ours)	**80.1**	**82.5**	**88.5**	**95.5**

The experiment is conducted on the FG-reID dataset. We replace $\mathcal{L}_{triplet}$ in Equation (14) with different loss functions. Rank-1 accuracy (%) and mAP score (%) are reported.

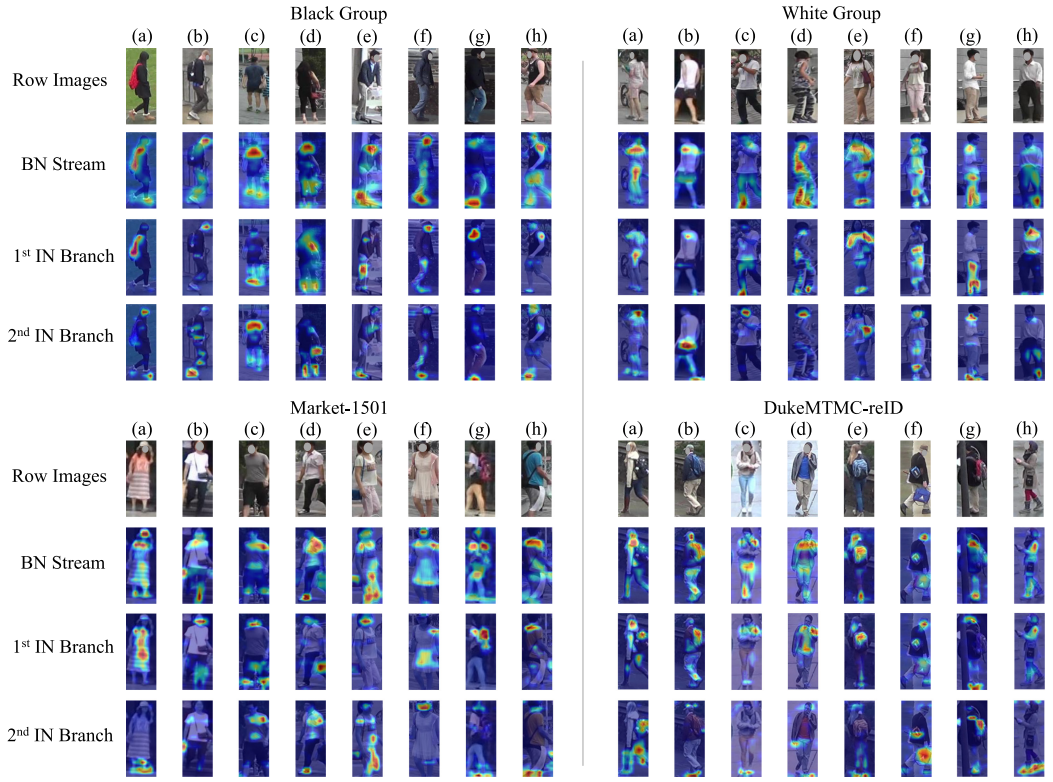

Fig. 7. The visualization of the attention maps of each stream on FG-reID, Market-1501, and DukeMTMC-reID datasets. Row images, attention maps of the BN stream, 1^{st} IN branch, and 2^{nd} IN branch are shown from top to bottom.

BN stream, 1^{st} IN branch, and 2^{nd} IN branch are shown from top to bottom. From Figure 7 we can find that the BN stream obtains bigger interesting regions than IN stream, which is corresponding to the opinion that IN eliminates individual contrast, but diminishes discriminative features at the same time. However, thanks to the attention exploration mechanism, the 1^{st} IN branch and 2^{nd} IN branch tend to focus on different semantic regions and complement each other. For example, in the (a) column of the black group, the BN stream pays attention to the head-shoulder region,

backpack and shoes. The 1^{st} IN branch only focuses on the backpack but the 2^{nd} IN branch tends to focus on the head and shoes for complementarity. Figure 7 illustrates that the attention exploration mechanism is able to increase the discrimination of the features extracted by the IN stream, and enable the IN stream to obtain almost the same interesting regions as the BN stream.

6 CONCLUSION

In this paper, we focus on the fine-grained (FG) person re-id problem and introduce an image-based FG person re-id dataset. To solve the FG person re-id problem, we propose the color-unrelated head-shoulder network (CUHS) for exploiting the head-shoulder feature and color-unrelated features. Furthermore, we design an attention exploration mechanism for solving the problem of inter-class discrimination reduction in the IN processing. Our method surpasses previous methods on FG-reID, Market-1501, and DukeMTMC-reID datasets, and is proven to be effective in solving both FG person re-id and conventional re-id problems.

ACKNOWLEDGMENTS

The authors would like to thank reviewers for providing valuable suggestions to improve this paper.

REFERENCES

[1] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. 2020. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition.* 98 (2020).

[2] Guangyi Chen, Tianpei Gu, Jiwen Lu, Jin-An Bao, and Jie Zhou. 2021. Person re-identification via attention pyramid. *IEEE Transactions on Image Processing* 30 (2021), 7663–7676.

[3] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji. 2021. Occlude them all: Occlusion-aware attention network for occluded person Re-ID. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 11833–11842.

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. 2019. ABD-Net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[5] Xiaodong Chen, Xinchen Liu, Wu Liu, Xiao-Ping Zhang, Yongdong Zhang, and Tao Mei. 2021. Explainable person re-identification with attribute-guided metric distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 11813–11822.

[6] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. 2021. IDM: An intermediate domain module for domain adaptive person re-ID. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 11864–11874.

[7] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. 2019. Batch DropBlock network for person re-identification and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629* (2016).

[9] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2360–2367.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[11] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[12] Lingxiao He and Wu Liu. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In *European Conference on Computer Vision.* Springer, 357–373.

[13] Lingxiao He, Wu Liu, Jian Liang, Kecheng Zheng, Xingyu Liao, Peng Cheng, and Tao Mei. 2021. Semi-supervised domain generalizable person re-identification. *arXiv preprint arXiv:2108.05045* (2021).

[14] Lingxiao He, Yinggang Wang, Wu Liu, Xingyu Liao, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *ArXiv* (2017).

[16] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. 2020. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 10 (2020), 3459–3471. https://doi.org/10.1109/TCSVT.2019.2948093

[17] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. PMLR, 448–456.

[18] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. 2021. Towards discriminative representation learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8526–8536.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial transformer networks. In *Conference and Workshop on Neural Information Processing Systems (NIPS)*.

[20] Haoxuanye Ji, Le Wang, Sanping Zhou, Wei Tang, Nanning Zheng, and Gang Hua. 2021. Meta pairwise relationship distillation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3661–3670.

[21] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[23] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2288–2295.

[25] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. 2012. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 7 (2012), 1622–1634.

[26] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. 2017. Learning deep context-aware features over body and latent parts for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 7398–7407.

[27] Shuangqun Li, Xinchen Liu, Wu Liu, Huadong Ma, and Haitao Zhang. 2016. A discriminative null space based deep learning approach for person re-identification. In *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*.

[28] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[29] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. 2018. Pose transferrable person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Jingke Meng, Wei-Shi Zheng, Jian-Huang Lai, and Liang Wang. 2021. Deep graph metric learning for weakly supervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[32] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[33] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*.

[34] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. 2018. MaskReID: A mask based deep ranking neural network for person re-identification. *ArXiv* (2018).

[36] Wei qi Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[37] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[38] Rodolfo Quispe and Helio Pedrini. 2021. Top-DB-Net: Top dropblock for activation enhancement in person re-identification. In *ICPR*.

[39] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision (ECCV) Workshops*.

[40] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, Yuhao Cheng, and Tao Mei. 2019. POINet: Pose-guided ovonic insight network for multi-person pose tracking. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*.

[41] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex Chichung Kot, and Gang Wang. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[44] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. 2023. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1613–1623.

[45] Chunfeng Song, Ying Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[46] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2017. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[47] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. *ArXiv* (2018).

[48] Xiaoxiao Sun and Liang Zheng. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*.

[49] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*.

[50] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[51] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. *Proceedings of the 26th ACM International Conference on Multimedia* (2018).

[53] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. 2022. NFormer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7297–7307.

[54] Xueping Wang, Min Liu, Dripta S Raychaudhuri, Sujoy Paul, Yaonan Wang, and Amit K Roy-Chowdhury. 2021. Learning person re-identification models from videos with weak supervision. *IEEE Transactions on Image Processing* 30 (2021), 3017–3028.

[55] Yanan Wang, Shengcai Liao, and Ling Shao. 2020. Surpassing real-world source training data: Random 3D characters for generalizable person re-identification. In *ACM MM*.

[56] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. 2022. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4754–4763.

[57] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 79–88.

[58] Jinlin Wu, Yuxin Yang, Zhen Lei, Yang Yang, Shukai Chen, and Stan Z. Li. 2023. Camera-aware representation learning for person re-identification. *Neurocomputing* 518 (2023), 155–164.

[59] Boqiang Xu, Lingxiao He, Jian Liang, and Zhenan Sun. 2022. Learning feature recovery transformer for occluded person re-identification. *IEEE Transactions on Image Processing* 31 (2022), 4651–4662.

[60] Boqiang Xu, Lingxiao He, Xingyu Liao, Wu Liu, Zhenan Sun, and Tao Mei. 2020. Black re-ID: A head-shoulder descriptor for the challenging problem of person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*.

[61] Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. 2022. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *ECCV*.

[62] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-aware compositional network for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[63] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. 2021. Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia* (2021).
[64] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying-Cong Chen, and Wei-Shi Zheng. 2021. Learning to know where to see: A visibility-aware approach for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11885–11894.
[65] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 2019. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[66] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. 2022. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14298–14307.
[67] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. 2020. Fine-grained person re-identification. *International Journal of Computer Vision* (2020), 1–19.
[68] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. 2021. UnrealPerson: An adaptive pipeline towards costless person re-identification. In *CVPR*.
[69] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. 2017. AlignedReID: Surpassing human-level performance in person re-identification. *ArXiv* (2017).
[70] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-aware global attention for person re-identification. *CVPR* (2020).
[71] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[72] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3219–3228.
[73] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[74] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. 2019. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing* 28, 9 (2019), 4500–4509.
[75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. *IEEE International Conference on Computer Vision (ICCV)* (2015).
[76] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jian-Huang Lai, and Shaogang Gong. 2015. Partial person re-identification. *IEEE International Conference on Computer Vision (ICCV)* (2015).
[77] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. 2021. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8371–8381.
[78] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *TOMM* 16, 2 (2020), 1–23.
[79] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A discriminatively learned CNN embedding for person reidentification. *TOMM* 14, 1 (2017), 1–20.
[80] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
[81] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
[82] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3702–3712.
[83] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-guided human semantic parsing for person re-identification. *ECCV* (2020).
[84] Jiaxuan Zhuo, Zeyu Chen, Jian-Huang Lai, and Guangcong Wang. 2018. Occluded person re-identification. *IEEE International Conference on Multimedia and Expo (ICME)* (2018).

Received 28 August 2022; revised 19 March 2023; accepted 9 May 2023