



# Fine-Grained Person Re-identification

Jiahang Yin<sup>1</sup> · Ancong Wu<sup>2,3</sup> · Wei-Shi Zheng<sup>4,5,6</sup>

Received: 18 December 2018 / Accepted: 25 October 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

Person re-identification (re-id) plays a critical role in tracking people via surveillance systems by matching people across non-overlapping camera views at different locations. Although most re-id methods largely depend on the appearance features of a person, such methods always assume that the appearance information (particularly color) is distinguishable. However, distinguishing people who dress in very similar clothes (especially the same type of clothes, e.g. uniform) is ineffective if relying only on appearance cues. We call this problem the *fine-grained person re-identification (FG re-id)* problem. To solve this problem, rather than relying on clothing color, we propose to exploit two types of local dynamic pose features: *motion-attentive local dynamic pose feature* and *joint-specific local dynamic pose feature*. They are complementary to each other and describe identity-specific pose characteristics, which are found to be more unique and discriminative against similar appearance between people. A deep neural network is formed to learn these local dynamic pose features and to jointly quantify motion and global visual cues. Due to the lack of a suitable benchmark dataset for evaluating the FG re-id problem, we also contribute a fine-grained person re-identification (FGPR) dataset, which contains 358 identities. Extensive evaluations on the FGPR dataset show that our proposed model achieves the best performance compared with related person re-id and fine-grained recognition methods for FG re-id. In addition, we verify that our method is still effective for conventional video-based person re-id.

**Keywords** Person re-identification · Fine-grained cross-view matching · Visual surveillance

---

Communicated by Greg Mori.

---

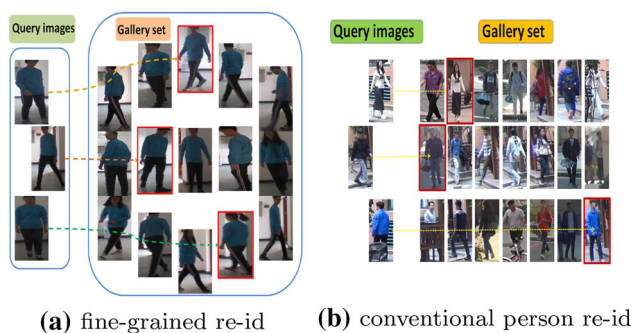
✉ Wei-Shi Zheng  
wszheng@ieee.org  
Jiahang Yin  
yinh5@mail2.sysu.edu.cn  
Ancong Wu  
wuancong@mail2.sysu.edu.cn

- <sup>1</sup> School of Data and Computer Science, SUN YAT-SEN University, Guangzhou, Guangdong, China
- <sup>2</sup> School of Electronic Information and Technology, SUN YAT-SEN University, Guangzhou, Guangdong, China
- <sup>3</sup> The Guangdong Province Key Laboratory of Information Security, Guangzhou 510275, P.R. China
- <sup>4</sup> Department of Computer Science, SUN YAT-SEN University, Guangzhou, Guangdong, China
- <sup>5</sup> Peng Cheng Laboratory, Shenzhen, China
- <sup>6</sup> Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

## 1 Introduction

Person re-identification (re-id) is an important field in video surveillance. As shown in Fig. 1, with some query images, we attempt to match images of the same person captured between two disjoint camera views. The common steps of re-id include person feature learning to find discriminative descriptions of pedestrians (Kviatkovsky et al. 2013; Liao et al. 2015; Farenzena et al. 2010; Zhao et al. 2013) and metric learning, which aims to measure the distance between features (Koestinger et al. 2012; Liao et al. 2015; Li et al. 2013; Zheng et al. 2013). Currently, the deep learning approach that unifies these two stages is dominant in re-id (Li et al. 2014; Ahmed et al. 2015; Xiao et al. 2016; Yi et al. 2014).

A primary characteristic of the existing re-id models is that they mostly assume that the appearance features are sufficient for distinguishing different people. Although this assumption is plausible in generic cases, it is problematic when people wear uniforms. For example, in some monitoring scenarios (banks, police stations, factories, etc.), people in the same section always have their own unique uniform and uniforms of different sections are of different colors or



**Fig. 1** Illustration of fine-grained person re-identification (FG re-id). For instance, for searching people in the query set (left part of the **a**), we match them with images of people who are wearing the same clothes in the gallery set (right part of the **a**) by measuring similarities. The greatest challenge in FG re-id is that people who are wearing very similar clothes are difficult to distinguish based on appearance

styles. In such a case, even humans experience difficulty in re-identifying a person, with other people in the same uniform, across camera views. In this work, we call this problem of distinguishing people who dress in very similar clothes (especially the same type of clothes, e.g. uniform) the *fine-grained person re-identification (FG re-id)* problem.

While many feature extractors for person re-id using either hand-crafted feature or deep learning have been developed (Kviatkovsky et al. 2013; Liao et al. 2015; Farenzena et al. 2010), they are more likely to find static (clothing) appearance difference between people. However, in FG re-id, the static (clothing) appearances of different people are very similar, and existing methods become ineffective. For example, when different people in similar clothes keep similar poses, like in the first column in Fig. 2, it is very difficult to distinguish them based only on clothing appearance.

Different from previous approaches, we find that the features extracted from local dynamic pose change caused by motion is more discriminative for telling different people apart in fine-grained person re-id, since although the local appearances of different people are similar, the local movements of different people can be different. For example, their moving postures, including the degree of their hand swing and the span of their feet, are unique motion habits, as shown each row in Fig. 2.

Therefore, we solve the FG re-id problem by exploiting discriminative local dynamic pose features along with extracting motion cues. In particular, we exploit two types of local dynamic pose features: motion-attentive local dynamic pose features and joint-specific local dynamic pose features. Specifically, we develop a multi-level attention network, which utilizes the feature maps of the motion stream to generate a series of fine-to-coarse attentions. These attentions, which are also called masks, can localize the key moving parts in an image, such as the head, hands and feet of a person so that the motion-attentive dynamic pose

feature can be modeled around the attentions. The visual results (i.e., feature maps) of some primary attentions (the first- and second-level attentions) are shown in Fig. 10. The region that receives more attention from our proposed model after adding our attentions is clearly observed. As the attentions could be biased by noise in the motion feature map due to the non-human moving in a video, we also investigate the joint-specific local dynamic pose feature around body joints, and such pose features could possess discriminative information complementary to the features extracted by the motion-attentive network, rather than global appearance features. In this work, we call the extracted two types of local dynamic pose features *local attentive dynamic pose features*, and we expect that they are more robust against the global motion variation of pedestrians who dress in similar clothes. A deep neural network is formed to learn all dynamic pose visual cues jointly with learning of representations from global to local. Note that although motion has been explored for person re-id (Xu et al. 2017; Chung et al. 2017), it has not previously been exploited to guide learning of the discriminant local dynamic pose features.

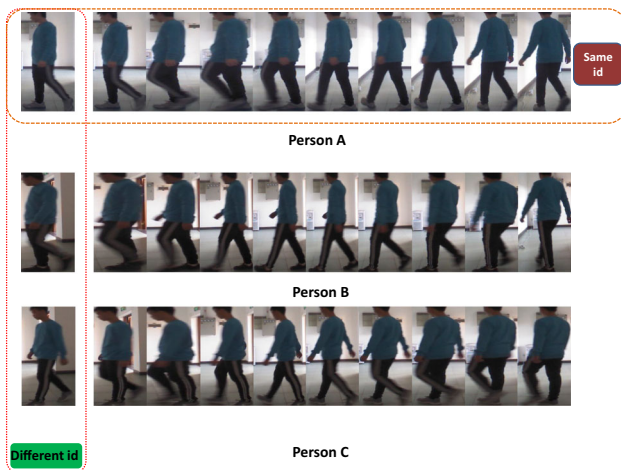
Since there is no existing dataset that is suitable for evaluating the FG re-id problem, we have collected a new dataset, named the FGPR dataset. This dataset contains 134,696 RGB images of 358 identities from 716 sequences. People are categorized into three groups with different colored clothes: blue, white and green. The numbers of corresponding identities are 200, 45 and 113, respectively. For each group, there are two camera views.

Based on the FGPR dataset, we have provided a new benchmark evaluation for studying the FG re-id problem. In comparison with some related methods, including deep and non-deep methods, the experimental results demonstrate the challenge of FG re-id and indicate that our proposed model has achieved the best performance. In addition, we show that our model is also effective for conventional video-based person re-id by the evaluation on MARS dataset (Zheng et al. 2016).

In summary, the main goals of this work are to identify the challenge of the FG re-id problem, which has rarely been investigated, and to provide an effective solution to this problem. For this purpose, we have collected a new dataset, named the FGPR dataset, which will be publicly available. To solve this problem, a deep neural network model is formulated that can achieve state-of-the-art performance on FG re-id.

## 2 Related Work

*Person re-identification* Person re-id is a challenging task for cross-view person matching. Most re-id methods assume that there is a clear difference in the color and style of clothes among people such that the appearance features are suffi-



**Fig. 2** Three human sequences with people dressed in blue uniform in the FGPR dataset. Different people appear extremely similar when they dress in the same clothes in the first column, and it is observed that the dynamic visual cues from their pose variations are more unique characteristics to distinguish different identities, as shown in columns 2–4 in the figure (Color figure online)

cient to distinguish different people. A large number of re-id methods based on handcrafted features and metric/subspace learning have been developed (Kviatkovsky et al. 2013; Liao et al. 2015; Farenzena et al. 2010; Zhao et al. 2013; Koestinger et al. 2012; Liao et al. 2015; Li et al. 2013). Recently, re-id models based on convolutional neural networks (Liu et al. 2017b; Li et al. 2018b; Simonyan and Zisserman 2014; He et al. 2016; Xiao et al. 2016; Sun et al. 2017) have achieved improved performance.

To some extent, video-based re-id (Xu et al. 2017; Chung et al. 2017; You et al. 2016; Liu et al. 2017c; Wu et al. 2018; Zhu et al. 2018; Zhang et al. 2019; Ye et al. 2019; Liu et al. 2019; Dai et al. 2018; Wu et al. 2019; Gou et al. 2016), which is closely related to our method, can either implicitly or explicitly include the gait feature (Rida et al. 2016; Wu et al. 2017; Makihara et al. 2017; Gou et al. 2016) for modeling (especially when appearance is not sufficient for classification) but without the requirement of good foreground segmentation, which is difficult for re-id due to occlusions and background clutter in surveillance videos. In general, the existing video-based person re-id methods either aim to extract spatial and temporal information, as in Xu et al. (2017), or to combine spatial and temporal information, as in Chung et al. (2017) and You et al. (2016). Recently, Zhang et al. (2019) proposed an intelligent feature aggregation method based on reinforcement learning to fuse video-level features. Liu et al. (2019) designed RRU and STIM to recover the missing parts, suppress noisy parts of the features and mine the spatial-temporal information. However, although non-color cues are exploited in the existing methods, they do not consider solving the fine-grained person re-id problem, and local dynamic pose features are

not exploited for this purpose. In addition, although the work in Xu et al. (2017) employed attention but for different purposes, the attention used in Xu et al. (2017) is for pooling features from different frames, while ours is for exploring attentive features in each frame by the guidance of motion and local feature extraction around joints.

Some approaches (Zhao et al. 2017; Su et al. 2017; Wei et al. 2017; Ge et al. 2018) combine pose information with person re-identification to extract discriminative features. Zhao et al. (2017) proposed Spindle Net, which is based on human body part and fuses the global and local features. Su et al. (2017) designed a two-stream network to learn global and part feature and utilized the proposed Feature Weighting Net to fuse the two types of features. Pose normalization (Liu et al. 2018; Pumarola et al. 2018; Qian et al. 2017) is also taken into consideration for re-id. They utilize generation models to achieve pose normalization and generate person image of the corresponding pose in order to overcome pose variation for person re-id. However, the objective of these works is different from ours. Our objective is not for pose alignment; instead, we aim to directly exploit discriminative local dynamic pose features for solving the FG re-id problem. Different from these works, we particularly localize the moving part by motion information except from the local part of static images and achieve a better performance.

Compared to video-based re-id and pose-guided re-id, FG re-id is considerably more challenging because of the similar appearances of different people. For the specific purpose of extracting fine-grained features, our method focuses more on local features and learns motion-attentive and joint-specific local dynamic pose features.

*Fine-grained classification* Studies on fine-grained image recognition can generally be divided into two steps: fine-grained feature learning and discriminative region localization. For discriminative region localization, previous works mainly focus on leveraging the extra bounding box annotations and part annotations to localize significant regions in fine-grained recognition (Huang et al. 2016; Branson et al. 2014; Lin et al. 2015; Johnson et al. 2016). However, this approach is not practical for large-scale real problems. Recently, there have been numerous emerging studies working towards a more general scenario and proposing the use of an unsupervised approach to learn part attention models (Fu et al. 2017; Zheng et al. 2017; Liu et al. 2017a; Zhang et al. 2016). Fu et al. (2017) and Zheng et al. (2017) constructed a recurrent attention convolutional neural network and a multi-attention convolutional neural network to localize the discriminative regions on objects based on unsupervised learning. For fine-grained feature learning, most of the recent recognition frameworks depend on convolutional networks (Fu et al. 2017; Zheng et al. 2017; Huang et al. 2016). Huang et al. (2016) proposed a part-stacked CNN architecture by



modeling the subtle differences between object parts. In our work, we design a method to extract local dynamic pose features, which are unique and effective characteristics for solving the FG person re-id problem.

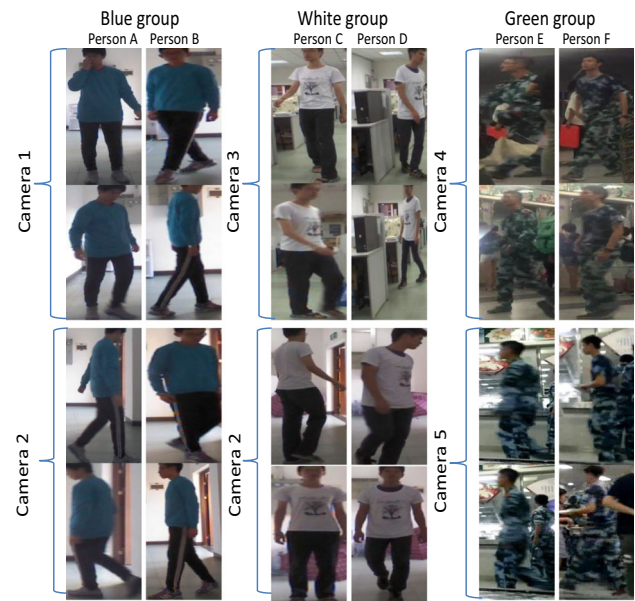
### 3 Fine-Grained Person Re-id (FGPR) Dataset

#### 3.1 Dataset Construction

To study the Fine-grained Person Re-id (FG re-id) problem, we have collected a new FGPR dataset. To the best of our knowledge, while there exist persons of similar clothing in existing datasets as pointed out by Gou et al. (2016), the FGPR dataset is the first benchmark for the FG re-id, which contains a number of persons who wear the same type of clothing but not just similar clothing. To obtain fine-grained data, we collected videos of people wearing uniform clothes. According to the clothes that the people wear, they are divided into three fine-grained groups, namely, “blue”, “white” and “green” groups. As shown in Figs. 3 and 4, people wear the same clothes in each fine-grained group. The identities in the blue group were captured by camera 1 and camera 2, the white group were captured in camera 2 and camera 3, and the persons in the green group only appeared in camera 4 and 5. The blue group and white group have one common camera view, and the green group has its unique two camera views. Hence, the number of camera views of FGPR dataset is 5. Note that, there is no overlapping on person identity between any two groups, and each identity has one sequence in the corresponding camera view. We clipped each video into frames and detected people by a person detector (He et al. 2017). For each identity, there are at least 150 consecutive frames. The “blue”, “white” and “green” groups contain 83,415, 31,691 and 19,590 images of 200, 45 and 113 identities, respectively. In total, our FGPR dataset includes 134,696 RGB images of 358 identities from 716 sequences. Note that there are more occlusions in the green group caused by other persons, while in the other two groups, the background is simpler. Thus, the green group is more challenging, and the performance is lower than the other groups. We present a comparison with existing re-id datasets in Table 1. Only our FGPR dataset contains fine-grained groups. Compared to common re-id, FG re-id is considerably more challenging since there is no clear difference in the style and color of clothes between different people.

#### 3.2 Evaluation Protocol

Following the protocols of other widely used re-id video datasets, like PRID2011 (Hirzer et al. 2011) and iLIDS (Wang et al. 2014), 10 train/test splits are conducted on the FGPR dataset. For each split, we randomly divided all of the 358 identities into 258 identities for training and 100



**Fig. 3** Examples of RGB images in our FGPR dataset. There are three groups: blue, white and green. Each group has two camera views. Every column is of the same identity, and each two rows are in the same camera view. The blue group and white group have one common camera view, and the green group has its unique two camera views. Hence, the number of camera views of the FGPR dataset is 5 (Color figure online)



**Fig. 4** Some examples of three groups. Compared with the blue and white groups, there are more occlusions in the green group which block the face, lower body or other body parts. Therefore, the green group is more challenging than the other two groups (Color figure online)

**Table 1** Comparison between the FGPR dataset and existing re-id datasets

Dataset	#ID	#Images	#Cameras	Tracklets	Evaluation	Fine-grained group
VIPER (Gray et al. 2007)	632	1264	2	–	CMC	No
CAVIAR (Cheng et al. 2011)	72	610	2	–	CMC	No
CUHK01 (Li et al. 2012)	972	1942	2	–	CMC	No
CUHK03 (Li et al. 2014)	1467	13,164	6	–	CMC	No
Market (Zheng et al. 2015)	1501	32,668	6	–	CMC + mAP	No
DukeMTMC-reID (Ristani et al. 2016)	1852	46,261	8	–	CMC + mAP	No
MSMT17 (Wei et al. 2018)	4101	126,441	15	–	CMC + mAP	No
iLIDS-VID (Wang et al. 2014)	119	42,459	2	600	CMC	No
PRID2011 (Hirzer et al. 2011)	200	35,942	2	400	CMC	No
MARS (Zheng et al. 2016)	1261	1,191,003	6	20,478	CMC + mAP	No
FGPR (proposed)	358	134,969	5	716	CMC + mAP	Yes

identities for testing. The numbers of testing identities for the “blue”, “white” and “green” groups are 60, 10 and 30, respectively.

In the training stage, all images of training identities obtained by the images clipped from the frames of all videos are utilized. In the testing stage, we have two settings: all-group setting and single-group setting. In the all-group setting, we evaluate the average performance of two cases. In the first case, since there are two cameras in each group, for each group, we select the videos from the first camera to form the gallery set and use videos from the second camera to form the probe set. In the second case, the cameras for the gallery set and probe set are exchanged. In the single-group setting, each group is individually tested in the same manner as in the all-group setting.

Given a probe sample, we perform the matching by computing the distance scores between the probe sample and all gallery samples and ranking the scores in ascending order to obtain a list of similar people. Then, we use the cumulative matching characteristic (CMC) and mean average precision (mAP) and compute the corresponding average result of 10 splits as our final performance.

## 4 Approach

Most existing person re-id methods assume that there are clear appearance differences between people for re-id. However, in the FG re-id problem, pedestrians in a group look very similar (e.g. dressing in very similar clothes), and thus, it is extremely difficult to distinguish them. The FG re-id problem has a considerably smaller interclass discrepancy of people wearing very similar clothes compared to conventional re-id, and it is difficult to completely overcome this problem by only using static appearance features. To overcome this problem, we consider that every person has his/her own unique

dynamic pose characteristics, especially local ones, in addition to clothing color. We therefore aim to solve this problem by exploring two types of dynamic pose features: motion-attentive and joint-specific local dynamic pose features in order to exploit discriminative features suitable for FG re-id.

### 4.1 Preliminary

Our development relies on two types of features, namely, global appearance features and motion features.

- *Global appearance feature* Global appearance information is the basic cue of the other streams in our network, and it can describe the color, texture and other abstract features of images. For each video sequence, we sampled segments, each of which lasts 10 frames long, as the inputs. All segments are fed into the global appearance stream and the global appearance features are obtained by averaging the features. In the global appearance stream, we take ResNet50 (He et al. 2016) as the basic network, and we extract global appearance feature maps of different levels, denoted as  $F_g = \{f_{g1}, \dots, f_{gn}\}$ , from each block in ResNet50 (He et al. 2016), where  $n$  is the sequence length.  $F_g$  will be fed into the motion-attentive local dynamic pose stream, which is introduced in the following. A cross entropy loss  $loss_{global}$  is applied to learn the global appearance features.
- *Motion feature* The motion feature that describes the moving patterns of a person is less relevant to the appearance of a person, and thus, it is suitable for distinguishing people who are wearing similar clothes. In the motion stream, we apply the *optical flow guided feature (OFF)* (Sun et al. 2018) as the motion feature. Two consecutive input segments, which belong to the same sequence but have a time delay of  $\Delta t$ , are separately fed into ResNet50 (He et al. 2016) to obtain different levels of feature maps

from different blocks, and the OFFs of two feature maps of two segments at the same level will be computed. In particular, each input segment of the appearance stream is corresponding to the segment of the motion stream of the same time step. The details can be found in Sun et al. (2018). A cross-entropy loss  $loss_{motion}$  is applied to learn the motion features.

In addition to obtaining the final fused motion feature, the motion feature maps of  $n$  different levels, denoted as  $F_m = \{f_{m1}, \dots, f_{m_n}\}$ , are extracted and fed into the motion-attentive local dynamic pose stream, which is introduced in the following.

## 4.2 Learning Local Attentive Dynamic Pose Features

In this section, for extracting discriminative local dynamic pose characteristics, we introduce two types of dynamic pose features: motion-attentive and joint-specific local dynamic pose features.

- *Motion-attentive local dynamic pose feature* Note that while the appearance may be similar, different people have different behaviors and walking styles, which can provide useful discriminative information for identification to avoid the interference caused by a similar appearance. Therefore, we speculate that the variant caused by motion of a person can be more useful and provide better discrimination than the appearance corresponding to no motion or less motion change. Thus, we first consider the motion-attentive local dynamic pose features, which utilize motion information to localize the sensitive moving parts in the feature map of global appearance. Since the motion within a detected person bounding box is mostly related to the human body, such as the head, torso and legs, we expect that cases in which a human body has a large motion variation will also generate more identity-specific pose variations, which are locally useful for distinguishing one person from another. Based on this assumption, we design motion-attentive local dynamic pose stream to learn the motion-attentive local dynamic pose features, as shown in Fig. 5. This network includes two parts, namely, *RNN-mask network* and *localization network* (LN), which attempt to find and extract the moving parts in global appearance feature maps, respectively.

On one hand, the RNN-mask network, which consists of five convolution layers for each feature map of different level and a LSTM network, aims to find the significant area by utilizing the extracted motion features. As shown in Fig. 5, the motion features denoted by  $F_m = \{f_{m1}, \dots, f_{m_n}\}$  are firstly

processed by a convolution operation to reduce the channel dimension. In the following, an RNN network takes these different levels of feature maps as inputs to generate fine-to-coarse masks, denoted as  $M = \{m_1, \dots, m_n\}$ , for finding the moving part. The procedure can be described as follows:

$$m_i = R_i(\phi(f_{m_i}), h_{i-1}), \quad i = 1, \dots, n \quad (1)$$

where  $\phi$  indicates the convolution operation to normalize the size of motion features,  $R_i$  denotes the unit of RNN of the  $i$ -th level, and  $h_{i-1}$  denotes the hidden layer output of RNN of the previous level. We feed different level features into the RNN for extracting motion-attentive information from a deeper layer based on the information from shallower layers. The value in the mask is between 0 and 1.

Note that we have indicated the final output of the RNN network by  $m_f$ , which is the coarsest motion-attentive information. For obtaining finer localization results, we add a finer output of the  $i$ -th level to the  $m_f$  in order to refine the localization results as follows:

$$m'_i = \alpha * m_i + (1 - \alpha) * m_f, \quad i = 1, \dots, n \quad (2)$$

where  $m'_i$  is the mask, which is considered as attention, after refinement,  $\alpha$  is a hyperparameter, and  $*$  denotes scalar multiplication.

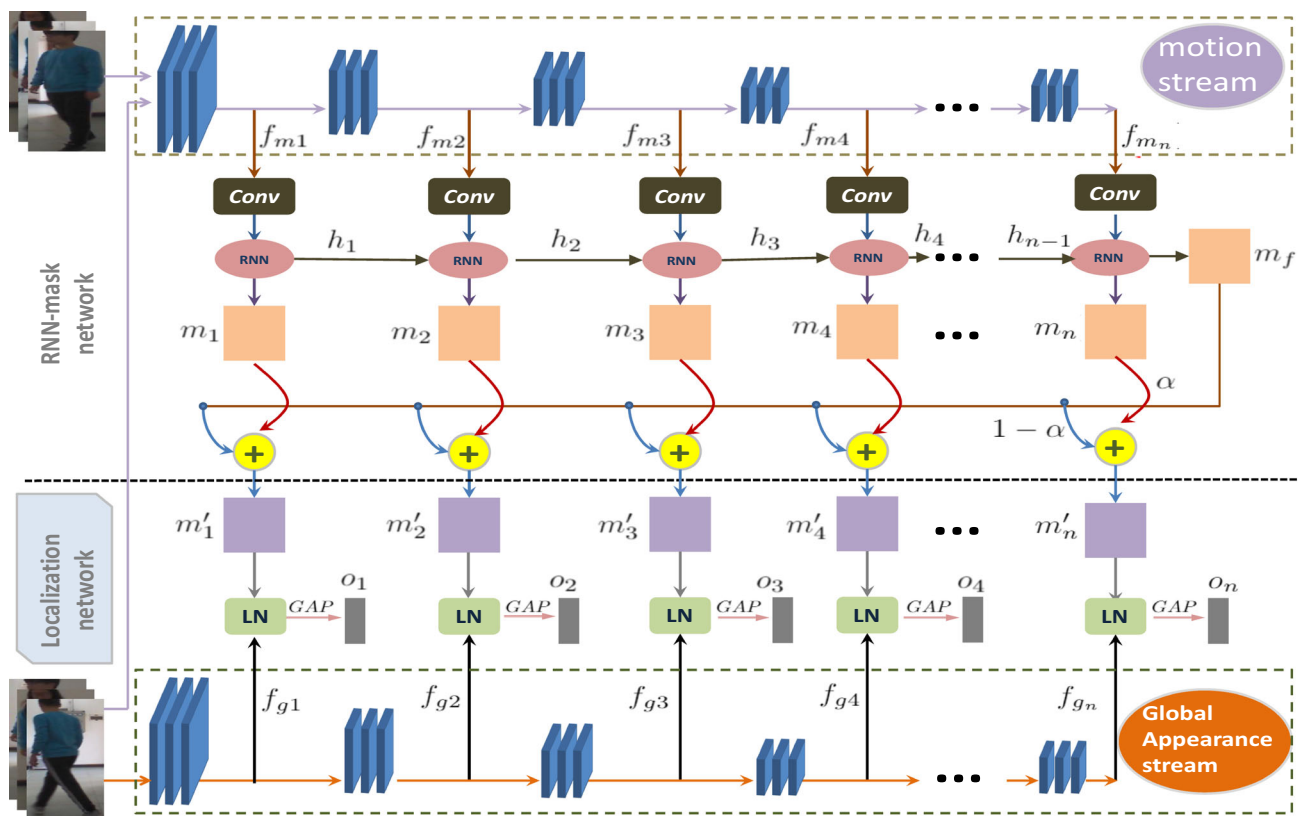
On the other hand, the localization network exploits the masks generated by the RNN-mask network to extract the discriminative regions in different levels of global appearance feature maps, denoted  $F_g = \{f_{g1}, f_{g2}, \dots, f_{g_n}\}$ . First, all the masks are fed into the localization network, denoted as LN in Fig. 5, including upsampling operations and element-wise multiplication operations to localize these regions by the following formula:

$$f_i = s(m'_i) \otimes f_{gi}, \quad i = 1, \dots, n \quad (3)$$

where  $s$  is the upsampling operation,  $f_{gi}$  indicates the feature map of the global appearance stream,  $f_i$  is the feature map after localization and  $\otimes$  denotes element-wise multiplication to localize the moving parts of human body.

After obtaining a series of fine-to-coarse feature maps of discriminative areas ( $f_i$ ), a global average pooling layer follows for each feature map, denoted as  $o_i$ ,  $i = 1, \dots, n$ . We concatenate them to form the final feature vector that can be matched with the identity entries, where the fully connected layers are quantified by a Softmax layer along with the loss function  $loss_{MA}$ .

Through such modeling, as shown in Fig. 10, the proposed motion-attentive local dynamic pose stream will learn an interesting local part of a body for discrimination.



**Fig. 5** The architecture of the motion-attentive local dynamic pose stream: First, we feed the first segment into the global appearance stream to obtain global appearance feature maps; then, the second segment, which has a time delay  $\Delta t$  to the first segment, is fed to the motion stream together with the first segment. The motion-attentive local dynamic pose stream takes the different-level feature maps of the global appearance stream and motion stream as inputs. The motion features  $f_{m1}, \dots, f_{m_n}$  are processed by an RNN to generate a series of masks, which we also called attentions, that aim to localize the discriminative area indicated by RNN in this figure. The symbols  $m_1, \dots, m_n$  denote the generated masks, and the symbol  $m_f$  denotes the last mask, which was illustrated in Eq. 2. The symbols  $m'_1, \dots, m'_n$  indicate the refined masks. Mean-

while, the global appearance feature maps  $f_{g1}, \dots, f_{g_n}$  and the masks  $m'_1, \dots, m'_n$  are fed into the localization subnetwork (denoted as LN in the figure) to localize significant regions, as illustrated in Eq. 3. Finally, the outputs of all the localization subnetworks will be fed into their corresponding global average pooling layer to attain motion-attentive local dynamic pose features at different levels, namely,  $o_1, o_2, \dots, o_n$ . These features are concatenated as the final pedestrian feature. For the localization network (LN), it includes both upsampling and element-wise multiplication operations. Note that the inputs of the motion stream are two consecutive segments that belong to the same sequence but have a time delay  $\Delta t$

- *Joint-specific local dynamic pose features* As complementary to the motion-attentive local dynamic pose feature, we further append a joint-specific local pose stream to extract local dynamic pose features for FG reid in order to help alleviate the effect of noise in the motion feature map caused by non-human moving that would bias the motion-attentive network learning. Such modeling is expected to enhance the robustness of local dynamic pose features.

The joint-specific local dynamic pose stream consists of three modules: (1) *pose joint local region extraction*, (2) *Fine-Grained Feature subNetwork (FG-FN)*, and (3) *Fine-Grained Feature Union Learning subNetwork (FG-ULN)*. Figure 6 shows the network structure for extracting joint-specific local dynamic pose features. Given an image, we

first locate 16 joints of the human body using the human pose estimation algorithm (Cao et al. 2017). The 16 joints are the following: *head, neck, left eye, right eye, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, left hip, left knee, left ankle, right hip, right knee, and right ankle*. We use five rectangular regions to cover the local body parts of a person, including the head, torso, left arm, right arm and lower body. By utilizing the coordinates of skeleton joints, we combine the left arm and right arm as one region and obtain four regions (head, body, hands and legs), as shown in Fig. 6.

Each region that we exploit is based on the position of the joint, and thus the joint-specific local features are more robust to the motion variation of a person. Then, the regions of the left arm and right arm are concatenated as one part of the image, and the other three regions are used individually



as part images. In this work, we resize all the part images to  $224 \times 224$ . The four parts are fed into the FG-FN network, as shown in Fig. 6. Each of them will be convolved by each CNN, and these CNNs have the same structure (including five convolution layers, BN layers, Relu layers and one average pooling) but do not share the weights. The extracted part-based features are then fused by the FG-ULN network. FG-ULN is composed of two convolutional layers, which aim to learn the connection between different parts through the first layer and perform dimensionality reduction through the second layer. For quantifying joint-specific local pose features, we utilize softmax classification loss, denoted as  $loss_{JS}$ . In both the training and inference stages, we perform average pooling on frame-level local pose features from the same video to obtain dynamic features.

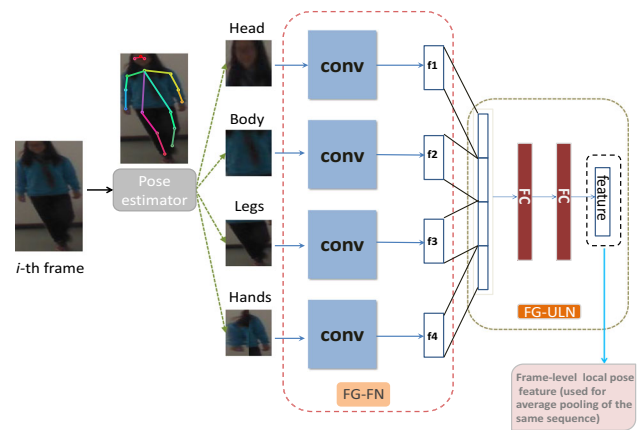
**Remarks** Learning both types of local features (i.e., motion-attentive local dynamic pose features and joint-specific local dynamic pose features) is necessary. For example, when some joint localizations are inaccurately detected due to the low resolution of images, the information from motion-attentive local dynamic pose features that partially consist of local dynamic pose features around joints become complementary. Similarly, the joint-specific local dynamic pose features will help enhance motion-attentive local dynamic pose features extracted around joints to alleviate the effect of noise extracted in the motion feature map. In this work, the motion-attentive local dynamic pose feature and the joint-specific local dynamic pose feature are called the local attentive dynamic pose features.

### 4.3 A Multiple Stream Network: Fusion of Global Appearance Feature and Motion Feature

In addition to extracting local attentive dynamic pose features, we extract global appearance feature via ResNet and motion features as a complement that is expected to describe the global shape and motion information of a person, as shown in Fig. 7. Therefore, based on the extracted global appearance and motion features, our solution to the FG re-id problem is to jointly learn all these features with the local attentive dynamic pose features including the motion-attentive local dynamic pose feature and the joint-specific local dynamic pose feature. In this manner, the local attentive dynamic pose features will be updated dynamically along with learning motion and global appearance features. The training strategy is illustrated in detail in Sect. 5.1.

When re-identifying the identities, we combine the output distances between different subjects  $O_{c_1}$  and  $O_{c_2}$  of each branch as follows:

$$Dis = Dis_{global}(O_{c_1}, O_{c_2}) + Dis_{local}(O_{c_1}, O_{c_2}) + Dis_{motion}(O_{c_1}, O_{c_2}) \quad (4)$$



**Fig. 6** The architecture of the joint-specific local dynamic pose stream. Given the  $i$ -th frame in a human sequence, we extract different discriminative regions with the pose estimation algorithm (Cao et al. 2017) and feed four different key parts of a human image into our fine-grained feature subnetwork (FG-FN) to obtain four local part features, which are independent of each other. Finally, these features are fed into the fine-grained union learning subnetwork (FG-ULN) to jointly learn the frame-level local pose features, which are used to form joint-specific local dynamic pose feature via the average pooling operation. Note that we combine the regions of left arm and right arm as the hand's region

where  $Dis_{global}$ ,  $Dis_{local}$  and  $Dis_{motion}$  are the distances between different subjects, computed via the Euclidean distance, corresponding to global appearance features, local attentive dynamic pose features and motion features, respectively. The distance of local attentive dynamic pose features,  $Dis_{local}(O_{c_1}, O_{c_2})$ , an addition summation of the following distance: the distance associated with motion-attentive local dynamic pose feature and the one associated with joint-specific local dynamic pose feature.

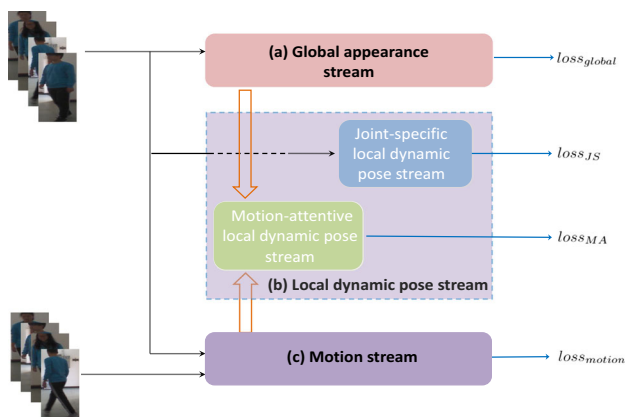
## 5 Experiment

### 5.1 Implementation Details

- *Implementation of our method* We implemented the whole network on the PyTorch framework. For the global appearance stream, we chose ResNet50 (He et al. 2016) as the base module which is initialized in ImageNet as usual. For the motion stream, the base module is initialized the same as the global appearance stream, and we set the time delay  $\Delta t$  to 5. The learning rate of the global appearance stream, motion stream and joint-specific local dynamic pose stream was set to 0.01, and for the motion-attentive local dynamic pose stream, it was set to 0.001. All RGB images were resized to  $224 \times 224$ . In this work, we set  $n$ , the number of attentions, to 5.

To better train our network, we implemented the following strategy. First, we pretrained the global appearance stream and joint-specific local dynamic pose stream by minimiz-





**Fig. 7** The architecture of our proposed method. Global appearance and motion features are obtained by the global appearance stream (a) and motion stream (c), respectively. The local dynamic pose stream consists of joint-specific local dynamic pose stream and motion-attentive local dynamic pose stream. The motion-attentive local dynamic pose stream in (b) contains an RNN-mask network and localization network, and we take the feature maps of the global appearance stream and motion stream as inputs. Joint-specific local pose feature is extracted by the joint-specific local dynamic pose stream in (b). Note that there are global average pooling layers at the end of (a) and (c) to obtain the features. Finally, for each stream, their corresponding loss will be computed, denoted  $loss_{global}$ ,  $loss_{JS}$ ,  $loss_{MA}$  and  $loss_{motion}$ , to train the corresponding stream. Note that the inputs of motion stream are two consecutive segments that belong to the same sequence but have a time delay  $\Delta t$

ing the corresponding loss function on the introduced FGPR dataset, with the batch size equal to 150 and epoch equal to 50. Second, we utilized the global appearance stream as the basic module and took the global appearance feature maps as input to train the motion stream. Third, by fixing the global appearance, joint-specific local dynamic pose and motion streams, we optimized the motion-attentive local dynamic pose stream using cross-entropy loss. Finally, we fine-tuned the whole network using the sum of the losses of all streams. For the second, third and final steps, we set both the batch size and epoch to 50.

Note that we do not employ other person re-id datasets to pretrain any part of our model before training on FGPR.

- *Testing* In the test stages, as indicated in Sect. 3.2, we have an all-group setting and a single-group setting. In the all-group setting, we evaluated the average performances of two cases. In the first case, we selected the videos from the first camera to form the gallery set and used videos from the second camera to form the probe set. In the second case, the cameras for the gallery set and probe set used in the first case were exchanged. The average performance is reported for the all-group setting. In the single-group setting, each group was individually tested in the same way as in the all-group setting. All the compared methods were tested based on the video-based

re-id protocol. For the conventional re-id model, we first extract the feature of each single video frame image, and then we obtain the sequence feature via the average pooling operation; for the metric learning method, we use the Euclidean distance to measure the distance between a pair of feature points.

## 5.2 Comparison with Related Work

- *Comparison with conventional re-id models* We compared our approach with conventional re-id models, including LBP (Guo et al. 2010), HOG (Dalal and Triggs 2005), LOMO (Liao et al. 2015) and GOG (Matsukawa et al. 2016). These methods mainly extract texture and color features. Table 2 shows that our model has a substantial improvement by a large margin, i.e., 60.8% on Rank 1 matching rate matching and 56.5% on mAP, compared with these models in the all-group setting. On the three single groups, our model has a great improvement, as reported in Table 2, which suggests that the conventional features are not effective for distinguishing people who are dressed in similar clothes since they pay more attention to quantifying appearance information such as color and texture.
- *Comparison with deep image-based re-id models* We compared our method with deep image-based re-id models, including VGG16 (Simonyan and Zisserman 2014), ResNet50 (He et al. 2016), JSTL (Xiao et al. 2016) and PCB (Sun et al. 2017). VGG16 (Simonyan and Zisserman 2014) and ResNet50 (He et al. 2016) are common CNNs, and JSTL (Xiao et al. 2016) is a type of joint learning method. For VGG16 (Simonyan and Zisserman 2014) and ResNet50 (He et al. 2016), we trained the models on the FGPR dataset directly. For JSTL (Xiao et al. 2016), we first trained ResNet50 (He et al. 2016) on a set of re-id datasets, including CUHK01 (Li et al. 2012), CUHK03 (Li et al. 2014), iLIDs (Wang et al. 2014), VIPER (Gray et al. 2007), and Market (Zheng et al. 2015), for obtaining a pretrained model. Then, we fine-tuned it on the FGPR dataset. We also evaluated the performance of PCB (Sun et al. 2017), which is a part-based model. During testing, all images of a gallery person sequence were used to form the gallery set, and all images of a query sequence were used as probe images. In the all-group setting, Table 2 shows that our proposed model surpassed PCB (Sun et al. 2017), which also learns local features and is a state-of-the-art model among other compared deep image-based re-id models, by 24.7% on Rank 1 matching rate and 18.3% on mAP. The local features extracted by PCB are not dynamic and focus more on static appearance, and thus our proposed model explores more suitable local

**Table 2** Performance (%) comparison with related work on FGPR

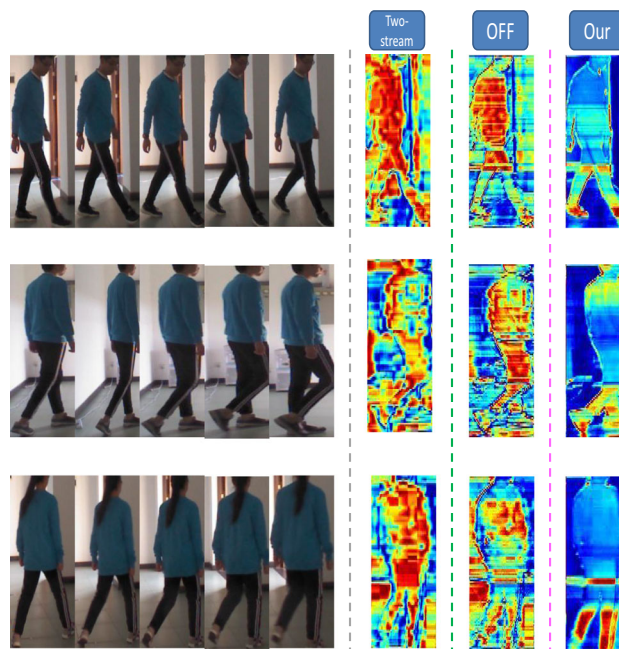
Methods	All groups			Blue group			White group			Green group		
	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP
LOMO (Liao et al. 2015)	15.7	28.3	24.6	14.2	22.1	21.4	11.0	62.0	33.9	21.6	55.9	34.2
HOG (Dalal and Triggs 2005)	24.6	39.2	32.1	12.9	22.9	15.9	21.0	59.0	38.6	44.8	63.1	54.9
LBP (Guo et al. 2010)	23.9	36.4	31.7	14.2	23.1	20.1	34.0	65.0	45.1	40.1	63.3	52.4
GOG (Matsukawa et al. 2016)	26.3	43.8	31.9	33.4	52.4	42.4	36.0	78.0	46.0	46.5	67.2	58.3
VGG16 (Simonyan and Zisserman 2014)	56.3	83.1	64.8	62.3	89.7	72.6	52.0	86.0	62.9	53.1	80.9	65.1
JSTL (Xiao et al. 2016)	53.2	80.4	62.4	59.2	88.1	71.3	53.0	90.0	61.1	53.7	78.6	65.7
ResNet50 (He et al. 2016)	57.3	84.7	66.8	62.9	83.3	76.1	48.0	79.0	61.5	51.3	76.3	64.2
PCB (Sun et al. 2017)	62.4	89.2	70.1	67.2	89.5	78.0	57.0	91.0	52.7	52.1	80.7	64.3
Part-stacked CNN (Huang et al. 2016)	64.9	88.2	74.1	70.8	89.1	81.2	44.0	85.0	59.3	55.7	82.4	66.2
RA-CNN (Fu et al. 2017)	61.5	85.9	70.9	69.9	87.4	74.1	37.0	87.0	48.4	52.0	82.6	69.4
MA-CNN (Zheng et al. 2017)	64.3	87.6	71.4	70.3	89.6	79.4	39.0	91.0	53.7	57.4	83.1	70.1
Two-stream (Chung et al. 2017)	67.2	89.8	72.8	71.2	90.3	81.1	54.0	94.0	61.4	52.5	80.4	66.1
OFF (Sun et al. 2018)	79.3	92.9	82.4	84.7	92.9	89.3	85.0	100	92.4	60.4	84.8	70.7
DSEPA (Li et al. 2018a)	80.1	93.1	84.2	87.1	94.1	91.4	87.0	100	93.1	64.1	86.9	75.1
JASTPN (Xu et al. 2017)	82.4	92.6	84.7	88.7	94.1	91.8	89.0	99.0	94.9	64.3	86.8	74.3
STMP (Liu et al. 2019)	83.7	93.9	86.1	92.1	94.7	92.1	91.0	100	96.7	63.4	85.9	75.4
GLAD (Wei et al. 2017)	80.6	93.3	83.4	87.9	94.4	91.5	86.0	99.0	90.6	64.3	87.1	75.3
SpindleNet (Zhao et al. 2017)	81.4	93.7	83.7	88.3	94.7	91.9	88.0	100	94.2	63.9	87.0	74.8
PDC (Su et al. 2017)	82.1	94.2	84.5	89.6	95.0	92.3	90.0	100	96.9	64.5	87.2	75.8
FD-GAN (Ge et al. 2018)	83.4	94.3	85.7	90.8	95.1	92.5	92.0	100	97.5	65.7	87.4	75.6
Our model	87.1	95.2	88.4	93.6	97.2	92.6	99.0	100	99.2	67.6	87.4	76.1

dynamic pose features. In the single-group setting, our model also achieved leading performance compared with these networks. The results indicate that our proposed deep model better addresses the FG re-id problem.

- *Comparison with video-based re-id models* We compared our method with recent video-based re-id models for which code is available, including two-stream networks (Chung et al. 2017), OFF (Sun et al. 2018), DSEPA (Li et al. 2018a) JASTPN (Xu et al. 2017) and STMP (Liu et al. 2019). Table 2 indicates that our method outperformed the best video-based re-id model that we compared by 3.4% matching rate on Rank 1 matching and 2.3% on mAP in the all-group setting, and in some case (e.g. the While Group), our method achieves 8% more on Rank 1 matching rate. Note that among the other compared methods, OFF (Sun et al. 2018) and DSEPA (Li et al. 2018a) extract not only global appearance features or spatial features but also motion information or temporal information, which is more useful for FG re-id to avoid interference caused by similar appearance. And JASTPN (Xu et al. 2017) also utilizes attention for pooling features after extracting global and motion information. STMP (Liu et al. 2019) designs RRU to recover the missing parts and suppress noisy parts. They achieved a better performance than other compared related methods. Note that when the optical flow information extracted by the two-stream network (Chung et al. 2017) is added as motion cues, a better performance is achieved by the two-stream network (Chung et al. 2017) compared to the conventional re-id approaches [LBP (Guo et al. 2010), HOG (Dalal and Triggs 2005), LOMO (Liao et al. 2015), GOG (Matsukawa et al. 2016)] and image-based re-id methods [VGG16 (Simonyan and Zisserman 2014), ResNet50 (He et al. 2016), JSTL (Xiao et al. 2016), PCB (Sun et al. 2017)], although its results are poorer than those of OFF (Sun et al. 2018) and DSEPA (Li et al. 2018a).

In our network, the extracted local attentive dynamic pose features further focus on the pose change caused by human motion. In the single-group setting, our model achieved considerably better performance on the white group specifically since our model explores effective local discriminative pose features. The results reported in Table 2 suggest that local attentive dynamic pose features can help to clearly improve the performance by utilizing the more identity-specific dynamic pose characteristics of different people, even though they have very similar appearances.

Additionally, we further show the visual results to validate the advantage of our model for solving the FG re-id problem in Figs. 8 and 10. The matching results of our method and the two-stream network (Chung et al. 2017) are presented in Fig. 9, where we used ResNet50 (He et al. 2016) as the



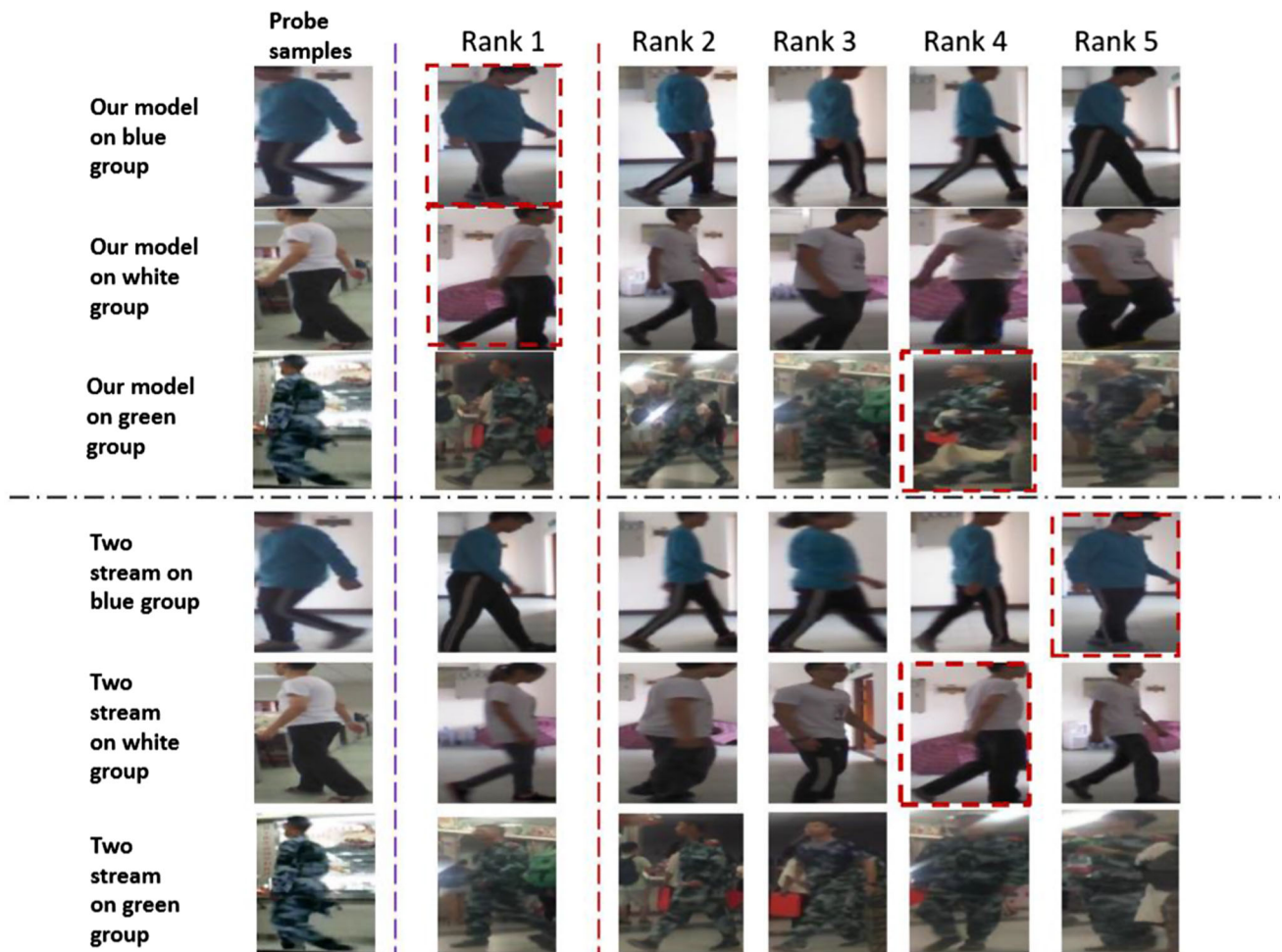
**Fig. 8** The feature maps of the motion-attentive local dynamic pose stream and the feature maps of the two-stream (Chung et al. 2017) and OFF (Sun et al. 2018). We present the image sequence of three different people in the first five columns. The last three columns are the feature maps of two-stream (Chung et al. 2017), OFF (Sun et al. 2018), and our model, respectively

backbone of the two-stream network (Chung et al. 2017) for a fair comparison. In particular, some failure cases are shown in Fig. 9, and these cases are probably due to the object occlusion, the illumination change and the low resolution in green group. Based on local dynamic pose features and motion features, our model still matched the query image within the top five images. After we add the attentions to the global appearance feature maps, the finer features can help our model better distinguish different people who have very similar appearances.

- *Comparison with pose-driven models* We compared three pose-driven methods, including GLAD (Wei et al. 2017), SpindleNet (Zhao et al. 2017) and PDC (Su et al. 2017), and we present the results in Table 2. Our method outperforms the best one among the three approaches by 5.0% on Rank 1 matching rate. The three compared methods only utilize the still pose information to extract still local features; and in comparison, our method particularly extracts the local dynamic pose feature by the guidance of motion cue.

In addition, we compared a pose-normalization methods FD-GAN (Ge et al. 2018), which utilizes GAN to generate a series of images possessing the same pose for data augmentation. FD-GAN outperformed the other three pose-driven





**Fig. 9** The matching results of our method and two-stream framework. We present the ranking images of the blue group and white group. The image with the red box is the correct identity (Color figure online)

methods but is inferior to ours—about 3.7% lower on Rank 1 matching rate.

- *Comparison with generic fine-grained models* We also compared our method with generic fine-grained models, including Part-stacked CNN (Huang et al. 2016), RA-CNN (Fu et al. 2017) and MA-CNN (Zheng et al. 2017), all of which have learned fine-grained features, where we used tResNet50 (He et al. 2016) as the backbone network of Huang et al. (2016), Fu et al. (2017), and Zheng et al. (2017) for a fair comparison. Table 2 indicates that our model gained a 22.2% matching rate improvement in terms of the Rank 1 matching rate and a 14.3% matching rate improvement in terms of the mAP compared to Part-stacked CNN (Huang et al. 2016), which achieved the best performance among the compared fine-grained models in the all-group setting. In the single-group setting, our model obtained better matching results, particularly on the white and green groups. This result is probably

mainly because the compared fine-grained models were designed for generic object recognition, but they have not found that dynamic pose features are important for overcoming the challenge of fine-grained recognition, and thus, no effective model for solving this problem has been previously presented in the literature. In this work, we mainly learn a deep fine-grained recognition network with motion-attentive local dynamic pose feature modeling and joint-specific local dynamic pose feature learning, and thus our model gains merits.

### 5.3 Ablation Study

- *The effect of each stream* We present the results when a branch is eliminated in Table 3. When using only appearance features (i.e., using global appearance stream), a significant performance degradation of (19.3% in terms of the Rank 1 matching rate and 18.6% in terms of the mAP matching rate in the all-group setting) compared to



**Table 3** Performance (%) of different components of our method on FGPR

Network	All groups			Blue group			White group			Green group		
	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP
Global appearance stream	67.8	83.2	69.8	73.2	93.4	80.1	53.0	85.0	64.5	53.4	79.6	64.7
Motion stream	75.1	90.7	80.6	84.3	93.9	89.5	71.0	94.0	86.4	61.4	84.2	70.5
Local dynamic pose stream	84.2	93.1	85.1	86.2	95.4	91.9	84.0	98.0	94.1	63.8	86.8	74.5
Full model	87.1	95.2	88.4	93.6	97.2	92.6	99.0	100	99.2	67.6	87.4	76.1

**Table 4** Evaluation performance (%) of local attentive dynamic pose features on FGPR

Network	All groups			Blue group			White group			Green group		
	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP	Rank 1	Rank 5	mAP
Without motion-attentive feature	80.4	91.2	80.2	88.4	92.6	90.7	73.0	87.0	78.2	62.9	81.4	71.5
Without joint-specific feature	82.1	93.3	84.7	91.5	94.5	91.3	85.0	94.0	85.2	66.8	86.3	75.9
Full model	87.1	95.2	88.4	93.6	97.2	92.6	99.0	100	99.2	67.6	87.4	76.1

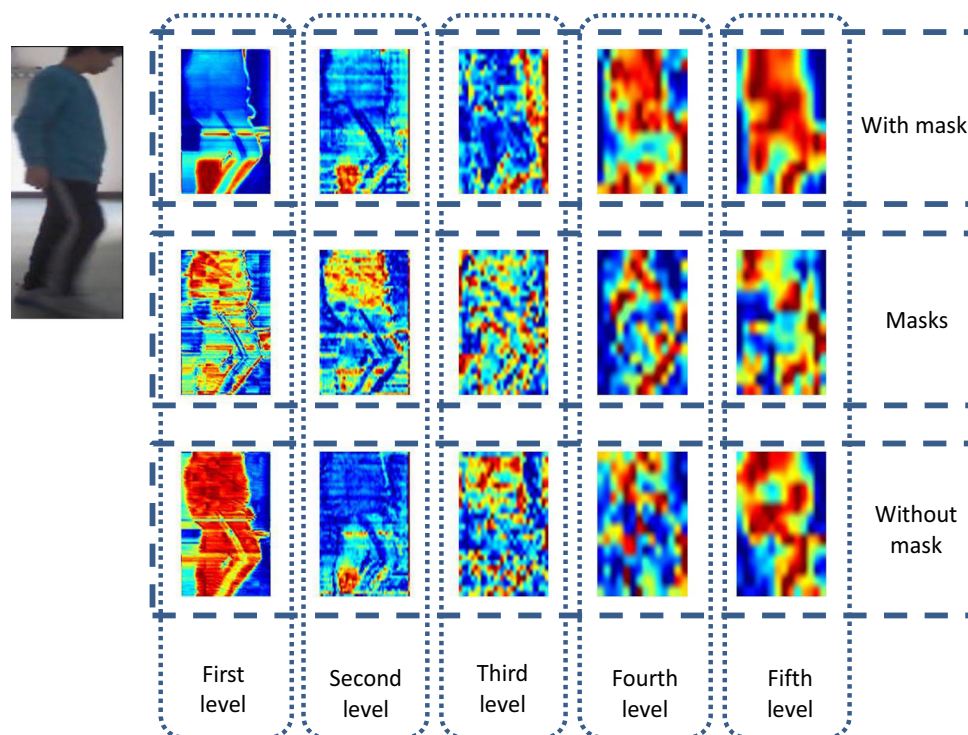
our Full Model is observed. This result indicates that the global appearance feature is not discriminative enough due to the similar appearances of different people globally. In particular, the Rank 1 rate decreased by 46.0%, and the mAP matching rate decreased by 34.7% in the white group setting. By comparing our Full Model with that using only motion features (i.e., motion stream), it is clear that motion features are much more effective than global appearance features, as our Full Model achieved improvements of 12.0% in terms of the Rank 1 matching rate and 7.8% in terms of the mAP matching rate in the all-group setting and improvements of 9.3% in terms of the Rank 1 matching rate and 9.3% in terms of the mAP matching rate in the blue group setting. However, the motion information is not sufficient to distinguish people who dress very similarly due to lack of modeling more fine-grained discriminant features. The local dynamic pose stream that extracts the local attentive dynamic pose features provides more identity-specific and robust important visual cues to distinguish different people who look very similar, and thus, a clear and notable improvement of approximately 16.4% in terms of the Rank 1 matching rate and 15.3% in terms of the mAP matching rate is observed compared with the global appearance stream in the all-group setting. Note that although the “Local Dynamic Pose Stream” is based on the global appearance stream and motion stream, the features extracted from the global appearance stream and motion stream are not used when we test the “Local Dynamic Pose Stream”.

- *The effect of modeling local attentive dynamic pose features* The local dynamic pose stream extracts local attentive dynamic pose features, including the motion-attentive local dynamic feature and joint-specific local

dynamic feature. We evaluate their effect in Table 4. We find that when the motion-attentive local dynamic features are removed in the all-group setting, the performance decreased by 6.7% in terms of the Rank 1 matching rate and 8.2% in terms of the mAP. In the single-group setting, the performance of the white group had a greater decrease of 26.0% in terms of the Rank 1 matching rate and 21.0% in terms of the mAP compared with the case for the blue group and green group. The results show that the proposed motion-attentive network can select more discriminative parts for distinguishing people who dress similarly. We present the feature map of the motion-attentive network in Fig. 10. We find that the motion-attentive network can focus on the moving part and learn motion-attentive local dynamic pose feature. We present some results of the motion-attentive network in Fig. 10. After adding the attentions, our network can pay more attention to the shape of a person and the moving parts; thus, the attention helps seek more discriminative dynamic pose features for solving the FG re-id problem.

In addition to the analysis of motion-attentive network, when the joint-specific dynamic pose feature network is removed from our model, the Rank 1 matching rate decreased by 5.0% and the mAP matching rate decreased by 3.7% in the all-group setting. In the single-group setting, the performance of the three groups all decreased, and the performance of the white group decreased by 14.0% on Rank 1 matching rate and 14.0% on mAP. These results show that the joint-specific local dynamic pose feature is also useful, and it is complementary to the motion-attentive one.

**Fig. 10** Comparison between the feature maps after adding attentions (masks) and the feature maps before adding attentions (masks), where we scale the feature map to the size  $196 * 128$  for better illustration. We also show the attentions of five levels indicated as ‘Masks’. After adding the attentions, in the first-level feature map, our network can easily find the shape of the person, and the moving parts are found in the first- and second-level feature maps (highlighted regions). For the third, fourth and fifth feature maps, our masks make the discriminative region more highlighted



- *On the fine-to-coarse attention* In our method, the fine-to-coarse attentions are applied. Now, we quantify the progress in Tables 5 and 6.

Firstly, we compared the effect of such attention when adding more coarse attentions step by step in Table 5. When we impose the first level attention, our result gained a 0.6% matching rate improvement on Rank 1 matching rate and 1.7% matching rate improvement on mAP. After adding more deeper layer features guided by more coarse attentions, the matching rate experienced an overall improvement of 4.8% on Rank 1 matching rate rate and 3.7% on mAP. The fine-to-coarse attention can be also called shallow-to-deep attention, since the attention focused on different level features. When features of more levels are incorporated, the motion-attentive local dynamic pose feature is more discriminative.

Secondly, as shown in Table 6, we evaluated the performance of individual layer. The results show that using attentions of all levels is clearly better than using attention of an individual level. Although it might be hard to tell which level of attention is better in theory, it is found that fusing the attention features of all levels makes the proposed model more stable and discriminative.

To further evaluate the effectiveness of our fine-to-coarse strategy, we compared it with the coarse-to-fine strategy, which has the inverse order of RNN to generate the attentions. As shown in Table 7, the result of the coarse-to-fine strategy is 2.8% inferior to our fine-to-coarse strategy on rank 1 matching rate and 3.3% lower on mAP. Since the fea-

**Table 5** Performance (%) of using different numbers of attention features on FGPR

The number of added attentions	All groups		
	Rank 1	Rank 5	mAP
0	82.3	92.8	84.3
1	82.9	92.4	85.4
2	84.3	93.7	86.9
3	84.2	93.4	87.1
4	85.4	94.2	88.4
5	87.1	95.2	88.4

tures of a neural network from shallower layer to deeper layer are intrinsically from low-level (fine) to high-level (coarse), it is therefore also a natural way to apply the fine-to-coarse strategy; and the results also validate that the fine-to-coarse strategy is more effective.

- *Selection of attention* To validate the effectiveness of our motion-guided attentions, we evaluated fine-to-coarse soft attentions without motion as a baseline (i.e., the Appearance-guided attentions in Table 8) and the CNN-guided attentions in Table 8. For the Appearance-guided attentions, we replace the inputs of motion features for RNN network with the feature maps of the global appearance stream. The specific operation is as follows: firstly, we utilize the model of global appearance stream and obtain five level feature maps of a video sequence with

**Table 6** Performance (%) of using individual attention features on FGPR

Index number of attentions	All groups		
	Rank 1	Rank 5	mAP
1	82.9	92.4	85.4
2	84.8	93.1	86.7
3	83.1	92.7	85.5
4	85.2	93.9	86.9
5	83.5	92.8	85.9
All attentions	87.1	95.2	88.4

**Table 7** Performance (%) of using individual attention features on FGPR

Networks	All groups		
	Rank 1	Rank 5	mAP
Coarse-to-fine strategy	84.3	94.7	85.1
Fine-to-coarse strategy (ours)	87.1	95.2	88.4

the length of 10. Then, these feature maps are fed into the RNN-mask network to learn a series of attentions. Finally, we localize the discriminative moving part by applying the generated attentions to the global appearance feature maps. The results are shown in Table 8. For the CNN-guided attentions, we have a similar operation by replacing the RNN network with CNN to generate attentions. When using the CNN-guided attentions, the results is 5.8% lower on the Rank 1 matching rate and 5.7% on mAP. And the result of using appearance-guided attentions is 3.5 % lower on the Rank 1 matching rate and 2.8% lower on mAP. The results indicate that our motion guided attention scheme is more effective to guide our model for learning more robust local dynamic pose features.

- *Analysis of the mask weight  $\alpha$  in the RNN-mask network* To explore the sensitivity of the mask weight in Eq. 2, we varied the value of from 0 to 1 with the interval 0.1. From Table 9, as the increase of  $\alpha$ , the Rank 1 matching rate decreases. When  $\alpha$  equals to 0.2 which is the default value, our method achieves the best performance on Rank 1 matching rate. Although  $\alpha$  is not sensitive, setting it appropriate small still refine the location result and this makes benefit.
- *Analysis of parameter of loss weighting* We also explore the sensitivity of the weighting parameters. Our loss function in the final fine-tune step is described as:

$$\begin{aligned} loss = loss_{global} + w_1 * (loss_{JS} + loss_{MA}) \\ + w_2 * loss_{motion} \end{aligned} \quad (5)$$

We fix the weight of global appearance stream to 1, and vary the weight value of the other two parts. Note that, we combine the loss functions of joint-specific local dynamic stream and motion-attentive local dynamic pose stream as a whole, which is corresponding to the distance measurement described below Eq. 4. The results are shown in the Table 10. When  $w_1$  equals to 1.2 and  $w_2$  equals to 1.4, the Rank 1 matching rate reaches the peak value 87.8%. When  $w_1$  equals to 1 and  $w_2$  equals to 1, the Rank 1 matching rate is 87.1%, which indicates that the default weight setting of our loss function can already achieve the performance that is comparable to the best case.

- *Analysis of the distance weighting parameter* We combine the distance of two types of local dynamic pose features by summation with equal weights in our development. We also evaluated using different weights for fusing the distances of two local dynamic pose features as follows:

$$Dis_{local} = a_1 * Dis_{MA} + a_2 * Dis_{JS}, \quad (6)$$

where  $a_1$  and  $a_2$  are the weights of the distance associated with motion-attentive local dynamic pose feature  $Dis_{MA}$  and the one associated with joint-specific local pose dynamic feature  $Dis_{JS}$ , respectively. The results in Table 11 show that the fusion with equal weights (i.e.,  $a_1 = a_2 = 1$ ) can already achieve the best performance. Therefore, we use equal weights for distance fusion in our implementation.

## 5.4 Evaluation on the MARS Dataset

To further verify whether our proposed model can still work on conventional video-based person re-id that does not mainly suffer from the fine-grained recognition problem, we evaluate our model on the MARS dataset (Zheng et al. 2016). MARS is a widely used video-based person re-id benchmark, including 1,261 identities and approximately 20,000 video sequences, the videos of each identity come from at least 2 camera views, and each identity has 13.2 sequences on average. For implementation, for a fair comparison, similar to DSEPA (Li et al. 2018a), we pre-trained our backbone (ResNet50) on DukeMCMT-reID dataset (Ristani et al. 2016), and then we fine-tuned our model on the MARS dataset (Note that we do not employ other person re-id datasets to pre-train our backbone and the whole model before training on FGPR). And we present the performance of our method compared to the state-of-the-art techniques on the MARS dataset in Table 12.

**Table 8** Performance (%) of different attentions on FGPR

Networks	All groups		
	Rank 1	Rank 5	mAP
CNN-guided attentions	81.3	91.6	82.7
Appearance-guided attentions (i.e., fine-to-coarse soft attentions without motion)	83.6	92.7	85.6
Motion-guided attentions (i.e., our model)	87.1	95.2	88.4

**Table 9** Exploration on parameter  $\alpha$  on FGPR

$\alpha$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Rank 1	86.6	86.9	87.1	86.5	86.3	86.9	86.7	86.3	86.1	86.2	86.3

**Table 10** Performance (%) of weighting the loss functions in our model on FGPR

Rank 1 $\backslash$ $w_2$	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	4.0	10.0
$w_1$ 0	67.8	69.3	68.4	72.4	73.9	75.6	76.4	77.2	75.7	74.1	72.6	73.5	73.3
0.2	72.1	74.6	75.5	78.9	79.3	79.8	79.6	80.6	78.1	76.7	75.2	73.4	73.1
0.4	75.6	77.4	79.3	80.8	80.9	81.1	81.9	81.4	79.3	77.0	76.3	74.1	75.2
0.6	78.5	79.1	80.8	81.6	82.2	81.9	82.4	83.1	80.3	78.4	77.1	76.5	76.7
0.8	82.2	83.2	84.6	84.2	85.4	84.7	85.2	86.6	83.2	83.0	81.9	80.2	79.6
1.0	85.9	85.6	86.6	86.3	86.9	<b>87.1</b>	87.3	87.	86.4	85.1	83.7	81.6	80.4
1.2	86.1	86.6	86.5	87.1	87.2	86.9	87.5	<b>87.8</b>	86.7	85.3	84.5	82.3	81.9
1.4	83.6	84.1	83.9	84.6	85.9	85.8	86.6	87.1	85.4	83.7	81.4	80.1	81.6
1.6	82.6	83.3	82.4	83.7	84.7	84.3	85.1	86.4	84.1	81.3	80.3	81.4	81.1
1.8	79.1	81.8	82.3	83.6	84.4	82.4	83.5	85.9	82.7	80.6	78.9	80.5	79.3
2.0	78.1	79.6	79.9	79.6	80.2	81.1	82.2	84.5	81.2	79.2	78.3	81.1	80.1
4.0	81.2	81.7	82.1	81.9	82.5	82.4	82.6	82.2	82.5	81.6	81.3	82.1	81.4
10.0	80.4	80.9	81.4	80.7	81.9	81.6	82.3	82.4	81.7	82.4	81.8	81.7	80.2

Bold values indicate our final results and the best results when the parameters vary

**Table 11** Performance (%) of weighting the local distances in our model on FGPR

Rank 1 $\backslash$ $a_2$	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	4.0	10.0
$a_1$ 0	78.4	78.6	79.5	80.3	80.1	80.4	80.6	80.5	80.2	80.4	80.9	80.8	80.7
0.2	79.2	79.8	80.4	80.7	81.3	81.9	81.7	81.7	81.7	81.5	81.4	81.2	81.1
0.4	79.9	80.1	80.8	81.5	82.4	82.7	82.4	82.4	82.3	82.4	82.1	81.9	81.9
0.6	81.4	81.9	82.1	82.4	83.6	84.9	84.1	84.1	83.5	82.9	82.9	82.8	82.5
0.8	81.7	82.7	83.5	84.7	85.9	86.4	85.6	85.4	85.3	84.4	83.4	83.5	83.2
1.0	82.1	83.9	84.3	85.7	86.4	<b>87.1</b>	86.5	85.2	85.6	85.1	84.4	84.6	83.7
1.2	82.3	83.7	84.3	85.5	86.7	87.2	87.1	86.8	86.4	85.4	84.9	84.5	84.4
1.4	81.9	83.5	84.1	84.5	86.1	86.5	<b>87.4</b>	86.9	85.8	85.4	84.3	84.4	84.4
1.6	82.4	83.5	83.8	84.3	85.5	85.9	86.9	86.4	85.6	85.3	84.8	84.5	84.1
1.8	82.2	83.4	83.8	83.5	84.9	85.5	86.3	86.1	85.4	85.1	84.8	84.4	83.8
2.0	81.6	83.2	83.4	83.4	84.7	85.5	86.1	86.1	84.6	84.7	84.6	84.5	84.5
4.0	81.8	82.8	83.4	83.1	83.9	84.2	84.2	85.2	84.6	84.6	84.5	84.3	84.5
10.0	82.4	82.6	83.1	82.9	83.6	82.8	83.7	83.9	84.5	84.4	84.1	84.3	84.6

Bold values indicate our final results and the best results when the parameters vary

– *Comparison with video-based deep models.* Table 12 shows the performance of our method on the MARS dataset compared to the performance of state-of-the-art techniques. Our approach attained a comparable performance: 82.9% in terms of the Rank 1 matching rate and 66.9% in terms of the mAP. Compared to the best performance reported by DSEPA (Li et al. 2018a), ours is 0.6%

greater in terms of the Rank 1 matching rate and 1.1% greater in terms of the mAP. Note that on the FG re-id problem, our proposed method performs clearly better than DSEPA, with a 7% higher Rank 1 matching rate and almost 4.2% higher mAP. Ours is 0.2 % matching rate lower on Rank 1 as compared with FARL (Zhang et al. 2019) and is 1.5 % matching rate lower on Rank 1 as



**Table 12** Performance (%) on MARS

Method	Rank 1(mAP)
CPS (Cheng et al. 2011)	49.6(26.3)
IDE (Zheng et al. 2017) + XQDA (Liao et al. 2015)	65.3(47.6)
Mars (Zheng et al. 2016)	68.3(49.3)
SeeForest (Zhou et al. 2017)	70.6(50.7)
QAN (Liu et al. 2017c)	73.7(51.7)
DSAN (Wu et al. 2018)	73.5(-)
P-SI <sup>2</sup> DL (Zhu et al. 2018)	75.3(-)
RQEN (Song et al. 2017)	77.8(71.1)
DuATM (Si et al. 2018)	78.7(62.3)
OFF (Sun et al. 2018)	79.4(63.1)
TGL (Dai et al. 2018)	80.5 (69.1)
EUG (Wu et al. 2018)	80.7(67.4)
DSEPA (Li et al. 2018a)	82.3(65.8)
FARL (TriNet) (Zhang et al. 2019)	83.1(69.9)
STMP (Liu et al. 2019)	<b>84.4(72.7)</b>
Our model	82.9(66.9)

Bold values indicate the best performance

compared with STMP (Liu et al. 2019); but for the FG re-id problem as shown in Table 2, our proposed method performs clearly better than STMP, with a 3.4% higher Rank 1 matching rate (especially 8% higher on the White group) and almost 2.1% higher on mAP. Therefore, the results suggest our model is still effective for the conventional person re-id problem.

- *Ablation study of our model on the MARS dataset* In addition, we also conducted ablation evaluation for motion-attentive feature and joint-specific feature in Table 13. Each stream is eliminated, and the results are presented in Table 13. As shown, compared with global appearance stream, our proposed local dynamic pose feature is more effective, which outperformed global appearance stream by 4.5% on Rank 1 matching rate and 4.4% on mAP as shown. These results show that global features are sensitive to occlusions and outliers in the sequences, while local features can to some extent alleviate these factors and are relatively more robust than global features.

When removing motion-attentive local dynamic pose feature, the performance is decreased by 2.2% on Rank 1 matching rate and 2.6% on mAP as compared with our full model in Table 13. When joint-specific local dynamic pose feature is removed from our model, the Rank 1 matching rate decreased by 1.3% and the mAP decreased by 1.2% as compared with our full model in Table. The results still show that our motion-attentive local dynamic pose feature and joint-specific local dynamic pose feature are also beneficial on MARS, although MARS is not formed to evaluate the fine-grained person re-id.

**Table 13** Performance (%) on MARS

Method	Rank 1(mAP)
Global appearance stream	77.4(60.3)
Local dynamic pose stream	81.9(64.7)
Motion stream	80.4(63.1)
Without motion-attentive feature	80.7(64.3)
Without joint-specific feature	81.6(65.7)
Full model	82.9(66.9)

## 6 Conclusions

Indeed, discriminating people who look extremely similar (e.g., who dress similarly) is very challenging due to ambiguities in appearance, and we therefore call this problem the FG re-id problem. To investigate this problem, we form the first benchmark dataset, called the FGPR dataset, for this problem. For solving the FG re-id problem, rather than relying on clothing color, we have proposed extracting motion-attentive local dynamic pose features and joint-specific local dynamic pose features, and they are learned simultaneously with global appearance and motion features using a deep neural network. Extensive results on the constructed FGPR dataset have validated the effectiveness of our model, especially the modeling of local attentive dynamic pose features that are more identity-specific and robust against similar appearances between people, for solving the FG re-id problem. In addition, our method is still effective on the conventional video-based person re-id problem.

**Acknowledgements** This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC (U1911401, U1811461), Guangdong Pro-vince Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong Project (No. 2018B030312002), Guangzhou Research Project (201902010037), Research Projects of Zhejiang Lab (No. 2019KD0A B03) and the Royal Society Newton Advanced Fellowship (NA150459). The principal investigator and corresponding author for this paper is Wei-Shi Zheng.

## References

- Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *CVPR*.
- Branson, S., Van Horn, G., Belongie, S., & Perona, P. (2014). Bird species categorization using pose normalized deep convolutional nets. arXiv preprint [arXiv:1406.2952](https://arxiv.org/abs/1406.2952).
- Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*.
- Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., & Murino, V. (2011). Custom pictorial structures for re-identification. In *BMVC*.
- Chung, D., Tahboub, K., & Delp, E. J. (2017). A two stream siamese convolutional neural network for person re-identification. In *CVPR*.
- Dai, J., Zhang, P., Wang, D., Lu, H., & Wang, H. (2018). Video person re-identification by temporal residual learning. *IEEE Transactions on Image Processing*, 28, 1366–1377.
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.
- Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.
- Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*.
- Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al. (2018). FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. In *NIPS*.
- Gou, M., Zhang, X., Rates-Borras, A., Asghari-Esfeden, S., Sznaiar, M., & Camps, O. (2016). Person re-identification in appearance impaired scenarios. In *BMVC*.
- Gray, D., Brennan, S., & Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*.
- Guo, Z., Zhang, L., & Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19, 1657–1663.
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *ICCV*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
- Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *SCIA*.
- Huang, S., Xu, Z., Tao, D., & Zhang, Y. (2016). Part-stacked CNN for fine-grained visual categorization. In *CVPR*.
- Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In *CVPR*.
- Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *CVPR*.
- Kviatkovsky, I., Adam, A., & Rivlin, E. (2013). Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 1622–1634.
- Li, S., Bak, S., Carr, P., & Wang, X. (2018a). Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*.
- Li, W., Zhao, R., & Wang, X. (2012). Human reidentification with transferred metric learning. In *ACCV*.
- Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*.
- Li, W., Zhu, X., & Gong, S. (2018b). Harmonious attention network for person re-identification. arXiv preprint [arXiv:1802.08122](https://arxiv.org/abs/1802.08122).
- Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., & Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *CVPR*.
- Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.
- Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In *ICCV*.
- Liu, J., Ni, B., Yan, Y., Zhou, P., & Cheng, S., Hu, J. (2018). Pose transferrable person re-identification. In *CVPR*.
- Liu, X., Wang, J., Wen, S., Ding, E., & Lin, Y. (2017a). Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *AAAI*.
- Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., et al. (2017b). Hydraplus-net: Attentive deep features for pedestrian analysis. arXiv preprint [arXiv:1709.09930](https://arxiv.org/abs/1709.09930).
- Liu, Y., Yan, J., & Ouyang, W. (2017c). Quality aware network for set to set recognition. In *CVPR*.
- Liu, Y., Yuan, Z., Zhou, W., & Li, H. (2019). Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*.
- Makihara, Y., Suzuki, A., Muramatsu, D., Li, X., & Yagi, Y. (2017). Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*.
- Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *CVPR*.
- Pumarola, A., Agudo, A., Sanfeliu, A., & Moreno-Noguer, F. (2018). Unsupervised person image synthesis in arbitrary poses. In *CVPR*.
- Qian, X., Fu, Y., Wang, W., Xiang, T., Wu, Y., Jiang, Y. G., & Xue, X. (2017). Pose-normalized image generation for person re-identification. arXiv preprint [arXiv:1712.02225](https://arxiv.org/abs/1712.02225).
- Rida, I., Jiang, X., & Marcialis, G. L. (2016). Human body part selection by group lasso of motion for model-free gait recognition. *IEEE Signal Processing Letters*, 23, 154–158.
- Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.
- Si, J., Zhang, H., Li, C. G., Kuen, J., Kong, X., Kot, A. C., & Wang, G. (2018). Dual attention matching network for context-aware feature sequence based person re-identification. arXiv preprint [arXiv:1803.09937](https://arxiv.org/abs/1803.09937).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Song, G., Leng, B., Liu, Y., Hetang, C., & Cai, S. (2017). Region-based quality estimation network for large-scale person re-identification. arXiv preprint [arXiv:1711.08766](https://arxiv.org/abs/1711.08766).
- Su, C., Li, J., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. In *ICCV* (pp. 3980–3989). IEEE.
- Sun, S., Kuang, Z., Sheng, L., Ouyang, W., & Zhang, W. (2018). Optical flow guided feature: A fast and robust motion representation for video action recognition. In *CVPR*.
- Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2017). Beyond part models: Person retrieval with refined part pooling. arXiv preprint [arXiv:1711.09349](https://arxiv.org/abs/1711.09349).
- Wang, T., Gong, S., Zhu, X., & Wang, S. (2014). Person re-identification by video ranking. In *ECCV*. Springer.

- Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*.
- Wei, L., Zhang, S., Yao, H., Gao, W., & Tian, Q. (2017). GLAD: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*.
- Wu, L., Wang, Y., Gao, J., & Li, X. (2018). Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, *21*, 1412–1424.
- Wu, L., Wang, Y., Shao, L., & Wang, M. (2019). 3-D person VLAD: Learning deep global representations for video-based person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, *30*, 3347–3359.
- Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., & Yang, Y. (2018). Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.
- Wu, Z., Huang, Y., Wang, L., Wang, X., & Tan, T. (2017). A comprehensive study on cross-view gait based human identification with deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 209–226.
- Xiao, T., Li, H., Ouyang, W., & Wang, X. (2016). Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*.
- Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., & Zhou, P. (2017). Jointly attentive spatial-temporal pooling networks for video-based person re-identification. arXiv preprint [arXiv:1708.02286](https://arxiv.org/abs/1708.02286).
- Ye, M., Li, J., Ma, A. J., Zheng, L., & Yuen, P. C. (2019). Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE Transactions on Image Processing*, *28*, 2976–2990.
- Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Deep metric learning for person re-identification. In *ICPR*.
- You, J., Wu, A., Li, X., & Zheng, W. S. (2016). Top-push video-based person re-identification. In *CVPR*.
- Zhang, W., He, X., Lu, W., Qiao, H., & Li, Y. (2019). Feature aggregation with reinforcement learning for video-based person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*. <https://doi.org/10.1109/TNNLS.2019.2899588>.
- Zhang, X., Xiong, H., Zhou, W., Lin, W., & Tian, Q. (2016). Picking deep filter responses for fine-grained image recognition. In *CVPR*.
- Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., & Tang, X. (2017). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*.
- Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-identification. In *CVPR*.
- Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*.
- Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., et al. (2016). Mars: A video benchmark for large-scale person re-identification. In *ECCV*. Springer.
- Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *CVPR*.
- Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q., et al. (2017). Person re-identification in the wild. In *CVPR*.
- Zheng, W. S., Gong, S., & Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 653–668.
- Zhou, Z., Huang, Y., Wang, W., Wang, L., & Tan, T. (2017). See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*.
- Zhu, X., Jing, X. Y., You, X., Zhang, X., & Zhang, T. (2018). Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. *IEEE Transactions on Image Processing*, *27*, 5683–5695.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.