IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 8, 2019, accepted July 19, 2019, date of publication August 13, 2019, date of current version September 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2935097

Spatial-Transformed Regional Quality Estimation Network for Large-Variance Person Re-Identification

MING JIANG[1], BIAO LENG[2,3], ZHIJUN MENG[4], AND GUANGLU SONG[1]

[1]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2]The Shenzhen Institute of Beihang University, Shenzhen 518057, China
[3]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China
[4]School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China

Corresponding author: Zhijun Meng (mengzhijun@buaa.edu.cn)

This work was supported in part by the Science, Technology and Innovation Commission of Shenzhen Municipality Foundation under Grant JCYJ20170307100750340, in part by the Beijing Municipal Natural Science Foundation under Grant 4182034, in part by the National Natural Science Foundation of China under Grant 61702023, and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT The video-based person re-identification aims to search a query video in a large video gallery. The major restrictions on the performance of this task are image misalignment and partial noises caused by detection error, occlusion, blur, and illumination. The misalignment between the different frames in a video caused by excessive background or part missing may play the devil with pedestrian matching. In addition, the influence of partial noises on pedestrian matching performance is also unfavorable. Since different spatial regions of a single frame have various qualities, and the quality of the same region also varies across frames in a tracklet. A good way to address the problem is to effectively aggregate complementary information from all frames in a sequence, using better regions from other frames to compensate for the influence of an image region with poor quality. To achieve this, we propose a novel Spatial-transformed Regional Quality Estimation Network (SRQEN), where a well-designed spatial-transformed unit is used to automatically learn the alignment from an identification procedure and another ingenious training mechanism enables the effective learning to extract the complementary region-based information between different frames. Visual examples indicate that pedestrians are better aligned with SRQEN and the proposed method can learn complementary information. Extensive experiments show that compared with other feature extraction methods, we achieved comparable results of 93.5%, 79.8%, and 74.85% on the PRID 2011, iLIDS-VID, and MARS, respectively.

INDEX TERMS Deep learning, person re-identification, regional quality estimation, spatial-transformed mechanism.

I. INTRODUCTION

The purpose of person re-identification (person re-id) is to identify a person by searching for similar images or videos between the probe and a gallery set [1], [2]. Person re-id has drawn a lot of attention in recent years due to its wide applications, such as intelligent surveillance system, video analysis, and multi-target tracking [3].

Video-based person re-id remains unsolved due to large intra-class and inter-class variations caused by lighting change, cameras' view change, occlusion, and misalignment. Feature extraction and metric learning are the two key components that most recent works are focused on.

Different kinds of features [4], [5] have been tailored and employed in previous feature extraction methods. Hand-crafted features mainly focus on two aspects: space-time features and appearance features. HOG3D [6] is a popular space-time feature. Color (e.g. HSV and LAB color histogram) and texture features (e.g. LBP [7], Schmid filters, and Gabor filters) are popular appearance features. Gray and Tao [8] employed an AdaBoost algorithm to select discriminative features from RGB, YCbCr, and HSV color space and nineteen texture features. Reference [9] proposes a Temporal Scale Pyramid to extract the difference among

The associate editor coordinating the review of this article and approving it for publication was Syed Muhammad Anwar.

 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

frames. Deep learning based models have been widely used for feature extraction. These models show superior performance in recent few years. Zhao *et al.* [10] proposes the local patch matching method that learns the mid-level filters to generate the local discriminative features for the person re-id. Reference [11] proposes Feature Fusion Net (FFN) for pedestrian image representation that hand-crafted histogram features can be complementary to Convolutional Neural Network (CNN) features. Reference [12] uses a recurrent neural network to learn the interaction between multiple frames in a video. A Siamese network is employed to learn the discriminative video-level features for person re-id. Reference [13] uses the Long-Short Term Memory (LSTM) network to aggregate frame-wise person features in a recurrent manner. Reference [14] uses both the global full-body and local body-parts features of the input persons to improve the person re-id performance. Reference [15] learns a robust feature representation with data from multiple domains for the same task. [16] utilizes the pedestrians' spatial-temporal patterns to improve performance. Reference [17] extracts sub-region features at different stages and proposes a tree-structured network for feature fusion.

Other works are devoted to learning discriminative distance metric functions for person re-id. Several methods have been proposed such as KISSME [18], LMNN [19], and Mahalnobis [20]. Reference [21] proposes the top-push distance learning model to enhance inter-class differences, and reduce intra-class differences. Reference [22] uses a more general way to learn a similarity metric from image pixels directly. FaceNet [23] and Ding *et al.* [24] apply this triplet framework to the face and person re-id task, respectively. References [25], [26] use the Siamese network architecture for feature mapping from raw images to a feature space where images from the same person are close while images from different persons are widely separated. Reference [27] proposes a systematic evaluation of person re-id task by incorporating recent advances in both feature extraction and metric learning.

Video-based person re-id has made a great improvement with the assistance of the powerful convolutional neural network (CNN) [28]–[31]. However, these methods may still be strongly affected by partial noises or significant misalignment. Compared to a single image, frames from a sequence provide richer information complementary to each other [32]. An intuitive way to aggregate information from a sequence is to simply take the average feature [33], but this can draw undesired noises into the feature such as partial occlusions existing in some specific images. In addition, the simple average method is not robust to the misalignment caused by significant body movements and the deviation of the person detectors. As shown in Fig. 1, some images in this sequence suffer both partial occlusion and misalignment. And using the simple average method will cause recognition failure because the effective information is weakened by noises and misalignment.

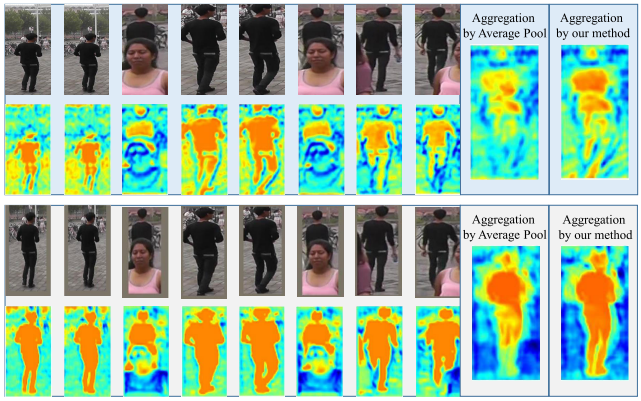


FIGURE 1. Illustration of different aggregation methods and the effectiveness of the spatial-transformed unit. The first row is an image sequence of pedestrian, the third row is the same sequence aligned by spatial-transformed unit, and the second row and the fourth row are the corresponding feature maps extracted by the baseline and our algorithm, respectively. Best viewed in color.

Reference [34] has tried to learn the image alignment from an identification procedure, but it mainly focuses on the alignment of global image and the partial noise information carried by the original image still exists in the image after alignment. In order to weaken the influence of partial noise information, Reference [35] has proposed a novel model which generates the adaptive score for each image in a sequence where the images with noise information will get lower scores. However, this score doesn't specify which part causes the major limitation. This method possibly leads to the dumping of a frame due to its unconsidered noisy part, causing the loss of valuable information from all other regions of that frame.

Considering all of the above problems, in this paper, we focus on the challenges that how to automatically align the person body in different frames and extract the complementary information between different frames via adaptive regional quality. A Spatial-transformed Regional Quality Estimation Network (SRQEN) is proposed to alleviate image misalignment and partial noises. SRQEN first employs the Spatial-Transformed Unit (STU) to align the input frames, then a well-designed Regional Quality Predictor (RQP) is used to compute the quality scores of different regions in a frame. The final video-level feature is generated by Set Aggregation Unit (SAU), using the quality scores and frame-level features. SRQEN could concentrate more attention on the effective images' regions in a sequence and aggregate the complementary regional information between different frames.

Another problem is that the current person re-identification datasets [25], [36]–[38] are flawed either in scale or cleanliness. Prevailing sizes range from 200 to 1500 identities, scanty for more extensive experiments. Relatively large datasets exist, but they possess poor cleanliness due to detection or tracking failure, which significantly undermines the performance of algorithms. Moreover, persons in most

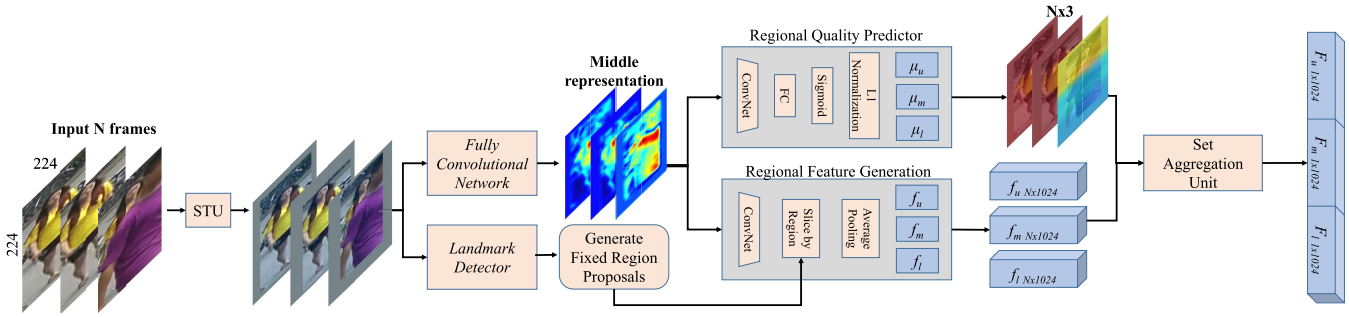


FIGURE 2. The inference of the proposed SRQEN. The input of this network is an image set belonging to the same person. Each of them first generates the aligned image through the spatial-transformed unit. Then the middle representation of the aligned image is generated by the fully convolutional network (FCN). The representation will be fed to the regional feature generation unit with landmarks and regional quality predictor. The scores of different regions in [0, 1] are indicated by the color map, i.e. from blue to red. Then the scores and features of all images will be aggregated by the set aggregation unit and the final representation of the image set will be produced. STU denotes spatial-transformed unit. $\{\mu_u, \mu_m, \mu_l\}$ and $\{f_u, f_m, f_l\}$ represent the regional quality scores and features of different regions in an image, respectively. $\{F_u, F_m, F_l\}$ means the video-level feature representation of a person.

datasets are well-aligned by hand-drawn bounding boxes. But in reality, the bounding boxes detected by pedestrian detectors may undergo misalignment or part missing [38]. To address these problems, it's important to introduce a large-scale dataset that is both clean and close to realistic settings.

For alleviating the scantiness of large-scale and clean datasets, we propose a new dataset named the "Labeled Pedestrian in the Wild (LPW)". It contains 2,731 pedestrians in three different scenes where each annotated identity is captured by from 2 to 4 cameras. The LPW features a notable scale of 7,694 tracklets with over 590,000 images as well as the cleanliness of its tracklets. It distinguishes from existing datasets in three aspects, containing large scale with cleanliness, automatically detected bounding boxes, and far more crowded scenes with greater age span. This dataset provides a more realistic and challenging benchmark, which facilitates the further exploration of more powerful algorithms.

To sum up, our contributions in this work are as follows:

1) Be the first to take into account the qualities of images' different regions to better aggregate complementary region-based information in a sequence, using the information of a certain image region with higher quality to make up for the same region of the other frames with poor quality.
2) We propose a well-designed pipeline to jointly learn the frame alignment and evaluate the qualities of different image regions. The proposed method achieves state-of-the-art performance on iLIDS-VID and PRID 2011 for video-based person re-identification.
3) Construct a new person re-identification dataset named "Labeled Pedestrian in the Wild (LPW)" with large scale and good cleanliness. It contains 7,694 tracklets and over 590,000 images, and can be available on `http://liuyu.us/dataset/lpw/index.html`.

II. PROPOSED METHOD

A. ARCHITECTURE OVERVIEW

The inference of the SRQEN is shown in Fig. 2. The input of SRQEN is a set $S = \{I_1, I_2, \cdots, I_n\}$ where each image belongs to the same person. The spatial-transformed unit (STU) is first used to align the input images. Then the fully convolutional network will be used to generate the middle representations of aligned images. The middle representations are sent into two different parts. In the regional feature generation part, the landmark detector [39] marks the key points of the human body. The middle representations are divided into different regions according to the detected key points. Due to the poor resolution of people images from real-life surveillance systems, the detected landmarks for a single image usually aren't accurate enough to directly give a satisfying division. Our solution is to calculate the standard person landmarks according to the whole training dataset and figure out a generally regional division according to the standard person. This requires that the dataset should be well aligned and thus the fixed division can be universal to an arbitrary person image. Let u, m, l represent the upper part, middle part and the lower part of images, respectively. Each part corresponds to a region of the original image and they generate feature vectors $F_{I_i} = \{f_u(I_i), f_m(I_i), f_l(I_i)\}$ using average pooling method. The other unit called regional quality predictor can generate the quality estimation $\mu_{I_i} = \{\mu_u(I_i), \mu_m(I_i), \mu_l(I_i)\}$ for the corresponding regional features. On the basis of the former work [35], we directly feed the output feature maps of the *pool2* layer into the regional quality generation unit. Each score of different regions $\mu_{part}(I_i)$ evaluates the corresponding regional feature $f_{part}(I_i)$. Eventually, the scores and their corresponding features generate the video-level feature representation $F_w(S)$ through the set aggregation unit. We now explicitly discuss these units in the following sections.

B. SPATIAL-TRANSFORMED UNIT

When generated by automatic detectors, person images usually contain misalignment problem including excessive background or part missing, which will compromise pedestrian matching accuracy. Several works [40], [34] utilize spatial transformation mechanism to align the feature maps of an image. However, the low-level part of the CNN still needs to distract from extracting robust feature from the misaligned

image. In order to alleviate the burden of processing the misaligned image, we design an improved spatial-transformed unit to apply on original image replaced with the feature maps. Further more, inspired by [41] where CNN can handle a certain rotation variance controlled at a certain range, based on the fact that almost all pedestrians do not have large angle tilt, we remove the learning of image rotation to reduce learning constraints of STU. The parameters in the STU can be learned without the supervision of transformation label.

The spatial-transformed unit takes the original image I as input, and outputs the transformed image I_T. The computing details are depicted as follows. Firstly, a lightweight network named localisation network utilizes the input image to compute the transformation parameter θ: $\theta = \mathcal{F}_{loc}(I; W_{loc})$, where $\mathcal{F}_{loc}$ is the localisation network function and W_{loc} is corresponding parameters. In this paper, the localisation network consists of four convolutional layers and two fully connected layers. Then the grid generator uses θ to create a regular sampling grid $G = \{G_i\}$, where $G_i = (x_i^t, y_i^t)$ forming the output feature map. As for affine transformation, the computational process can be formulated as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}. \quad (1)$$

where (x_i^s, y_i^s) are the source coordinates in the input feature maps, and (x_i^t, y_i^t) are the target coordinates of the regular grid in the output feature maps. Finally, the transformed image I_T is generated by the sampler using the sampling grid $G = \{G_i\}$ and the input image I. The aligned images are more suitable for person re-id task as illustrated in Fig. 1. Note that in our implementation, we fix θ_{12} and θ_{21} to be 0 in (1), because CNN can handle a certain rotation variance controlled at a certain range and almost all pedestrians do not have large angle tilt.

C. STRATEGY OF GENERATING REGION PROPOSALS

Let $P_i = \{p_1, p_2, \cdots, p_m\}$ represents the landmark set of image I_i. m is the total number of landmarks which is set to 14 in our method. Fig 3 shows some samples from iLIDS-VID dataset. Because of the low resolution and occlusion of some images, in many cases, accurate landmarks are not available. A number of such detection failures exist in the datasets under surveillance. Taking into account these factors, in order to ensure more accurate generation of region proposals, we compute the standard person landmarks by averaging the key points over the acceptable threshold in the whole aligned dataset. According to this standard landmarks, the full area of an image is divided into a generally fixed three-part-division $I_i = \{I_u, I_m, I_l\}$ in the height direction computed by location ratio (3:2:2 in our experiment) of different regions. As CNN has considerable translation invariance and the use of STU, though regions are fixed, most body movements can be tolerated which can effectively filter landmark detection errors.

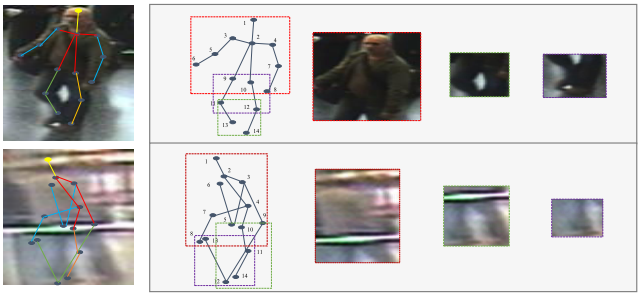


FIGURE 3. Samples of 14 landmarks detected on iLIDS-VID. Landmarks of the first line are detected successfully and image of the second line with low resolution and blur failed to detect.

D. REGIONAL QUALITY PREDICTOR

The quality of the image has a great influence on the feature extraction of the neural network. Images of a sequence may contain a lot of noises such as blur, occlusion, and deformation. Discarding the information of the whole image containing the regional noise often leads to the loss of critical regional information which plays a key role in feature matching. Thus we propose the regional quality predictor as shown in Fig. 2. The middle representation of the input image is sent into a convolutional network consisting of convolution layers and fully connected layer. In the generation of video-level features, the regional quality scores can help depose some noisy information and thus the complementary information between different frames can be efficiently aggregated. The fully connected layer generates the original scores $\mu_{ori}(I_i)$ corresponding to images' different regions The scores are scaled to $[0, 1]$ by the sigmoid function $\sigma(\cdot)$. The score of each region is generated by its corresponding feature map. In the last stage, scores belonging to the same region of the sequence are normalized, in order to facilitate video-level complementary information aggregation.

E. SET AGGREGATION UNIT

An image set $S = \{I_1, I_2, \cdots, I_n\}$ is mapped to a representation with a fixed dimension by the set aggregation unit. For an image, the regional feature generation part generates the different regions' representation $F_{I_i} = \{f_u(I_i), f_m(I_i), f_l(I_i)\}$ and the regional quality predictor part generates the corresponding scores $\mu_{I_i} = \{\mu_u(I_i), \mu_m(I_i), \mu_l(I_i)\}$. The feature representation of an image sequence can be denoted as: $F_w(S) = \{\mathcal{F}_u(S), \mathcal{F}_m(S), \mathcal{F}_l(S)\}$, where $\mathcal{F}(\cdot)$ is the set aggregation function. $\mathcal{F}(\cdot)$ generates feature representation with fixed dimension for a sequence with various size by incorporating all frames in a weighted manner where highly scored areas contribute more information. $\mathcal{F}(\cdot)$ is formulated as:

$$\mathcal{F}_{part}(S) = \sum_{i=1}^{n} \mu_{part}(I_i) f_{part}(I_i) \quad (2)$$

where $\mu_{part}(I_i)$ and $f_{part}(I_i)$ represent different regions' scores and features, respectively and n is the number of images. All

the scores are normalized, so there are $\sum_{i=1}^{n} \mu_u(I_i) = 1$, $\sum_{i=1}^{n} \mu_m(I_i) = 1$ and $\sum_{i=1}^{n} \mu_l(I_i) = 1$.

F. LEARNING COMPLEMENTARY INFORMATION VIA JOINT TRAINING FRAME-LEVEL AND VIDEO-LEVEL FEATURES

The representation of information is not independent between different frames. The complementary information of different image regions in pedestrian image sets can effectively improve the performance of recognition. The extraction of complementary regional information should weaken the influence of noise information and fuse more effective information of pedestrian expression. This can be accomplished via the SRQEN we proposed with the assist of regional quality scores.

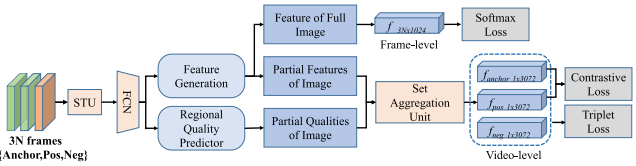


FIGURE 4. The training stage of SRQEN. The frame-level features are supervised by Softmax loss with pedestrian identities, while video-level features are supervised by Contrastive Loss and Triplet loss.

Fig. 4 shows the training stage of SRQEN. The overall loss of SRQEN can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_c + \mathcal{L}_t, \tag{3}$$

Note that $\mathcal{L}_s$ is the Softmax Loss, $\mathcal{L}_c$ is the Contrastive Loss, and $\mathcal{L}_t$ represents the Triplet Loss:

$$\mathcal{L}_t = [d(F_w(S_i^o), F_w(S_i^+)) - d(F_w(S_i^o), F_w(S_i^-)) + \tau]_+. \tag{4}$$

where the $d(\cdot)$ is L_2-norm distances, $[\cdot]_+$ indicates $max(\cdot, 0)$ and τ is the margin of triplet loss. The input of the network is a triplet $S_i =< S_i^o, S_i^+, S_i^- >$, and the network maps S_i into a feature space with $F_w(S_i) =< F_w(S_i^o), F_w(S_i^+), F_w(S_i^-) >$. We define S_i^o as anchor set, S_i^+ as positive set, and S_i^- as negative set.

By using the supervision of pedestrian identities, the frame-level features can be closer to the pedestrian center in feature space. The more robust video-level features can be extracted via joint training frame-level and video-level features.

In the stage of regional quality score predictor, T represents input middle representation, $\mu_{ori}(I_i)$ stands for partial scores before normalization. The generation formulation is

$$\mu_{ori}(I_i) = \sigma(W \cdot T + b), \tag{5}$$

where W and b are the corresponding parameters and $\sigma(\cdot)$ is the sigmoid function scaling the scores to [0, 1]. The stage of normalization is

$$\mu_{part}(I_i) = \frac{\mu_{ori}(I_i)}{\sum_{i=1}^{n} \mu_{ori}(I_i)}. \tag{6}$$

TABLE 1. The identities captured by each view in LPW.

Scene1		Scene2		Scene3	
	Identities		Identities		Identities
View1	662	View1	874	View1	166
View2	756	View1	1501	View1	203
View3	410	View1	1751	View1	224
View4	–	View1	996	View1	151

The back propagation can be formulated as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu_{ori}(I_i)} &= \frac{\partial \mathcal{L}_t}{\partial \mu_{ori}(I_i)} + \frac{\partial \mathcal{L}_c}{\partial \mu_{ori}(I_i)} \\ &= (\frac{\partial \mathcal{L}_t}{\partial \mathcal{F}_{part}(S)} + \frac{\partial \mathcal{L}_c}{\partial \mathcal{F}_{part}(S)}) \\ &\quad \cdot (\frac{f_{part}(I_i)}{\sum_{i=1}^{n} \mu_{ori}(I_i)} - \frac{\mathcal{F}_{part}(S)}{\sum_{i=1}^{n} \mu_{ori}(I_i)}),\end{aligned} \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial f_{part}(I_i)} = (\frac{\partial \mathcal{L}_t}{\partial \mathcal{F}_{part}(S)} + \frac{\partial \mathcal{L}_c}{\partial \mathcal{F}_{part}(S)}) \cdot \mu_{part}(I_i). \tag{8}$$

After back propagation through set aggregation unit, the gradient of $\mu_{ori}(I_i)$ can be calculated by (7). Notice that there is the dot product of $\nabla \mathcal{F}_{part}(S)$ and vector $f_{part}(I_i) - \mathcal{F}_{part}(S)$. If the angle between them is $(-90°, 90°)$, the gradient of $\mu_{ori}(I_i)$ will be positive and the $f_{part}(I_i)$'s quality score $\mu_{ori}(I_i)$ will increase after this back propagation process. Obviously, features of noisy parts, which usually deviate from those of normal parts, tend to receive low quality scores. The gradient of $f_{part}(I_i)$ is shown in (8). We can notice that the gradients of low-scored-parts with occlusion, blur and noise will be weakened so that their influence on parameters in the feature generation part will become negligible during back propagation. Such scheme guarantees the feature generation part to concern more about the good parts of images, which benefits video-based matching.

III. THE LABELED PEDESTRIAN IN THE WILD DATASET

A. DATASET DESCRIPTION

In this section, a new large scale person re-id dataset, the LPW dataset, is introduced. The dataset is collected in three different crowded scenes. There are three cameras placed in the first scene and four cameras in the other two scenes. During collection, the cameras with the same set of parameters were placed at the intersections of the streets. The LPW consists of 2,731 different pedestrians in total. We make sure that each annotated identity is captured by at least two cameras so that cross-camera search can be performed. Some sample images from different perspectives of view are shown in Fig. 5.

The conversion from raw video to tracklets is threefold. First, pedestrians are captured by an elaborate detector. Then the overlaps between bounding boxes in two consecutive frames are computed to correctly assign boxed person images into their tracklets. Finally, we manually clean up the dataset, eliminating contaminated images caused by wrong detection and impure tracklets with tracking errors. As a result, a total of 7,694 image sequences are generated with an average of 77 frames per sequence. There are 1058, 1114 and

TABLE 2. Comparing LPW with existing datasets containing MARS [32], iLIDS-VID [36], PRID2011 [37], MSMT17 [42], Market [38], Airport [?], CUHK03 [25], and PKU-Reid [43]. The DT failure stands for whether detection or tracking failure in sequences exist. The symbol # represents the number of corresponding aspects.

Datasets	LPW	MARS	iLIDS-VID	PRID2011	MSMT17	Market	Airport	CUHK03	PKU-Reid
#identities	2,731	1,261	300	200	4,101	1,501	1,382	1,360	114
#BBoxes	590,547	1,067,516	43,800	40,000	126,441	32,668	39,902	13,164	1,824
DT failure	no	yes	no	no	no	yes	no	no	yes
#tracklets	7,694	20,715	600	400	–	–	–	–	–



FIGURE 5. Sample images from the different cameras in the LPW. There are various postures and large span of age.

559 pedestrians captured by two, three and four cameras, respectively. The identities in each view are shown in Table 1.

Overall, our dataset has three characteristics. First, from the comparison in Table 2 we can see that most existing datasets are either too small in size or large in size but not clean, which negatively affects the performance of deep learning based methods and the implementation of many new ideas. Our dataset not only has a large scale in the number of pedestrians but also bears cleanliness over sequences, which meets the training requirements of many deep networks. Second, other large datasets exist in many cases where the images are not well aligned and area of pedestrian account for only a small proportion of the image. In LPW the bounding boxes are fine-tuned to make sure that area of a pedestrian is dominant. Third, this dataset is collected in three different crowded scenes with a large span of age, frequent occlusion, and various postures, all reflecting real-world conditions, which is more challenging. The large-scale dataset with good cleanness can be fully exploited to obtain more discriminative information about the person of interest.

B. EVALUATION PROTOCOL

The LPW is divided into training and testing sets, containing 1,975 and 756 identities, respectively. The persons in the second scene and the third scene are used for training and the first scene is used for testing. In the testing set, the videos from *view*2 in the first scene are used as probe set and the other two views are used as the gallery set. The reason for using *scene*1 as a testing set is that the first scene contains all the challenging cases such as occlusion and significant changes of pedestrian pose in the other scenes. There are 756 probes and 1,072 galleries, thus for the same probe, there are multiple galleries. The images of *scene*2 and *scene*3 are used for training. Given the abundance of samples in this dataset, it is adequate to conduct one single experiment with fixed train/test partitioning, while repeated trials with random partitioning are considered unnecessary.

The Cumulative Matching Characteristic (CMC) curve is employed as the evaluation metric. Given a query image sequence, all gallery identities are assigned a similarity score and the galleries are ranked according to their similarity to the query.

IV. EXPERIMENTS

In this section, we evaluate the performance of SRQEN on several video datasets. We first conduct an ablation study to analyze the effectiveness of each component in SRQEN, then we compare our results with the state-of-the-art methods and verify the generalization ability of our model on cross-dataset testing. Finally, the compatibility of generative scores with different image parts and the complementarity of information are discussed in detail.

A. DATASETS AND EXPERIMENTAL SETTINGS

In this paper, we conduct experiments on three publicly available video datasets and the proposed large-scale dataset: the PRID 2011 dataset [37], iLIDS-VID dataset [36], MARS [32] and the proposed LPW dataset. Sample images from these datasets are shown in Fig. 5 and Fig. 6. We briefly introduce the three public datasets and the details of the LPW dataset is depicted in Section III.

1) PRID 2011

The PRID 2011 dataset contains image sequences captured by two non-overlapping cameras. One camera view has 385 persons and the other has 749 persons. Only 200 persons appeal in both two cameras. The length of the image sequences varies from 5 to 675. In order to train our models, following [36], the sequences of 178 persons are used which have more than 21 frames in the sequence.



(a) Samples from PRID 2011



(b) Samples from iLIDS-VID



(c) Samples from MARS

FIGURE 6. Sample image sequences from three benchmark dataset. Some of them suffer from misalignment and noises.

2) ILIDS-VID

The iLIDS-VID dataset contains 300 persons, and each person has two different sequences captured by two disjoint cameras. This dataset is challenging because the training data is limited and there are heavy occlusions and similar clothes in some cases.

3) MARS

The MARS dataset is collected using six non-overlapping cameras in a university. It is one of the largest video datasets, which contains 1,261 persons who are captured by at least 2 cameras. There are 20,715 image sequences in total and each identity has 13.2 tracklets on average. 3,248 image sequences are distractor tracklets due to false detection or tracking results, which makes this dataset more challenging.

4) EVALUATION SETTINGS

In our experiments, we adopted a multi-shot experiment setting. The evaluation of iLIDS-VID and PRID 2011 is same as CNN+RNN [12] and QAN [35]. We use 50% of persons for training and the rest for testing. To obtain statistically reliable results, we repeat the experiment ten times on iLIDS-VID and PRID 2011, and report the average of the results. The evaluation setting of MARS is the same as [32]. We use the protocol mentioned in Section III-B for the evaluation of LPW. Features are measured using cosine distance. The results are reported in terms of Cumulative Matching Characteristics (CMC) curves.

TABLE 3. Ablation study on iLIDS-VID and PRID2011.

Conf	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
+RU		√					√	√
+RM			√				√	√
+RL				√			√	√
+QFix					√			
+MP						√		
+STU								√
	iLIDS-VID							
CMC1	64.8	58.2	69.1	54.9	72.6	43.6	77.1	**79.8**
CMC5	85.5	85.7	88.9	77.2	91.6	68.2	93.2	**94.8**
CMC10	89.9	92.2	94.0	85.7	96.4	76.5	97.7	**97.7**
CMC20	95.3	98.0	97.3	92.9	99.1	83.0	99.4	**99.5**
	PRID2011							
CMC1	81.6	76.9	86.1	75.7	90.0	39.7	92.4	**93.5**
CMC5	96.5	95.6	97.2	91.7	98.8	64.9	98.8	**99.3**
CMC10	98.3	99.1	99.1	96.1	99.8	76.1	99.6	**99.8**
CMC20	99.1	99.8	99.6	97.9	100.0	85.7	100.0	**100.0**

5) MODEL SETTINGS

We use GoogLeNet [30] as the backbone network. The input images are resized to $224 \times 224 \times 3$. The Stochastic Gradient Descent (SGD) is employed to optimize the network, with the batch size set to 24 and base learning rate set to 0.001. The max iteration $= 15,000$, $15,000$, $80,000$, and $10,0000$ on PRID 2011, iLIDS-VID, MARS, and LPW, respectively.

B. ABLATION STUDY

In this subsection, we compare the influence of different units in SRQEN. Table 3 shows the comparing results on iLIDS-VID dataset and PRID 2011 dataset. The results of ablation study experiments on large-scale LPW are shown in Fig. 7. We remark that in Table 3 and Fig. 7 all models are trained with the same parameter setting and tested using the same configuration.

1) THE PERFORMANCE OF THE BASELINE

Method (a) is the GoogLeNet [30] with batch normalization [44] initialized with the ImageNet model. Benefiting from stronger generalization ability of GoogLeNet with batch normalization, the performance of our baseline has achieved a high accuracy. The top-1 accuracy is 64.7% and 81.6% on iLIDS-VID and PRID 2011, respectively. The performance is higher than many existing works [12], [32], [45].

2) THE PERFORMANCE OF DIFFERENT REGIONAL FEATURES

In method (b), (c) and (d), $+RU$, $+RM$ and $+RL$ represent the features of different regions as mentioned above. In the result, we can notice that the middle part has a wealth of information.

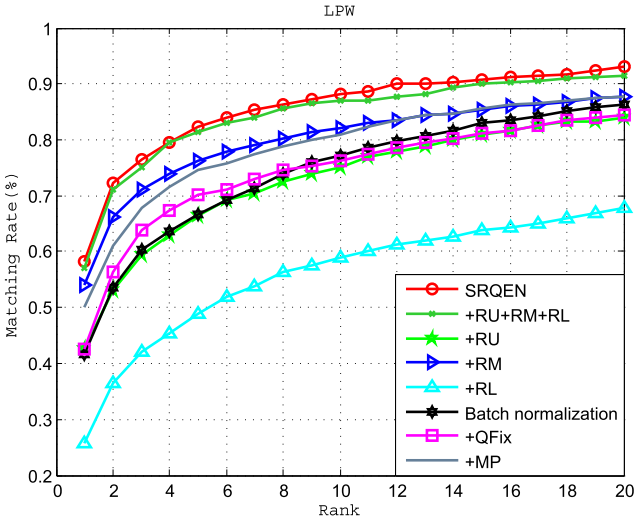


FIGURE 7. Performance of SRQEN, baseline and ablation study experiments on LPW dataset.

The results fit our perception that the upper part and middle part of the body plays a better role in identifying a person.

3) THE EFFECTIVENESS OF THE REGIONAL QUALITY PREDICTOR

In order to verify the effectiveness of regional quality predictor part, we performed two other comparisons method (e) and (f). In the first experiment *+QFix*, we fixed the quality generation unit by setting all the quality scores to 1. In method (e), joint training of image-level and video-level features improves performance to a certain extent. With the assistance of the regional quality predictor, the performance has been further improved. Considering that SRQEN has increased the number of parameters and in order to explore the impact of parameter increase on performance, we configure *+MP* which has the same number of parameters with SRQEN. The branches generated scores are used for generating features and we directly concatenate the scalar with the feature produced by these branches to be the final feature. Over-fitting of the network caused by insufficient data results in the decrease of performance of method (f) on iLIDS-VID. In method (g), we use all the region proposals and the corresponding regional quality scores. On both two datasets, the performance has been improved by a large margin compared to the baseline, which demonstrates the effectiveness of the proposed method.

4) THE EFFECTIVENESS OF THE SPATIAL-TRANSFORMED UNIT

Method (h) is the configuration of SRQEN, the STU is added on the basis of method (g). Compared to method (g), the top-1 accuracy has been improved by 2.7% and 1.1% on iLIDS-VID and PRID 2011, respectively. Although both iLIDS-VID and PRID 2011 are manually cropped and the misalignment problem is not much serious than MARS and LPW, the performance has also been improved when STU is added.

TABLE 4. Comparison results with state-of-art methods on PRID 2011 dataset. rank-1,5,10,20 accuracy (%) are shown. The best results are shown in boldface.

Matching rate(%)	PRID2011			
Rank	1	5	10	20
SRQEN	**93.5**	**99.3**	**99.8**	**100.0**
QAN [35]	90.3	98.2	99.3	100.0
ST^2N [49]	87.8	97.4	–	99.3
CNN+SRM+TAM [47]	79.4	94.4	–	99.3
CNN+XQDA [32]	77.3	93.5	–	99.3
DSAN+KISSME [50]	77.0	96.4	99.2	99.4
ASTPN [48]	77.0	95.0	99.0	99.0
CNN+RNN [12]	70.0	90.0	95.0	97.0
STA [46]	64.1	87.3	89.9	92.0
TDL [21]	56.7	80.0	87.6	93.6

TABLE 5. Comparison results with state-of-art methods on iLIDS-VID dataset. rank-1,5,10,20 accuracy (%) are shown. The best results are shown in boldface.

Matching rate(%)	iLIDS-VID			
Rank	1	5	10	20
SRQEN	**79.7**	**94.8**	**97.7**	**99.5**
QAN [35]	68.0	86.8	95.4	97.4
ASTPN [48]	62.0	86.0	94.0	98.0
DSAN+KISSME [50]	61.9	86.8	94.7	98.6
CNN+RNN [12]	58.0	84.0	91.0	96.0
ST^2N [49]	57.7	81.7	–	94.1
TDL [21]	56.3	87.6	95.6	98.3
CNN+SRM+TAM [47]	55.2	86.5	–	97.0
CNN+XQDA [32]	53.0	81.4	–	95.1
STA [46]	44.3	71.7	83.7	91.7

TABLE 6. Comparison results with state-of-art methods on MARS dataset. rank-1,5,20 accuracy (%) and mAP (%) are shown. The best results are shown in boldface.

Matching rate(%)	MARS			
Rank	1	5	20	mAP
SRQEN	**74.85**	88.03	92.53	57.78
IDE+Re-ranking [51]	73.94	–	–	**68.45**
DSAN+KISSME [50]	73.5	85.0	97.5	–
CNN+SRM+TAM [47]	70.6	**90.0**	**97.6**	50.7
CNN+XQDA [32]	65.0	81.1	88.9	49.3
ASTPN [48]	44.0	70.0	81.0	–

C. COMPARISON WITH PUBLISHED RESULTS

In the stage of the experiment, results of the evaluation are shown in Table 4, Table 5, Table 6 and Fig. 7. The result is compared with other state-of-the-art methods including CNN+XQDA [32], CNN+RNN [12], STA [45], TDL [21], CNN+SRM+TAM [46], ASTPN [47], ST^2N [48], QAN [35], DSAN+KISSME [49], and IDE+Re-ranking [50]. *Matching rate*(%) at Rank *i* means the accuracy of the matching within the top *i* gallery classes.

On PRID 2011 dataset, SRQEN increases top-1 accuracy by 3.2% compared with the state-of-the-art. On iLIDS-VID dataset, SRQEN increases top-1 accuracy by 11.7% compared with the state-of-the-art. Notice that the performance on iLIDS-VID has been improved by a large margin because the

TABLE 7. Cross-dataset testing performances on PRID 2011 dataset. CD [51] Is trained On Shinpuhkan 2014 dataset and other methods are trained On iLIDS-VID. The PRID 2011 Is used for testing.

Matching rate(%)	PRID 2011			
Rank	1	5	10	20
SRQEN	**62.8**	**86.5**	**95.3**	**98.5**
QAN [35]	34.0	61.3	74.0	83.1
Batch normalization [45]	34.3	61.2	76.4	88.8
CNN+RNN [12]	28	57	69	81
CD [52]	17	–	43	52

iLIDS-VID dataset has more images with partial occlusion and deformation. And our model can be effective for selecting the complementary information of the images' regions with high confidence, abandoning the inherent noise of some regions in other frames.

On the large-scale dataset MARS, SRQEN achieves the comparable result with the state-of-the-art methods. Note that the IDE+Re-ranking method uses re-ranking to further improve the performance. The top-1 accuracy and mAP of SRQEN also outperform other state-of-the-art methods which don't use re-ranking.

On the more challenging LPW, we also configure the ablation study on large-scale dataset LPW. From Fig. 7, we can notice that the SRQEN still has stable performance and increase top-1 accuracy by 16.8% compared with the baseline. $+QM$ has a wealth of information for recognition. In order to verify the effectiveness of regional quality generation part, we performed two other comparisons. In the $+QFix$, it has a similar performance in large scale dataset which meets normal expectations. In the $+MP$, the growth of parameters does bring performance improvements but still not as good as SRQEN. The experimental results show that the SRQEN does improve performance effectively by automatically learning the alignment of different frames and the complementary information between different frames in a sequence.

D. CROSS-DATASET TESTING

Cross-dataset testing is a better way to confirm the validity of our model. Therefore to better understand how well our proposed system generalizes, we also perform cross-dataset testing. For comparing with other methods in cross-dataset evaluation, we adopt the same protocol as [12]. We train the SRQEN on the whole iLIDS-VID dataset, and use half of the PRID 2011 dataset for testing. The result is shown in Table 7. We can notice that the SRQEN outperforms the average pooling method of the baseline and other state-of-the-art methods by a large margin. The experimental result demonstrates the generalization and the robustness of the proposed method.

E. PARAMETER ANALYSIS

In this subsection, we evaluate the impact of several parameters in SRQEN on the iLIDS-VID dataset.

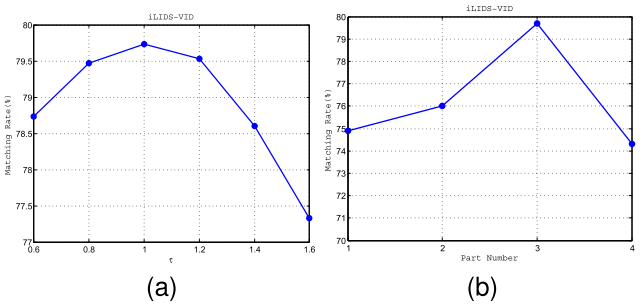


FIGURE 8. Parameter analysis on iLIDS-VID dataset. (a) Evaluation of the impact of the triplet loss margin τ. The change of matching rate with respect to τ is shown. (b) Evaluation of the impact of the part numbers in SRQEN. The change of matching rate with respect to part number is shown.

1) THE EFFECTIVENESS OF THE TRIPLET LOSS MARGIN τ

We first evaluate the effectiveness of the triplet loss margin τ in (4), and the result is shown in Fig. 8a. The matching rate first improves with the increase of τ, then the performance begins to drop when τ is larger than 1.0. In this paper, we set $\tau = 1.0$ in all the experiments. When τ varies within an appropriate scope (e.g. [0.8,1.2]), the change of the matching rate is relatively small. But if τ is set too small or too large, the performance drops significantly. On the one hand, if τ is small, the effectiveness of the triplet loss is limited, the distance between different persons may be still small. On the other hand, if τ is too large, the network may mainly concentrate on optimizing the triplet loss, and the frame-level features may be suboptimal, which also leads to a drop of performance. Choosing a suitable τ through the validation set is important to train the network.

2) THE EFFECTIVENESS OF PART NUMBER

Fig. 8b shows the results of SRQEN with different part numbers on the iLIDS-VID dataset. The performance is highest when the part number is set to 3, which shows that it is appropriate to divide the human body into three parts in Section II. If the part number is set to 1 or 2, it may be too coarse to generate regional features and the corresponding scores. When the human body is divided into more than 3 parts, the performance also drops a lot, this may be because the discriminative ability of each regional feature drops as the increase of part numbers.

F. COMPLEMENTARY INFORMATION EXTRACTION

We generated the quality scores of the test set using SRQEN used for intuitive exploration. Some partially occluded images are selected for evaluation. The results in Fig. 9 provide a vivid illustration of how the scores reflect the corresponding regional quality. The qualities of images' regions with some noise such as occlusion and deformation are poor and the corresponding scores are lower.

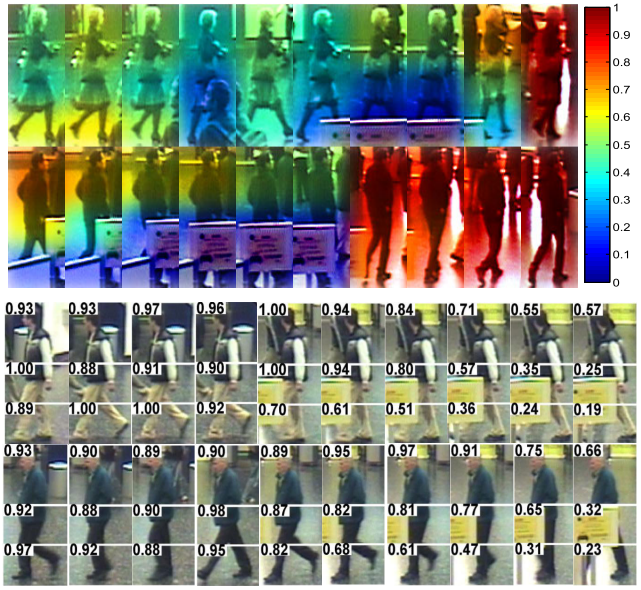


FIGURE 9. The quality scores of selected images' different parts. In the first row, the quality scores in [0,1] are mapped to the colors in the transition from blue to red. In the second row, the quality scores of each part are shown in the images. Best viewed in color.

Fig. 1 clearly demonstrates the advantages of SRQEN in the aggregation of sequence features. Compared with the average pooling method, it can make more effective use of complementary information in different images' parts of the sequence. The effective information in the region of one frame can make up for missing portions of the same region's information in other frames. The more robust feature representations can be extracted via SRQEN with the assist of partial quality scores and get better performance.

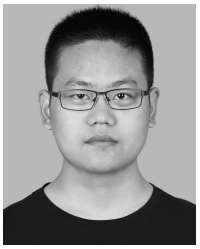
V. CONCLUSION

In this paper, we mainly focus on the problem of image misalignment and partial noises in video-based person re-identification task. To address these two problems, we propose the Spatial-transformed Regional Quality Estimation Network (SRQEN). SRQEN can learn to align different images and partial qualities of each image and aggregate the images' complementary partial information of different frames in an image sequence. We design an end-to-end training strategy with ingenious gradient design and jointly train the network with classification and verification signals. The SRQEN leads to the state-of-the-art results on PRID 2011 and iLIDS-VID. In addition, we also propose a large-scale and clean dataset named "Labeled Pedestrian in the Wild (LPW)" which contains 7,694 tracklets with over 590,000 images. The dataset, which has occlusion, large posture change, and huge age span, is more challenging and suitable for practical use.

REFERENCES

[1] N. Aparajita, P. K. Sa, S. K. Choudhury, S. Bakshi, and B. Majhi, "A neuromorphic person re-identification framework for video surveillance," *IEEE Access*, vol. 5, pp. 6471–6482, 2017.
[2] J. Guo, Y. Zhang, Z. Huang, and W. Qiu, "Person re-identification by weighted integration of sparse and collaborative representation," *IEEE Access*, vol. 5, pp. 21632–21639, 2017.
[3] T. Li, L. Sun, C. Han, and J. Guo, "Salient region-based least-squares log-density gradient clustering for image-to-video person re-identification," *IEEE Access*, vol. 6, pp. 8638–8648, 2018.
[4] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. ICCV*, Oct. 2007, pp. 1–8.
[5] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," 2017, *arXiv:1710.00870*. [Online]. Available: https://arxiv.org/abs/1710.00870
[6] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, Leeds, U.K., Sep. 2008, pp. 1–10.
[7] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[8] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, Oct. 2008, pp. 262–275.
[9] Z. Chen, X. You, B. Zhong, J. Li, and D. Tao, "Dynamically modulated mask sparse tracking," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3706–3718, Nov. 2017.
[10] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 144–151.
[11] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. WACV*, Mar. 2016, pp. 1–8.
[12] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1325–1334.
[13] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 701–716.
[14] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proc. CVPR*, Jun. 2016, pp. 1335–1344.
[15] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1249–1258.
[16] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7948–7956.
[17] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 907–915.
[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[19] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. ACCV*. Queenstown, New Zealand: Springer, 2010, pp. 501–512.
[20] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 247–267.
[21] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1345–1353.
[22] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, Aug. 2014, pp. 34–39.
[23] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, Jun. 2015, pp. 815–823.
[24] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.
[25] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[26] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. CVPR*, Jun. 2015, pp. 3908–3916.

[27] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," 2016, *arXiv:1605.09653*. [Online]. Available: https://arxiv.org/abs/1605.09653
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[32] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," in *Proc. ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 868–884.
[33] S. Karanam, Y. Li, and R. J. Radke, "Sparse Re-ID: Block sparsity for person re-identification," in *Proc. CVPR*, Jun. 2015, pp. 33–40.
[34] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," 2017, *arXiv:1707.00408*. [Online]. Available: https://arxiv.org/abs/1707.00408
[35] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. CVPR*, Jul. 2017, pp. 5790–5799.
[36] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. ECCV*. Zürich, Switzerland: Springer, 2014, pp. 688–703.
[37] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandin. Conf. Image Anal.* Ystad, Sweden: Springer, 2011, pp. 91–102.
[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, Jun. 2015, pp. 1116–1124.
[39] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. CVPR*, Jun. 2016, pp. 4724–4732.
[40] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 2017–2025.
[41] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and Fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2884–2893.
[42] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 79–88.
[43] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun, "Orientation driven bag of appearances for person re-identification," 2016, *arXiv:1605.02464*. [Online]. Available: https://arxiv.org/abs/1605.02464
[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456.
[45] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Proc. ICCV*, 2015, pp. 3810–3818.
[46] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4747–4756.
[47] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4743–4752.
[48] J. Dai, P. Zhang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," 2018, *arXiv:1802.07918*. [Online]. Available: https://arxiv.org/abs/1802.07918
[49] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1412–1424, Oct. 2018.
[50] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3652–3661.
[51] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li, "Cross dataset person re-identification," in *Proc. ACCV*. Singapore: Springer, 2014, pp. 650–664.



MING JIANG received the bachelor's degree in computer science and engineering from Beihang University, Beijing, China, in 2018, where he is currently pursuing the master's degree with the School of Computer Science and Engineering. His research interests include person re-identification and video analysis.



BIAO LENG received the bachelor's degree from the School of Computer Science and Technology, National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2009. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing. His current research interests include image processing, pattern recognition, and 3D model retrieval.



ZHIJUN MENG received the B.E. and Ph.D. degrees from Beihang University, Beijing, China, in 2004 and 2009, respectively, where he is currently an Associate Professor. His research interests include object recognition, image processing, and automatic control of unmanned systems.



GUANGLU SONG received the bachelor's degree in computer science and engineering from Beihang University, Beijing, China, in 2017. He is currently pursuing the master's degree with the School of Computer Science and Engineering, Beihang University. His research interests include person re-identification and object detection.