Hindawi
Journal of Sensors
Volume 2022, Article ID 3852243, 12 pages
https://doi.org/10.1155/2022/3852243

Research Article

IIPA-Net: Joint Illumination-Invariant and Pose-Aligned Feature Learning for Person Reidentification

Senquan Yang,[1,2] **Fan Ding**,[1] **Haoxiang Wen**,[1] **Pu Li**,[1,3] **and Songxi Hu**[1,3]

[1]*School of Intelligent Engineering, Shaoguan University, Shaoguan 512005, China*
[2]*Foshan Nanhai Guangdong Technology University CNC Equipment Cooperative Innovation Institute, Foshan 528225, China*
[3]*Guangdong Provincial Key Laboratory of Technique and Equipment for Macromolecular Advanced Manufacturing, South China University of Technology, 510641, China*

Correspondence should be addressed to Pu Li; lipu_300@sgu.edu.cn

Received 3 March 2022; Revised 9 April 2022; Accepted 26 May 2022; Published 21 June 2022

Academic Editor: Han Wang

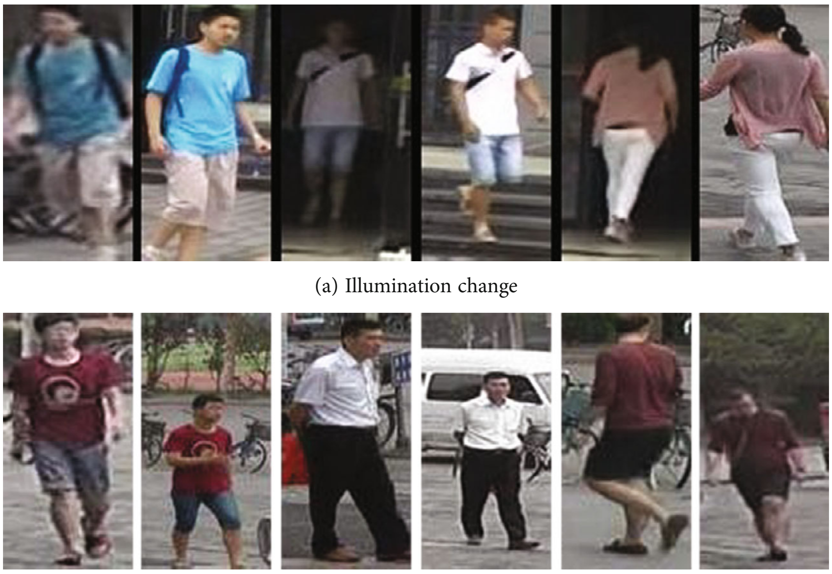
Person reidentification (re-id) has gained significant progress and aroused great interest in computer vision. However, due to the effect of weak illumination and poor alignment, person re-id is still a challenging task. Many previous works focus on either illumination enhancement methods or pose estimation. However, those methods are difficult to apply in real-world scenarios, which usually contain various interference factors. To improve the performance of re-id, we propose an Illumination-Invariant and Pose-Aligned Network (IIPA-Net). The illumination change is handled by a retinex decompose network, and the pose variation problem is solved by a local feature matching method. Based on the multimodal nature of a person, we propose a part attention module to optimize the global feature. Finally, a data-driven training strategy is proposed to train the proposed architecture effectively. Experiments show that the proposed framework outperforms other state-of-the-art approaches on both normal- and low-light datasets.

1. Introduction

Person reidentification (re-id) is aimed at identifying a specific person (probe query image) from a gallery of candidate images captured by multiple cameras with overlap or nonoverlap fields of view. The increasing need for safety and security, combined with the growing availability of surveillance cameras, makes person reidentification an increasingly explored area [1]. However, it is very challenging since the interest person images captured by surveillance cameras usually have significant variations in different viewpoints, illumination, human pose, and so on [2]. Low resolution, partial occlusions, and blurring increase the difficulty of person re-id [3].

Since person images are captured by different cameras under unknown lighting conditions, the appearance of the same person contains various variants, making the re-id task extremely difficult. In order to eliminate the effect of illumination, many methods rely on the statistics of color distribution and project image to color constant space [4]. However, the prior information of lighting is unpredictable in real-world scenarios. An alternative solution is to simulate the real-world illumination and use data augmentation techniques, which is expensive and needs a lot of labeled data [5]. Pose misalignment, which is caused by changed viewpoint or inaccurate detection boxes, is another interference of person re-id framework [6]. A straightforward solution to this pose variation is to apply human pose estimation, which parses a person image into different semantic parts. However, pose estimation requires massive labeled data to train the model [7]. What is more, the re-id accuracy degrades substantially for inaccurate estimation. Figure 1 shows some examples of illumination change and pose misalignment.

Convolutional neural networks (CNNs), which have powerful representation and invariant embedding capabilities, have boosted the performance of person re-id [8]. CNN-based person re-id methods can be divided into two aspects: discriminative feature representation learning and deep metric learning [9]. In the first category, majority of the methods generally concentrate on extracting discriminative



(a) Illumination change

(b) Pose variation

Figure 1: Examples of challenging images in re-id.

features, then formulate the person re-id as a classification problem [10]. In the second category, a robust metric between positive (the same) and negative (the different) persons is learned to deal with the matching problem [11]. In this paper, we focus on extracting discriminative feature representation. To achieve this aim, we propose a joint CNN framework that couples global and local feature learning to suppress interference, especially illumination and pose variations. Firstly, motivated by deep retinex illumination decomposition [12], we adopt a lightweight estimation to eliminate the effect of illumination and enhance the global person feature. Secondly, inspired by AlignedReID++ [13], which aligns local information to learn more discriminative features, we introduce a local feature matching to align different parts of person image, which is able to solve the pose variation problem. We find that the illumination-invariant feature can guide the local feature matching to align different person image parts. Thirdly, since the detected person has two significant modes [14], we concatenate the low-level feature of CNNs and the two-peak Gaussian map to design an attention mechanism. Consequently, the proposed IIPA-Net can boost the performance of the re-id in both normal- and low-light datasets. In summary, the contributions of this paper are threefold:

(i) We build a novel network framework, which contains a retinex decomposition net and a weight-shared Resnet50 backbone CNN and achieves illumination-invariant and pose-aligned re-id

(ii) We propose a part attention module to reweight the CNN output and extract the most informative parts of a person

(iii) A data-driven training strategy is introduced to train the network effectively and speed up the training process

2. Related Work

The main challenges of reidentification are changes in illumination, viewpoint, and pose across cameras. Many works focus on extracting the most discriminative visual feature of a person, including color [14], texture [15], and shape [16]. Kviatkovsky et al. [14] use shape context descriptors as a color-based signature to represent a person, which is divided into two significant modes. However, they assume that the silhouette of a person can be always obtained, which is not the case in real-world applications. Deep learning has revolutionized the techniques for person reidentification [17]. Li et al. [18] successfully apply deep learning to extract the features for person reidentification. Xiao et al. [19] propose a new deep learning framework that jointly handled both person detection and reidentification in a single convolutional neural network. Wu et al. [20] improve the discriminative feature representation of CNNs by exploiting unlabeled tracklets. The major limitation of this framework is that they either have handcrafted features or employ single scene images, thus making them less robust to various lighting conditions and changed human pose. Retinex theory is widely used for illumination estimation [21]. Many retinex-based re-id algorithms had achieved competitive performances [22, 23]. Specially, Liao et al. utilize the retinex transform and a scale invariant texture operator to handle illumination variations [23]. Huang et al. propose a retinex decomposition network to address the illumination variation problem and achieved a competitive re-id performance in low-light condition [22].

In [24], a new synthetic dataset, which contains hundreds of illumination conditions, is introduced to simulate the real-world lighting. The above methods reduced the adverse effects of illumination variant. However, they ignore the matching of local feature and failed to learn the aligned information, which effectively eliminate the influence of pose variant.

To reduce the negative impact of pose variant, some works apply human pose estimation to extract pixel-level body regions [8, 25]. Zheng et al. adopt the pose estimation confidence of input image to build a pose-invariant embedding (PIE) descriptor [8]. In [25], Zhao et al. represent a person with a discriminative feature, which is learned from different semantic regions of a person. On the other hand, some works focus on utilizing horizontal stripes or grids to extract pose-invariant features [13, 26]. Sun et al. design a Part-based Convolutional Baseline (PCB) network to learn discriminative part-level features [26]. Using the dynamic programming to match horizontal stripes of person images, Luo et al. propose a deep model to address the misalignment issue [13]. Additionally, Miao et al. propose an occluded person re-id framework by incorporating the pose information [27]. In spite of the great progress in re-id performance, the above methods still could be optimized by integrating the advantages of different architectures.

Different from existing frameworks, we focus on addressing issues of illumination and pose change simultaneously. Then, we propose a novel framework that is able to learn illumination invariance and pose alignment in a multitask manner.

3. Methodology

In this section, we firstly describe the retinex decomposition net and the part attention module. Then, the details of the proposed structure and training strategy are introduced.

3.1. Retinex Decomposition Net. To simulate the human color perception, retinex theory decomposes the observed image into two components: reflectance and illumination [21]. Mathematically, the source image **S** can be denoted as follows:

$$\mathbf{S} = \mathbf{R} \circ \mathbf{I}, \tag{1}$$

where **R** and **I** represent the reflectance and illumination components, respectively, and ° represents element-wise multiplication. The reflectance map described the intrinsic person property and is invariant to light change.

Thus, it is active to extract illumination-invariant discriminative features from the reflectance map. The illumination map, which represents various light environments, is harmful to re-id performance and ignored in this paper.

Unlike deep retinex net [12] that performs both reflectance and illumination decomposition to enhance low-light images, we only perform retinex decomposition net to extract the consistent feature of a person. As shown in Figure 2, the retinex decomposition net includes 8 layers. The first layer is a 3×3 convolutional layer, which extracts convolutional features from the input image. The second to sixth layers are 3×3 convolutional layers with a Relu activation function. The seventh layer is a 3×3 convolutional layer which maps **R** and **I** from feature space. The last layer is a sigmoid function that normalizes **R** and **I** to $[0, 1]$.

To extract **R** from different lightness images, the decomposition network is fed in paired normal/low-light images each time. During the training stage, the paired images, instead of their corresponding ground truth, are taken to train the retinex decomposition net. However, it can predict **R** and **I** in the test stage.

The loss L_{R} for retinex decomposition net consists of reconstruction loss L_{recon} and invariable reflectance loss L_{ir}:

$$L_{\mathrm{R}} = L_{\mathrm{recon}} + \lambda_{\mathrm{ir}} L_{\mathrm{ir}}, \tag{2}$$

where λ_{ir} is used to balance the consistency of reflectance. The reconstruction loss L_{recon} is defined as

$$L_{\mathrm{recon}} = \sum_{i=\mathrm{low,normal}} \sum_{j=\mathrm{low,normal}} \lambda_{ij} L_{\mathrm{ir}} \left\| \mathbf{R}_i \circ \mathbf{I}_j - \mathbf{S}_j \right\|_1, \tag{3}$$

where $\mathbf{S}_{\mathrm{low}}$ and $\mathbf{S}_{\mathrm{normal}}$ denote the input low-light and normal-light images, respectively. $\mathbf{R}_{\mathrm{low}}$ and $\mathbf{I}_{\mathrm{low}}$ denote the reflectance and illumination of $\mathbf{S}_{\mathrm{low}}$, as well as $\mathbf{R}_{\mathrm{normal}}$ and $\mathbf{I}_{\mathrm{normal}}$ of $\mathbf{S}_{\mathrm{normal}}$. The invariant reflectance loss L_{ir} is defined as

$$L_{\mathrm{ir}} = \| \mathbf{R}_{\mathrm{low}} - \mathbf{R}_{\mathrm{normal}} \|_1. \tag{4}$$

3.2. Part Attention Module. In order to extract discriminative features, many re-id methods introduce the attention mechanism to highlight the informative parts of person images, while suppressing cluttered background [9, 28]. The goal of the attention mechanism is to produce a saliency map to reweight CNN output. Given a 3-D $X \in \mathbb{R}^{C \times H \times W}$, where C, H, and W indicate the number of pixels in the channel, height, and width dimensions, respectively, the reweight process can be formulated as

$$\mathbf{Y} = A(\mathbf{X}) \odot \mathbf{X}, \tag{5}$$

where **Y** is the reweighted map and $A(\mathbf{X})$ is the output of the attention module. Combined with the state-of-the-art detector, there is an intuitive assumption that the detected persons lie in the middle of images. In real-world scenarios, a person usually has different clothing for lower and upper parts. Based on their multimodal nature, we introduce a two-peak Gaussian map M_f, defined as Equation (6), to deal with the intradistribution of person appearance:

$$M_f = (2\pi)^{-1/2} |\Sigma|^{-1/2} \left(e^{-1/2 (r-\mu_1)^T \Sigma^{-1} (r-\mu_1)} + e^{-1/2 (r-\mu_1)^T \Sigma^{-1} (r-\mu_1)} \right), \tag{6}$$

where $\mu_1 = [H/3, W/2]$ and $\mu_2 = [2 \times H/3, W/2]$ represent the peak centers of the Gaussian map.

As shown in Figure 3, we concatenate M_f and the 4th layer of Resnet-50. Subsequently, six 3×3 convolution layers are added to extract the discriminative feature. Finally, a softmax classifier is implemented with a Fully Connected (FC) layer.

3.3. IIPA-Net Architecture. As shown in Figure 4, the proposed IIPA-Net can be divided into two parts: global branch and local branch.

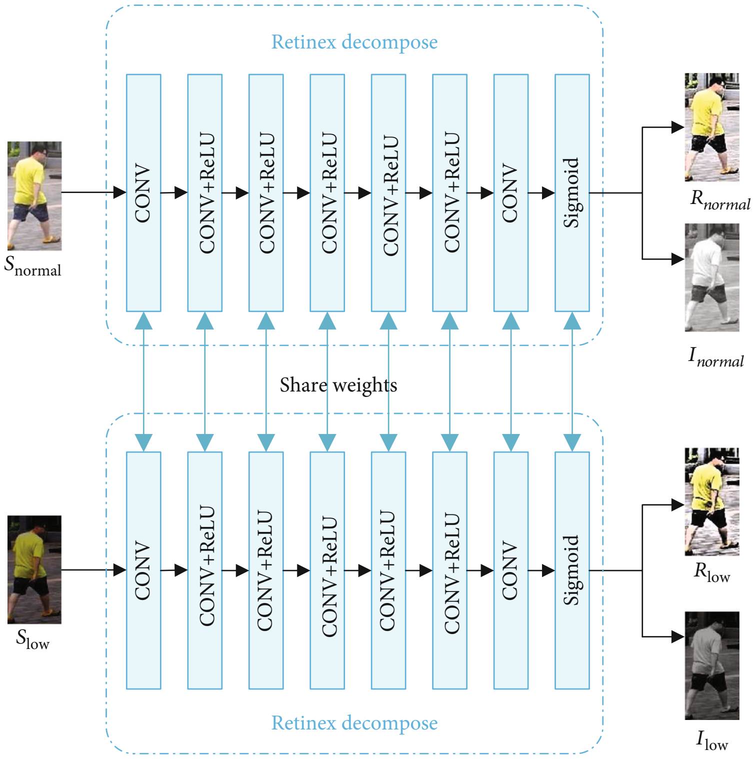


FIGURE 2: Retinex decomposition net.

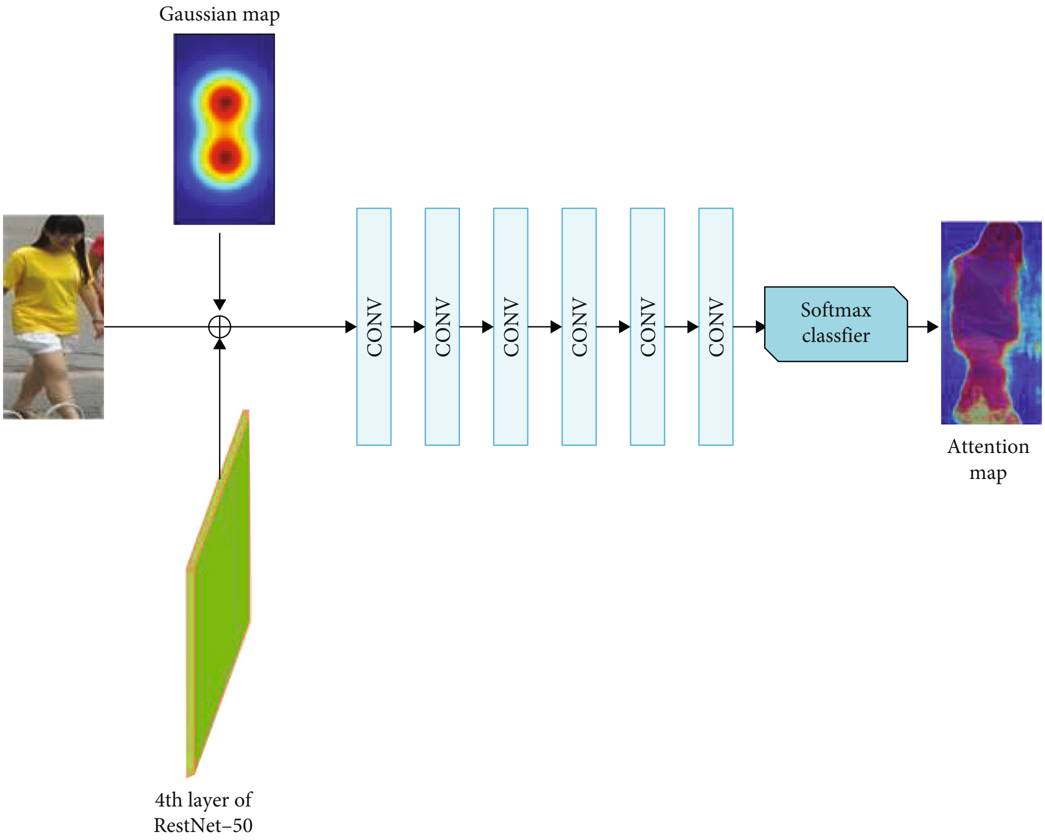


FIGURE 3: The proposed part attention module.

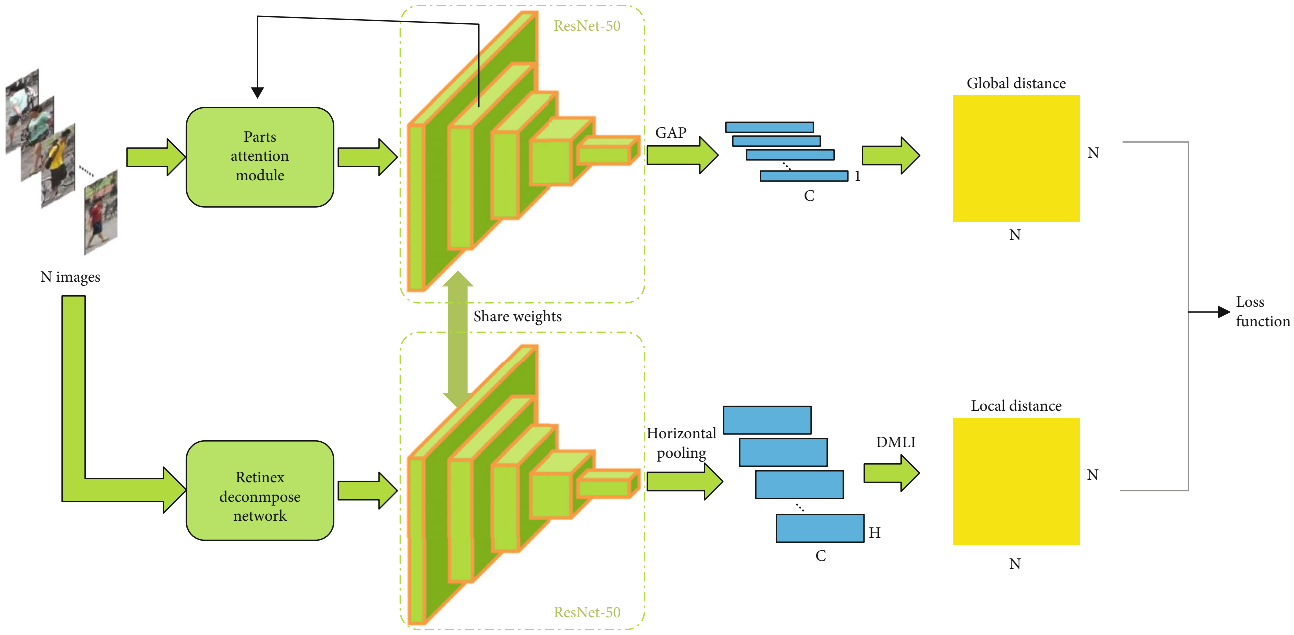


FIGURE 4: The proposed network architecture.

For the first branch, the most discriminative image parts of a person are extracted by the part attention module. In the second branch, the person images are enhanced by preserving the reflectance map of retinex decomposition net. Both of the two branches are sent into the weight-shared Resnet50 backbone CNNs, which makes the proposed model more flexible and easy to train. The output of Resnet50 is a$\mathbb{R}^{c\times h\times w}$ feature map, where c represents the feature channel and $h \times w$ is the spatial size. We extract a global discriminative feature vector $I \in \mathbb{R}^{c\times 1}$ using Global Average Pooling (GAP). Then, the global feature distance can be calculated by

$$d_I(A, B) = \|I_A - I_B\|_2, \tag{7}$$

where I_A and I_B denote the global feature of images A and B. The global feature is able to learn holistic information from the person image. However, it fails to address the pose-misalignment issue for the reason that the local representation is still unexploited. To learn the pose-aligned local feature, the output feature map of Resnet50 is transferred into $c \times h$ size using horizon horizontal average pooling. Let $P_A = \{p_A^1, p_A^2, \cdots, p_A^h\}$ and $P_B = \{p_B^1, p_B^2, \cdots, p_B^h\}$ denote the local feature of images A and B. We can have the distance of the ith vertical part of A and jth vertical part of B as follows:

$$d(i, j) = \frac{e^{\|p_A^i - p_B^i\|_2} - 1}{e^{\|p_A^i - p_B^i\|_2} + 1}. \tag{8}$$

We further have the distance matrix D, where its elements are $d(i, j)$. As described in [13], the local pose-aligned feature distance $d_p(A, B)$ can be derived by dynamically matching local information (DMLI), which could dynamically align different part features. Finally, we obtain the total distance of A and B by

$$d(A, B) = d_I(A, B) + d_p(A, B). \tag{9}$$

The total loss function of the framework is

$$L_{\text{total}} = L_{ID} + L_T^I + L_C^P, \tag{10}$$

where L_{ID} and L_T^I denote softmax loss and triplet loss [29] of the global feature and L_C^P denotes the circle loss [30] of the local pose-aligned feature. The performance of different loss functions is described in Section 4.3.

3.4. Training the Network. Since there is a lack of explicit ground truth for the training part attention module and retinex network, it is difficult to optimize the network for various scenes. Therefore, we try to train the network in a date-driven way. The whole network is trained in four stages, as illustrated in Algorithm 1.

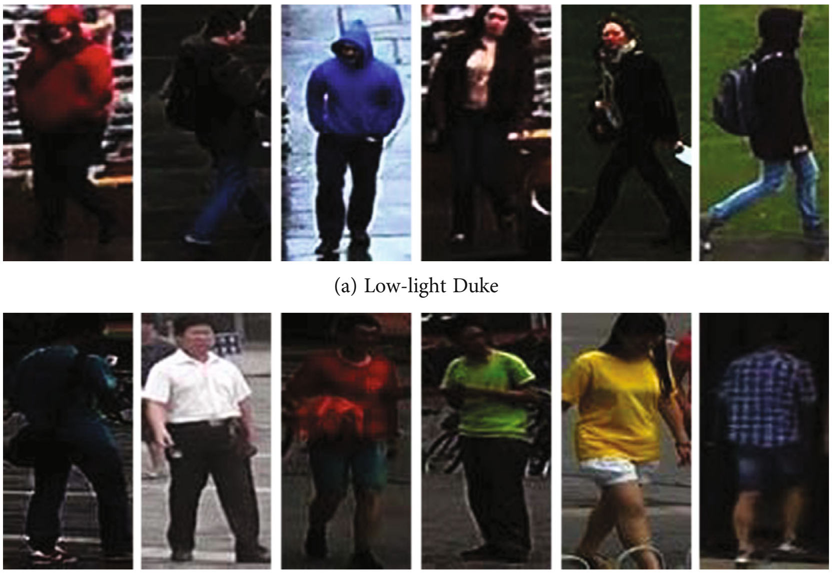
(i) First, the backbone network, Resnet-50, is initialized by the ImageNet [31] pretrained model and trained to convergence under the supervision of triplet loss

(ii) Second, the synthetic low-light image sets based on PASCAL VOC, together with their original images, are fed to the Retinex decompose network, as described in Section 3.3. This training step is finished after 200 epochs

(iii) Third, all the layers in Resnet-50 are fixed; only the part attention module is trainable. Then, the IIPA-Net is retained with the softmax and triplet loss on

1. The shared-weights Resnet-50 is trained to convergence with triplet loss.
2. All synthetic images, together with their original images, are fed into the Retinex decomposition network.
3. Parts attention module is trained using the training images set.
4. The whole network is fine-tuned with Equation (10).

Algorithm 1: Training steps of the proposed network.



(a) Low-light Duke

(b) Low-light Market

Figure 5: Examples of synthetic low-light image.

the training set. The learning rate is decayed for 40 epochs

(iv) Finally, we set all the layers trainable and fine-tune the IIPA-Net to convergence again

4. Experiments

4.1. Datasets and Evaluation Measures. Our experiments are based on two real-world and popular person re-id datasets: Market1501 [32] and DukeMTMC-reID [33]. To better present the advantages of the proposed illumination-invariant feature, we adopt two manual low-light re-id datasets named low-light Market and low-light Duke. The Market1501 includes 32,668 images of labeled people captured by six cameras. Specially, there are 12,936 images of 751 identities in the training set and 19,732 images of 750 identities in the testing set. The DukeMTMC-reID contains 25,272 images, which are extracted from the DukeMTMC dataset [34] captured by eight cameras. There are 6,522 images of 702 identities in the training set and 18,750 images of 1110 identities in the testing set. The low-light Market and low-light Duke are built from Market1501 and DukeMTMC-reID, respectively. Following [22], we use gamma correction to simulate low-light conditions. Each image in the datasets is processed with a gamma value, which is randomly picked from $\{1, 2, 3, 4\}$. Figure 5 shows examples of synthetic low-light images. To evaluate the performance of different algorithms, we use Cumulative Matching Characteristic (CMC) curves and mean Average Precision (mAP) [32] as the evaluation criteria. CMC is defined as a function of Rank-r [35].

$$q(r) = \frac{|C(r)|}{|\mathcal{P}_{\mathrm{g}}|}, \tag{11}$$

where $|\mathcal{P}_{\mathrm{g}}|$ represents the total number of person images in the gallery, and the query set $C(r)$ is defined as

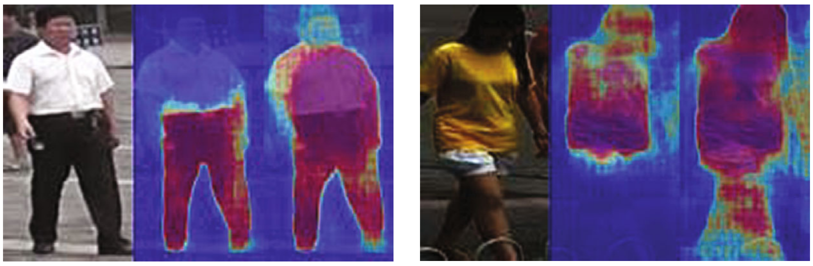
$$C(r) = \{p_i : \text{rank } (p_i) \le r\}, \quad \forall p_i \in \mathcal{P}_g. \tag{12}$$

mAP is calculated based on the Average Precision (AP) and defined as

$$\text{mAP} = \frac{\sum_{k=1}^{n} \text{AP}(k)}{n}, \tag{13}$$

where $\text{AP}(k)$ represents the precision-recall curve area of the kth query and n represents the size of the query set.

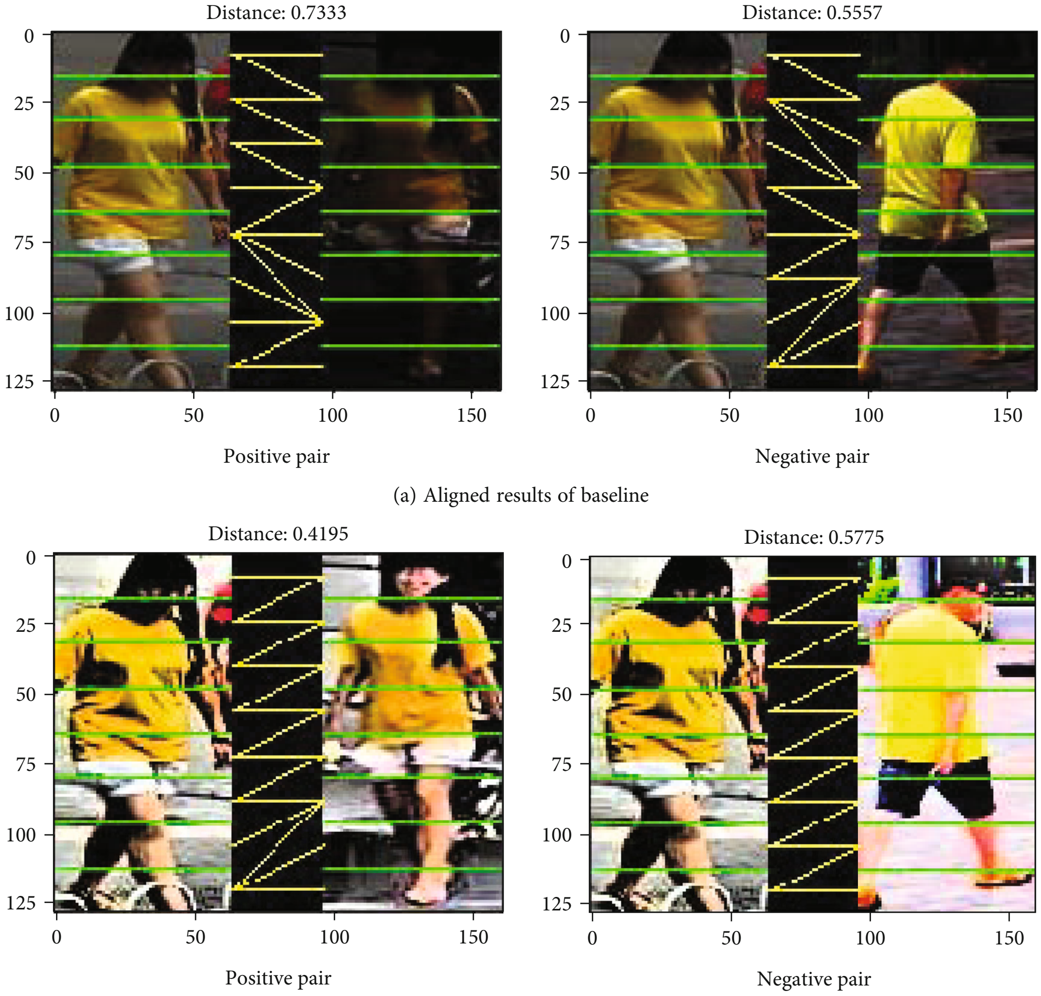
4.2. Experimental Setup. We implement all experiments using an Intel Xeon e5-2630 v3 2.4 GHz machine with



(a) Normal-light example

(b) Low-light example

Figure 6: Illustration for the parts attention module. The first column shows the input images. The second and third columns show the attention map results of normal and two-peak Gaussian.



Positive pair

Negative pair

(a) Aligned results of baseline

Positive pair

Negative pair

(b) Aligned results of our framework

Figure 7: The example aligned results of a positive pair and a negative pair from low-light Market.

32 GB RAM and one NVIDIA GTX Titan 12 GB GPU. The training patch size is set to be 32; λ_{ir} is set to be 0.001. λ_{ij} is set to 1, when $i = j$. Otherwise, λ_{ij} is 0.001. Each input image is resized to 256×128. Random horizontal flipping and cropping tricks are preformed to augment data. We use Adam optimizer with learning rate 10^{-4}.

4.3. Experimental Results. In this subsection, we firstly evaluate the part attention module. The two-peak Gaussian map can better guide the main body information of a person. Then, the effect of low light is analyzed. We can see that the low-light condition has a negative impact on pose alignment. Then, we evaluate the performance of our proposed IIPA-Net compared with other state-of-the-art re-id methods.

4.3.1. Evaluation of Part Attention. To better illustrate the effect of the proposed part attention module, we visualize the attention maps of the model with normal and two-peak Gaussian maps. In Figure 6, we can observe that the two-peak Gaussian map can pay attention to both upper and down parts of a person, while the normal one only to either upper (Figure 6(a)) or down (Figure 6(b)) part. The introduction of two-peak Gaussian makes part attention work more effective with the multimodal nature of a person. Figure 6 third columns show that the proposed part attention is able to produce similar predicted attention under different light conditions.

4.3.2. Effect of Low Light. As shown in Figure 7(a), using AlignedReID++ [13] as the baseline model, the fifth block of the left image is aligned to the fourth and sixth blocks of the right image and the distance of the two images is 0.7333, which is greater than the negative pair (0.5557). However, after decomposing the illumination, our proposed method is able to align the head, chest, foot, etc., of the positive pair images, and the distance is reduced to 0.4195, which is less than the negative pair (0.5775), as illustrated in Figure 7(b). The wrong connections of the baseline can be attributed to the negative impact of the low illumination. This indicates that the proposed approach eliminates the effect of weak illumination and learns the illumination-invariant features.

4.3.3. Performance of Different Loss Functions. We train four models with softmax+triplet loss ($L_{ID} + L_T^I + L_T^P$), softmax +instance [36] loss ($L_{ID} + L_I^I + L_I^P$), softmax+circle loss ($L_{ID} + L_C^I + L_C^P$) and the proposed loss. The performance on Market1501 is presented in Table 1. L^I and L^P represent the loss of the global and local features, respectively. We can observe that Softmax+Instance and Softmax+Circle loss achieve the similar Rank-1 accuracy. Compared with Softmax+Triplet, the proposed loss improves the Rank-1 and mAP arropminately 0.3 and 0.2, respectively. We believe that the Circle loss works on some hard local features.

Table 1: The performance of different loss functions on Market1501.

Loss function	Rank-1	mAP
$L_{ID} + L_T^I + L_T^P$	95.9	90.1
$L_{ID} + L_I^I + L_I^P$	95.7	89.8
$L_{ID} + L_C^I + L_C^P$	95.6	90.0
$L_{ID} + L_T^I + L_C^P$	**96.2**	**90.3**

Bold: best results.

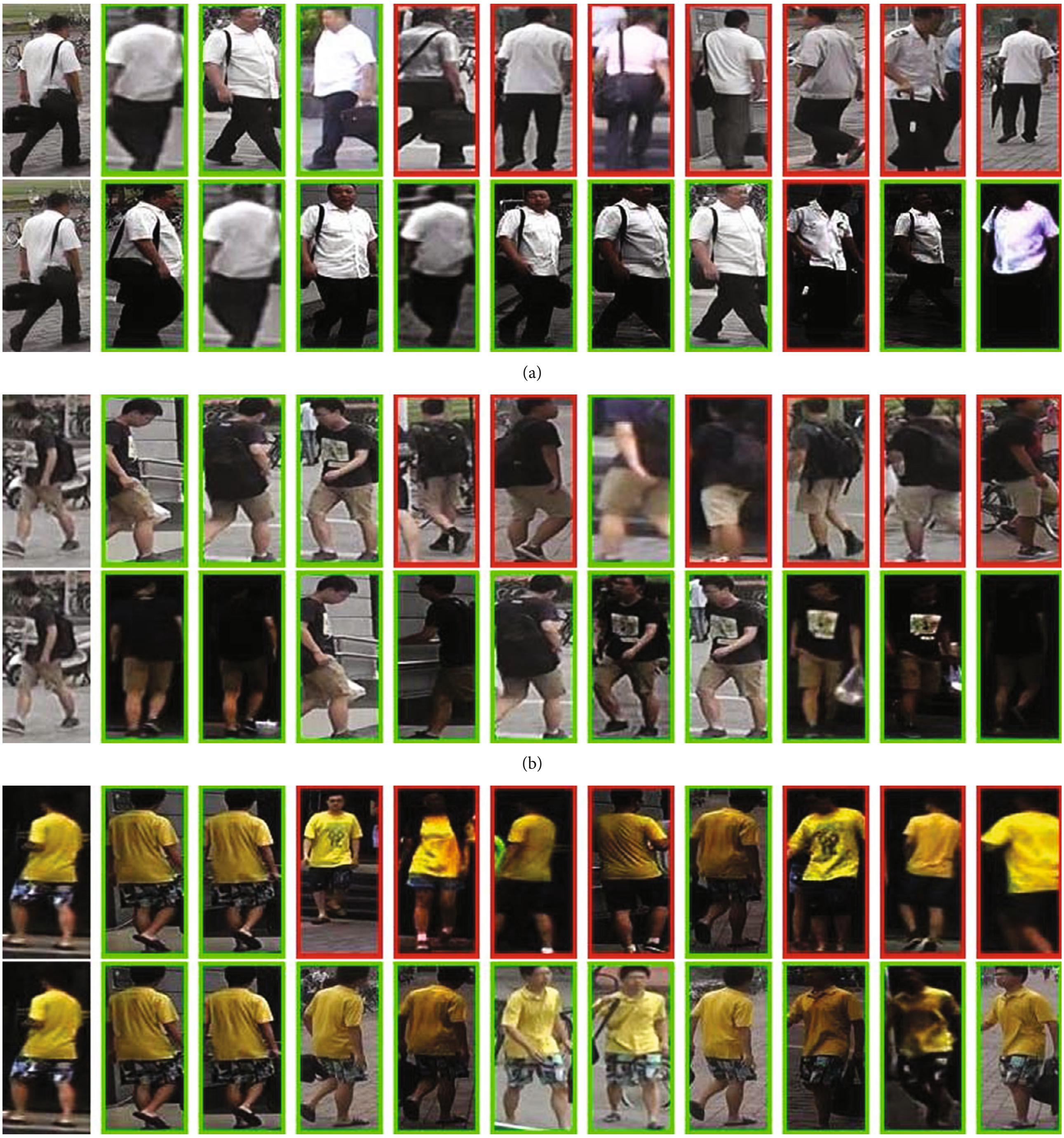
4.3.4. Comparison with State-of-the-Art. To evaluate the performance of the proposed IIPR-Net, we report the experimental results with some state-of-the-art methods. Our baseline is AlignedReID++ [13], which focuses on solving the pose change problem. In order to demonstrate the advantage of the proposed framework, we also report the results of baseline with a low-light enhancement method. Both training and testing image sets are enhanced with MSRCP [37] and then fed into the baseline.

Table 2: Experiment results of our framework compared to other state-of-the-art methods.

Dataset	Method	Rank-1	mAP
Market1501	IDE [41]	85.3	68.5
	Baseline [13]	92.8	89.4
	PCB [26]	92.3	77.4
	MHN-6 [3]	95.1	85.0
	DSA [38]	95.7	87.6
	FlipReID [39]	95.8	94.7
	st-ReID [40]	**98.0**	**95.5**
	IIPA-Net	96.2	90.3
DukeMTMC-ReID	IDE [41]	73.2	52.8
	Baseline [13]	80.7	68.0
	PCB [26]	81.7	66.1
	MHN-6 [3]	89.1	77.2
	DSA [38]	86.2	74.3
	FlipReID [39]	93.0	90.7
	st-ReID [40]	**94.5**	**92.7**
	IIPA-Net	90.8	83.3
Low-light Market	Baseline [13]	33.4	14.1
	Baseline+MSRCP	49.4	15.7
	PCB [26]	48.5	16.2
	IIPA-Net	**60.5**	**27.7**
Low-light Duke	Baseline [13]	36.2	12.4
	Baseline+MSRCP	40.4	18.3
	PCB [26]	48.4	21.0
	IIPA-Net	**51.6**	**24.3**

Bold: best results.

As shown in Table 2, our proposed framework outperforms most state-of-the-art methods on all four datasets. Specially, the proposed framework achieves 96.2% Rank-1 for Market1501 and 90.8% Rank-1 for Duke MTMC-reID, outperforming other attention-based methods, i.e., MHN-6 [9] and DSA [38]. Although FlipReID [39] and st-ReID [40] achieve the best performance, the extra data, for instance, spatial and temporal information, are utilized to train the network. For low-light Market and Duke datasets, the Rank-1 accuracy of the proposed method is increased by 10.1% and 11.2%, and the mAP increased by 9.5% and 6.0%, respectively. This demonstrates that our joint framework not only eliminates the impact of low light but also explores pose-invariant local features for person re-id. Figure 8 depicts five examples of queries together with the top 10 retrieved results of baseline and IIPA-Net on the low-light Market dataset. As we can see, the IIPA-Net

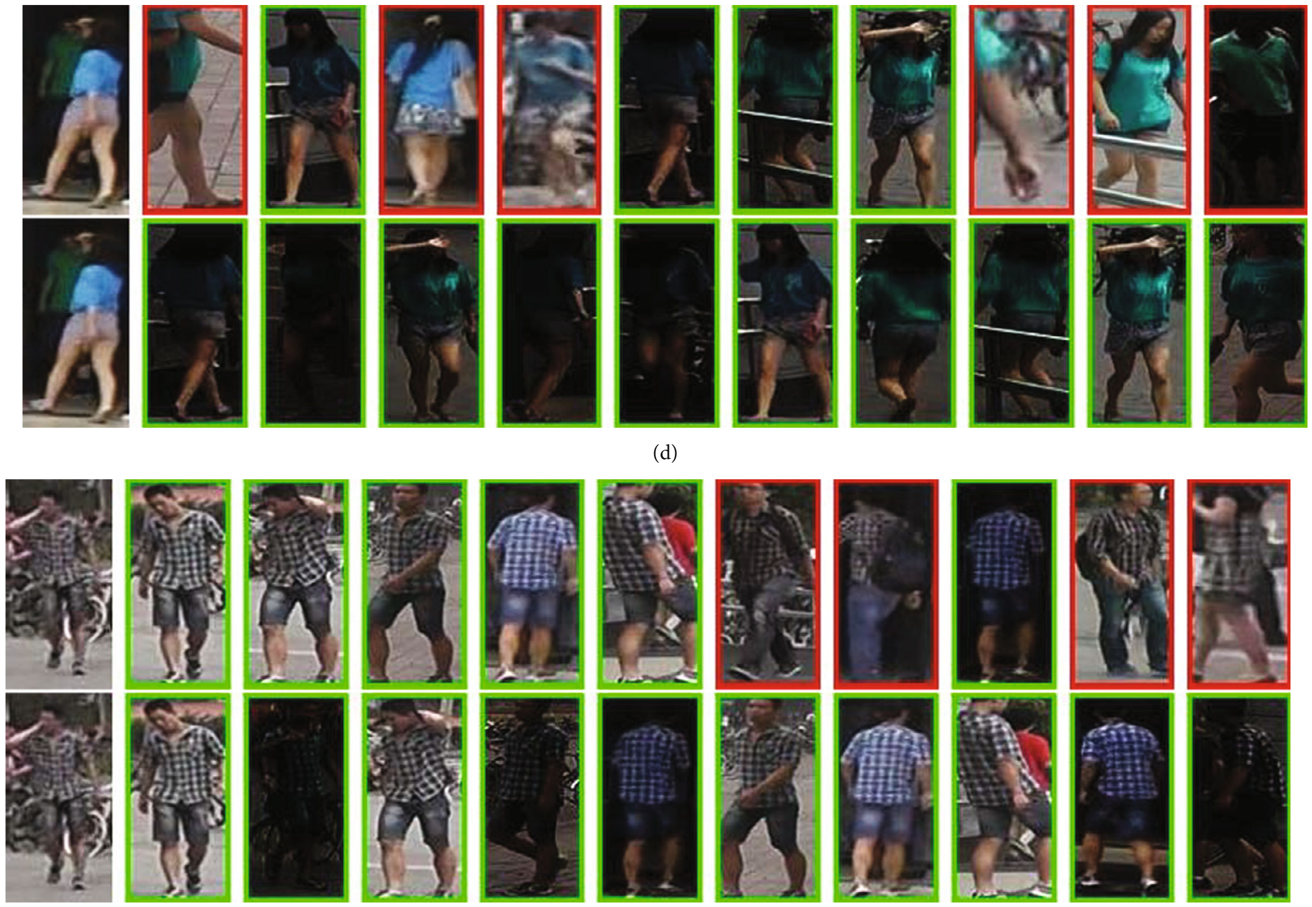


(a)

(b)

(c)

Figure 8: Continued.


(d)

(e)

Figure 8: Retrieved results of baseline and IIPA-Net on the low-light Market dataset. The first column images are the query. For each example, the upper row images are the results of baseline and the lower row images are of IIPA-Net. The images with a green border are the correct retrieved results, and those with a red border are the incorrect results.

Table 3: Ablation study on normal- and low-light market datasets.

Condition	Method	Rank-1	mAP
Normal	Ours w/o attention	94.4	89.5
	Ours w/o retinex	92.6	88.3
	Ours	**96.2**	**90.3**
Low light	Ours w/o attention	44.7	17.3
	Ours w/o retinex	56.3	22.4
	Ours	**60.5**	**27.7**

w/o: without; bold: best results.

outperforms the baseline and accurately retrieves the target in spite of illumination and pose variants.

4.3.5. Ablation Study. To verify the contribution of each component, we perform the ablation study on normal- and low-light Market datasets. Table 3 shows the results of each component of IIPA-Net. We note that the attention component achieves better results on the Market1501 dataset. However, retinex is better in low-light conditions. The combination of the retinex and attention achieves the best performance on both datasets. The reason is that IIPA-Net is able to learn both illumination and pose-invariant features.

5. Conclusions

In this paper, we proposed a jointly illumination-invariant and pose-aligned learning framework for person re-id. Motivated by retinex theory, we introduce a retinex decomposition net to eliminate the impact of different lights and extract an illumination-invariant feature. To tackle the problems of pose alignment, dynamically matching local information is utilized to align local feature, which is transferred from the deep learning feature map. Based on the nature of a person, we proposed a part attention mechanism to extract the most discriminative global feature. The joint framework is trained in a four-stage fashion. Experiments demonstrate that the proposed framework achieves better performance on both normal- and low-light datasets. In the future, we will focus on long-term re-id scenarios which present more complex scene variations.

Data Availability

All data included in this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant 52105268, Natural Science Foundation of Guangdong Province under Grant 2022A1515011409, Key Platforms and Major Scientific Research Projects of Universities in Guangdong under Grants 2019KTSCX161 and 2019KTSCX165, Key Projects of Natural Science Research Projects of Shaoguan University under Grants SZ2020KJ02 and SZ2021KJ05, Project of Guangdong Provincial Key Laboratory of Technique and Equipment for Macromolecular Advanced Manufacturing under Grant 2020kfkt07, and the Science and Technology Program of Shaoguan City of China under Grants 2019sn056, 200811094530423, 200811094530805, and 200811094530811.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[2] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto- reid: searching for a part-aware convnet for person re- identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3749–3758, Seoul, Korea (South), 2019.

[3] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, 2019.

[4] R. Prates, C. R. S. Dutra, and W. R. Schwartz, "Predominant color name indexing structure for person re-identification," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 779–783, Phoenix, AZ, USA, 2016.

[5] A. J. Ma, J. Li, P. C. Yuen, and P. Li, "Cross-domain person reidentification using domain adaptation ranking svms," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1599–1613, 2015.

[6] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re- identification," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3980–3989, Venice, Italy, 2017.

[7] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 542–551, Seoul, Korea (South), 2019.

[8] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4500–4509, 2019.

[9] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 371–381, Seoul, Korea (South), 2019.

[10] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6036–6046, Salt Lake City, UT, USA, 2018.

[11] Y. Lin, L. Zheng, Z. Zheng et al., "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, no. C, pp. 151–161, 2019.

[12] W. Chen, W. Wenjing, Y. Wenhan, and L. Jiaying, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference*, British Machine Vision Association, 2018.

[13] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "AlignedReID++: dynamically matching local information for person re- identification," *Pattern Recognition*, vol. 94, pp. 53–61, 2019.

[14] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[15] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Personal identification based on iris texture analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1519–1533, 2003.

[16] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *IEEE International Conference on Computer Vision*, pp. 1–8, Rio de Janeiro, Brazil, 2007.

[17] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Computer Vision and Pattern Recognition*, pp. 3908–3916, Boston, MA, USA, 2015.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: deep filter pairing neural network for person re-identification," in *IEEE International Conference on Computer Vision*, pp. 152–159, Columbus, OH, USA, 2014.

[19] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3376–3385, Honolulu, HI, USA, 2017.

[20] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5177–5186, Salt Lake City, UT, USA, 2018.

[21] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, 1977.

[22] Y. Huang, Z. J. Zha, X. Fu, and W. Zhang, "Illumination-invariant person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, pp. 365–373, Nice France, 2019.

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re- identification by local maximal occurrence representation and metric learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2197–2206, Boston, MA, USA, 2015.

[24] S. Bak, P. Carr, and J. F. Lalonde, "Domain adapta- tion through synthesis for unsupervised person re- identification," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C.

Sminchisescu, and Y. Weiss, Eds., pp. 193–209, Springer International Publishing, Cham, 2018.

[25] H. Zhao, M. Tian, S. Sun et al., “Spindle net: person re-identification with human body region guided feature decomposition and fusion,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 907–915, Honolulu, HI, USA, 2017.

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

[27] J. Miao, Y. Wu, and Y. Yang, “Identifying visible parts via pose estimation for occluded person re-identification,” *IEEE Transactions on Neural Networks and Learning Systems*, no. article 3059515, pp. 1–11, 2021.

[28] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in *2018 IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition*, pp. 2285–2294, Salt Lake City, UT, USA, 2018.

[29] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017, http://arxiv.org/abs/1703.07737.

[30] Y. Sun, C. Cheng, Y. Zhang et al., “Circle loss: a unified perspective of pair similarity optimization,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6397–6406, Seattle, WA, USA, 2020.

[31] O. Russakovsky, J. Deng, H. Su et al., “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: a benchmark,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, Santiago, Chile, 2015.

[33] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3774–3782, Venice, Italy, 2017.

[34] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *2016 European Conference on Computer Vision (ECCV)*, G. Hua and H. J’egou, Eds., pp. 17–35, Springer International Publishing, Cham, 2016.

[35] P. Grother, R. J. Micheals, and P. J. Phillips, “Face recognition vendor test 2002 performance metrics,” in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA’03*, pp. 937–945, Guildford, UK, 2003.

[36] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y. D. Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–23, 2020.

[37] A. B. Petro, C. Sbert, and J. M. Morel, “Multiscale retinex,” *Image Processing On Line*, vol. 4, pp. 71–88, 2014.

[38] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, “Densely semantically aligned person re-identification,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 667–676, Long Beach, CA, USA, 2019.

[39] X. Ni and E. Rahtu, “Flipreid: closing the gap between training and inference in person re-identification,” in *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6, Paris, France, 2021.

[40] G. Wang, J. Lai, P. Huang, and X. Xie, “Spatial-temporal person re-identification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8933–8940, 2019.

[41] L. Zheng, Y. Yang, and A. Hauptmann, “Person re- identification: past, present and future,” 2016, http://arxiv.org/abs/1610.02984.