Illumination-Invariant Person Re-Identification

Yukun Huang
kevinh@mail.ustc.edu.cn
University of Science and Technology of China

Zheng-Jun Zha*
zhazj@ustc.edu.cn
University of Science and Technology of China

Xueyang Fu
xyfu@ustc.edu.cn
University of Science and Technology of China

Wei Zhang
davidzhang@sdu.edu.cn
Shandong University

ABSTRACT

Due to the effect of weak illumination, person images captured by surveillance cameras usually contain various degradations such as color shift, low contrast and noise. These degradations result in severe discriminant information loss, which makes the person re-identification (re-id) more challenging. However, existing person re-identification approaches are designed based on the assumption that the pedestrians images are under well lighting conditions, which is impractical in real-world scenarios. Inspired by the Retinex theory, we propose a illumination-invariant person re-identification framework which is able to simultaneously achieve Retinex illumination decomposition and person re-identification. We first verify that directly using weak illuminated images can greatly reduce the performance of person re-id. We then design a bottom-up attention network to remove the effect of weak illumination and obtain the enhanced image without introducing over-enhancement. To effectively connect low-level and high-level vision tasks, a joint training strategy is further introduced to boost the performance of person re-id under weak illumination conditions. Experiments have demonstrated the advantages of our method on benchmarks with severe lighting changes and low light conditions.

KEYWORDS

Person re-identification, Retinex, Image enhancement, Weak illumination, Deep neural networks

ACM Reference Format:
Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. 2019. Illumination-Invariant Person Re-Identification. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350994

*Corresponding Author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350994

Figure 1: Person images under different illumination conditions.

1 INTRODUCTION

Recently, due to the powerful representation learning capabilities of deep convolutional neural networks (CNNs), significant progress has been achieved in person re-identification (re-id). To cope with the challenges caused by view variations, pose variations and occlusion, existing methods generally focus on extracting and matching local features of pedestrians [11, 28, 33, 35]. However, these methods rarely consider the influence of illumination variations. The main reason may be that most publicly available person re-id benchmarks, such as Market-1501 [36] and DukeMTMC-reID [38], were collected from limited areas within fixed time period. Therefore, the illumination diversity is not taken into consideration. While in real-world scenarios, illumination is a non-negligible interference factor. As shown in Figure 1, images with the same identity have different appearances due to various lighting conditions. Especially, weak illuminated images contain color shift, low contrast and noises. These degradations make person re-id task extremely difficult even for human beings.

A straightforward solution to this specific illumination problem is to use larger datasets that cover as many lighting conditions as possible. However, this solution is expensive and unpractical to label the massive surveillance videos to support supervised learning [24]. An alternative solution is to utilize data augmentation techniques, such as color jitter and gamma correction. To better simulate the real-world lighting conditions, Bak *et al.* [2] collected varieties of High Dynamic Range (HDR) environment maps for the rendering of virtual humans and built a new synthetic dataset

SyRI. The above mentioned methods can help the model learn illumination-invariant features in a data-driven fashion. However, the network architecture designed in this way may be non-optimal.

Different from existing methods, we eliminate the effect of illumination changes by explicitly decomposing the illumination map. Both the intrinsic pedestrian image and illumination-invariant features can be simultaneously obtained to improve the re-id performance. To achieve the above goal, it is natural to estimate illumination first. It is well known that estimating the illumination from single image is an ill-posed problem, which requires some priors and constraints to handle it. For instance, Retinex-based algorithms are required to satisfy the Lambertian scene assumption, and the illumination map should be piece-wise smooth. In previous works [7, 27, 30], Retinex theory was widely used for illumination estimation and low light enhancement. By adopting CNNs in a data-driven way, there are many methods [27, 30] achieved competitive results compared to other non-deep learning methods.

However, existing methods [7, 27, 30] of illumination estimation and low-light enhancement mainly focus on improving subjective visual quality, rather than serving subsequent high-level vision tasks. To tackle this issue, we first design a lightweight illumination estimation network to enhance the pedestrian images. Then a bottom-up attention mechanism is proposed to suppress the over-enhancement in extreme dark areas. To further combine the two incompatible low-level and high-level vision tasks, i.e., illumination estimation and person re-id, we intentionally build a novel network framework with a joint training strategy. Experimental results show that our approach have a substantial superiority compared to other existing methods in low-light re-id.

In summary, our contributions are as follow:

- We specifically build new low-light image datasets for the person re-id community. Based on our datasets, we verify that weak illumination can actually reduce the re-id performance.
- To obtain effective illumination-invariant features for re-id, we design a novel bottom-up attention mechanism to avoid the over-enhancement that usually contained in dark areas.
- We propose a novel CNNs framework for boosting the performance of person re-id under weak illumination conditions. Besides, a joint training strategy is introduced to effectively connect low-light image enhancement and person re-id tasks.

2 RELATED WORK

In recent years, person re-id, including image-based re-id [18, 19, 28] and video based re-id [4, 17], has received a lot of attention from both academic and industrial communities. The main difficulty in person re-id is how to learn a robust person feature representation to resist the interference caused by variations in views, pose, occlusion, illumination and so on. Inspired by existing low-light image enhancement methods, we focus on the illumination issue in the re-id task in this paper. In the following subsections, we will briefly review recent advances in person re-identification and low-light image enhancement.

2.1 Person Re-Identification

Most person re-id methods focus on reducing the adverse effects of various pose and view by combining global and local features. Specifically, the methods of modeling local features generally include explicitly extracting regions of body parts [11, 34], attention-based implicit local feature learning [15, 20, 35], heuristic pre-defined image partitioning, such as grid [1] and horizontal stride [28, 33], which usually accompanied by the steps of inter-area alignment [16, 33]. Due to coarse bounding box part detection, Kalayeh *et al.* [11] use semantic parsing network to perform pixel-level body regions extraction. Inspired by the attention mechanism, Zhao *et al.* [35] learn part-aligned local feature representations by utilizing the similarity between pedestrians without additional supervision information. Sun *et al.* [28] extract the part-level feature by dividing horizontal stripes. Zhang *et al.* [33] find that the alignment of local features according to the shortest path, which is calculated by a dynamic programming method. Additionally, Ge *et al.* [6] propose feature distilling generative adversarial network to learn pose-unrelated person representations without extra auxiliary pose information during inference. Despite the significant progress in re-id, above mentioned works mainly focus on addressing the feature matching problem of high-level semantic information while ignoring the matching of information on the underlying visual perception. However, in real-world scenarios, some underlying visual factors, such as illumination, resolution and weather, can also have a seriously negative impact on the person re-id task.

To the best of our knowledge, there are few re-id works focus on addressing the issue of illumination variations. To enrich the illumination diversity of training samples, Bak *et al.* [2] synthesize data from a variety of lighting conditions and use cycleGAN [39] to perform cross-domain transformations. Similar to the illumination issue, cross-resolution is another common problem [5, 9, 29] in real-world re-id. Eliminating the influence of illumination and cross-resolution can be considered as low-level vision tasks. In this work, we only focus on the illumination problem.

2.2 Low-Light Image Enhancement

In general, low-light image enhancement methods can be mainly categorized into three types: gamma correction-based, histogram equalization-based and Retinex-based. The first two kinds of methods only conduct intuitive and straightforward pixel-level mapping, while Retinex-based methods perform more image analysis and processing. The Retinex theory [13], which was proposed by Land in 1977, is firstly used for color constancy. Based on the simplification of Retinex, an observed image can be modeled as the pixel-wise multiplication of the illumination and the reflectance.

With the rapid development of deep learning in the computer vision community, a series of deep CNNs-based low-light image enhancement methods have been proposed. Lore *et al.* [21] propose a data-driven approach to simultaneously achieve low-light image enhancement and denoising. Lv *et al.* [22] implement a multi-branch network structure in which features from different levels are combined to extract rich and detailed information. In addition, deep neural networks based on Retinex theory have also appeared. Wei *et al.* [31] design the loss function for estimating reflectance and illumination from single input image. In [27], the classical Multi-scale Retinex algorithm is considered as a feed-forward convolutional neural network with different Gaussian convolution kernels, then MSR-Net is proposed to directly learn an end-to-end mapping from dark images to bright versions.

3 METHODOLOGY

In this section, we first introduce our Retinex decomposition net, the basic part of our approach. Then a bottom-up attention mechanism is well designed to solve the over-enhancement in dark areas. Moreover, a illumination-invariant feature learning framework is further proposed to connect Retinex decomposition and person re-id.

3.1 Retinex Decomposition Net

As illustrated in Figure 2, our Retinex decomposition network contains two sub-networks: **Light Estimation Net** (LE-Net) and **Light Decomposition Net** (LD-Net). The former is used to produce the illumination map of the original image, while the later aims to generate the reflectance.

Network Architecture. To build a lightweight architecture for practical applications, we choose the dehazing AOD-Net [14] as the backbone of our LE-Net and LD-Net. We modify the AOD-Net to make it suitable for our Retinex decomposition task. Specifically, for LD-Net, we set the number of output channel to 1 and use sigmoid as the activation function. Subsequently, a Gaussian blurring layer is added to satisfy the priori of the pixel-wise smoothness of the illumination. For LD-Net, the output channel dimension is set to 3 and Rectified Linear Unit (ReLU) is used as the activation

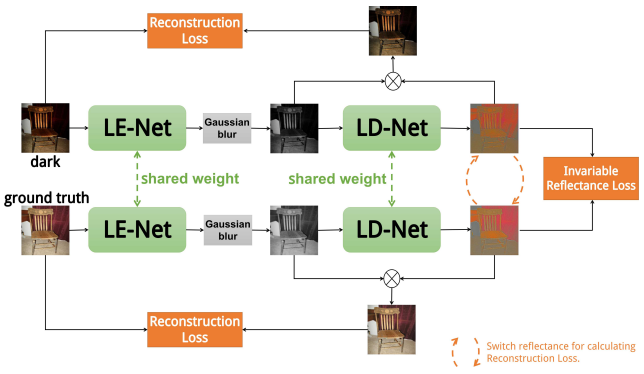


Figure 2: Retinex Decomposition Net.

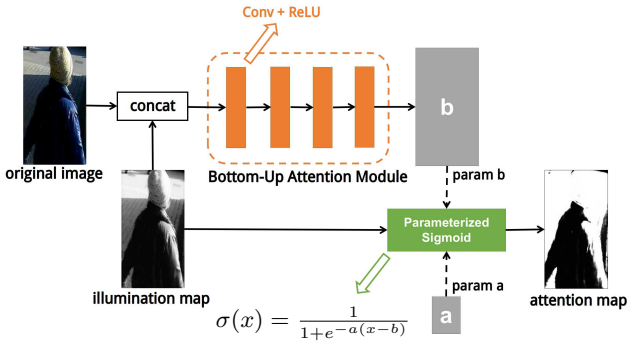


Figure 3: Bottom-Up Attention Module.

function. It is worth mentioning that, unlike Retinex-Net [30] that simultaneously estimates both illumination and reflectance, our method performs Retinex decomposition in a two-stage manner. This makes our deep model more flexible and easy to be optimized.

Loss Function. Based on the Lambertian scene assumption, each image pixel S captured by the camera can be modeled by

$$S(x, y) = L(x, y) \times R(x, y), \tag{1}$$

where L and R represent illumination and reflectance, respectively. For person re-id, the reflectance map describes the intrinsic property that is related to person identity and should be preserved or restored. The illumination map reflects the light conditions of environment, which is usually considered as an interference factor affecting the performance of re-identification and thus should be removed.

To effectively train the network, a low-light image is fed into the network to predict illumination and reflectance, and its corresponding ground truth is used to calculate the reconstruction loss L_{recon}:

$$L_{recon} = \sum_{i=low,gt} \sum_{j=low,gt} \lambda_{ij} ||R_i \circ L_j - S_j||_1, \tag{2}$$

where $\circ$ denotes the element-wise multiplication, and the invariable reflectance loss L_{ir} is defined as:

$$L_{ir} = ||R_{low} - R_{gt}||_1. \tag{3}$$

The total loss L for Retinex decomposition net is:

$$L = L_{recon} + \lambda_{ir} L_{ir}, \tag{4}$$

where λ_{ir} is used to control the consistency of reflectance, as described in [30].

3.2 Bottom-Up Attention for Enhancement

Existing Retinex-based methods enhance an image by adjusting the estimated illumination. This operation makes the enhanced image more natural and has a better subjective visual effect. For person re-id task, however, the appearance properties of the person itself are more important. Therefore, we directly choose the reflectance map, which contains boosted image characteristics, as the enhanced result.

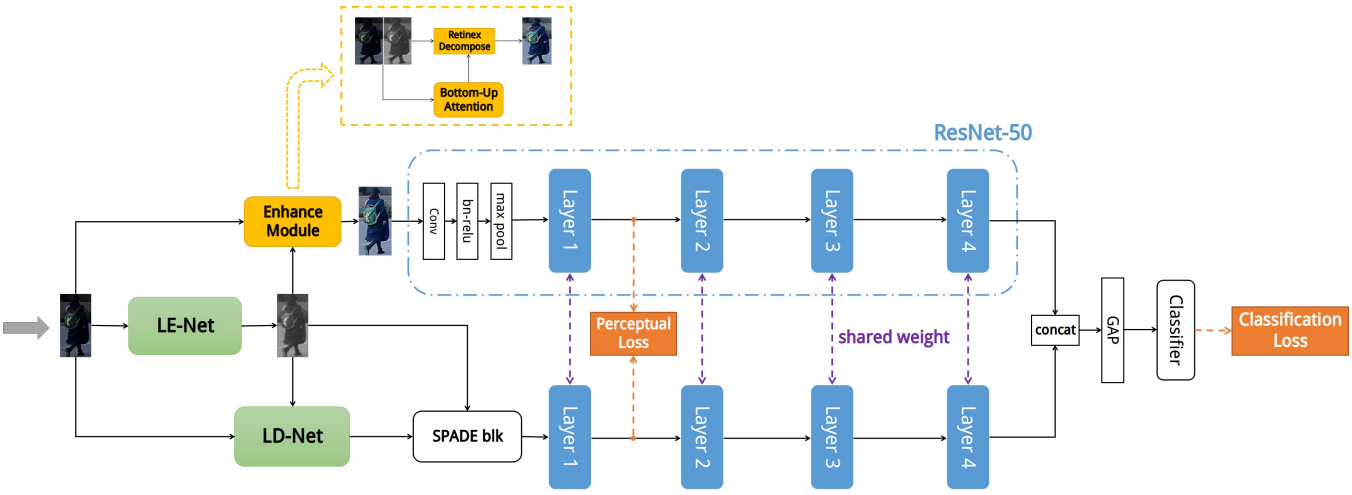


Figure 4: Framework of our illumination-invariant person re-identification.

However, the over-enhancement issue usually happens in dark areas where pixel values equal or close to 0. According to Equation 1, directly calculating R by using L as the denominator leads to severe distortion and amplified noises.

In fact, since the information is almost lost, there is no need to enhance these dark areas. Based on this observation, we introduce a bottom-up attention mechanism to deal with the over-enhancement problem by not enhancing dark areas. We design a parameterized sigmoid function, i.e., Equation 5, to generate the attention map and use it to determine which areas need to be enhanced.

$$\sigma(x) = \frac{1}{1 + e^{-a(x-b)}}. \tag{5}$$

Since a fixed scalar parameter treats all pixels equally, which reduces the model flexibility for different lighting conditions. Therefore, we use several convolution layers to predict the parameter b in Equation 5. Consequently, the attention map, which represents the intensity of the illumination enhancement, is obtained by the bottom-up attention module (Figure 3). Then the illumination map is adjusted by:

$$L_{adj}(x, y) = A(x, y) \times L(x, y) - A(x, y) + 1, \tag{6}$$

where L_{adj} denotes the adjusted illumination map, A denotes the attention map. Finally, we calculate the result E by:

$$E(x, y) = S(x, y)/L_{adj}(x, y). \tag{7}$$

3.3 Learning Illumination-Invariant Feature

Again, we emphasize that the motivation of this work is to improve the performance of person re-id rather than generate visually pleasing images. Although Retinex-based methods can obtain the illumination-independent reflectance map, only pixel-level image processing is performed. As mentioned in [9], a direct combination of low-level vision and re-id models may suffer from suboptimal compatibility. This is because generic-purpose low-level image process methods are designed to improve visual fidelity rather than the re-id performance. Therefore, there is still a gap between the low-level and high-level vision tasks. To further connect the Retinex decomposition and person re-id, we propose a joint training framework for learning illumination-invariant features, as illustrated in Figure 4.

Network Architecture. Our illumination invariant feature learning framework can be divided into two parts: pixel-level branch and feature-level branch. For the pixel-level branch, we perform image enhancement by removing the illumination map as described in the previous two subsections. For the feature-level branch, we first remove the last convolutional layer of the LD-Net, which directly output 32-channel feature maps rather than 3-channel image. Then, these illumination-independent features are down-sampled and fed into a SPADE block [23] for further feature transformation. The obtained features from two branches are then sent into the weight-shared resnet50, respectively. At the last layer of resnet50, the feature vectors are concatenated to perform classification.

Loss Function. The total loss function L consists of the re-id loss L_{id} and the perceptual loss L_{per},

$$L = L_{id} + L_{per}. \tag{8}$$

For re-id loss, we use the cross-entropy loss for multi-identity classification. For perceptual loss [10], different from the general practice of using pre-trained deep classification networks, i.e., VGG-16, we use the re-id backbone for obtaining both feature extraction and perceptual loss. In this way, both low-level and high-level constraints can be simultaneously taken into consideration during training. This can help the network learn illumination-invariant and discriminative features for re-id. Note that there is no extra computational and memory cost to use this training strategy.

Specifically, the feature maps from the two branches are directly used to calculate the perceptual loss,

$$L^j_{per}(\hat{x}, x) = \frac{1}{C_j H_j W_j} ||\hat{x}_j - x_j||_2^2, \quad (9)$$

where $\hat{x}_j$ and x_j denotes the feature maps, which are size of $C_j \times H_j \times W_j$, from the jth layer of the pixel-level branch and feature-level branch, respectively.

4 EXPERIMENTS

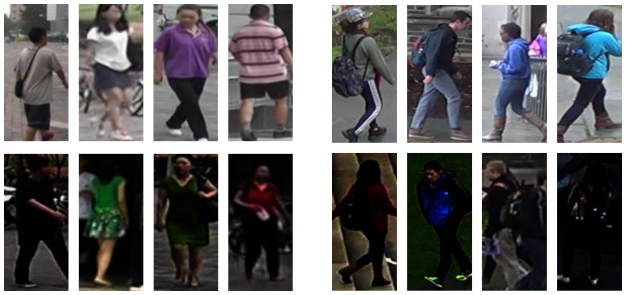
4.1 Datasets and Evaluation Measures

Datasets. Our evaluation is performed on two real-world person datasets, i.e., MSMT17 [32] and 3DPeS [3], which have severe lighting changes. We also adopted two simulated low-light person datasets which are based on Market-1501 and DukeMTMC-reID, respectively. Some examples are shown in Figure 5.

The **MSMT17 dataset** [32] is collected by the surveillance camera on the campus, including 12 outdoor cameras and 3 indoor cameras. To cover different time period, four days with different weather conditions in one month as well as three hours in the morning, noon and afternoon were selected to collect the raw videos. The whole dataset consists of 32,621 bounding boxes of 1,041 distinct persons for training and 93,820 bounding boxes of 3,060 persons for testing.

The **3DPeS dataset** [3] includes 1,011 images of 192 individuals captured from 8 outdoor cameras on the campus and each person has 2 to 26 images. The illumination variations could be very strong since people were recorded in bright and shadowy areas over the course of several days.

The **low-light Market-1501** (low-light Market) dataset is based on a publicly available person re-id benchmark, Market-1501. Market dataset has totally 32,668 images of 1,501 people captured by 5 high-resolution and one low-resolution camera. During dataset collection, a total of six cameras are placed in front of a campus supermarket in the daytime, so the illumination changes are not significant. To simulate the low-light conditions in the surveillance scene, we randomly select a process method for each image in the test set, including gamma correction with gamma value randomly picked from $\{2, 3, 4\}$ or no processing. Some examples are shown in Figure 5(a).



(a) Low-light Market-1501 (b) Low-light DukeMTMC-ReID

Figure 5: Examples of synthetic low-light person datasets.

(a) Produced by baseline model (b) Produced by our approach

Figure 6: Illustration for the illumination-invariant features. Top: normal images, bottom: low-light images.

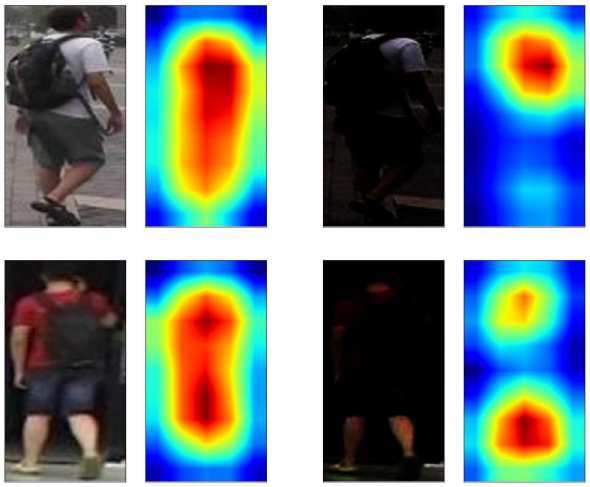
The **low-light DukeMTMC-reID** (low-light Duke) is built from DukeMTMC-ReID dataset, all images of which are extracted from the DukeMTMC [26] tracking dataset collected by 8 high-resolution cameras. Specifically, there are 6,522 images of 702 persons in the training set and 18,750 images associated to 702 identities in the test set. We randomly selected the images in the test set for low-light processing, as for Market-1501.

Evaluation Measures. We evaluate the performance of different person re-id methods using Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP). The standard single-shot setting is performed in our experiments.

4.2 Training the networks

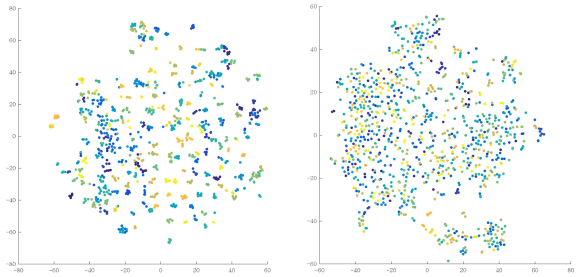
Our framework integrates multiple vision tasks: Retinex decomposition, image enhancement and person re-id. We train our network in a two-stage fashion.

Stage 1: Retinex decomposition. To train our Retinex decomposition network shown in Figure 2, we synthesize a large set of low-light images based on the PASCAL VOC images dataset. The experimental settings are followed by [22]. Specifically, we use Adam optimiser [12] with learning rate 10^{-4} and set mini-batch size to 32, λ_{ij} to 0.25 and λ_{ir} to 10^{-3}. The training is finished after 100 epoch. The input image is resized to 256×256. All parameters are randomly initialized by [8].



(a) normal-light condition (b) low-light condition

Figure 7: Illustration for the activation maps under different illumination conditions.



(a) normal-light condition (b) low-light condition

Figure 8: Visualization of feature embedding under different illumination conditions.

Stage 2: Joint training for person re-id. In this stage, all the parameters in the framework are fine-tuned by the re-id loss and perceptual loss (Equation 8). Specifically, the Retinex decomposition network is pre-trained in stage 1 and the re-id backbone is initialized with ImageNet pre-trained weights. We use SGD with Nesterov momentum and set the initial learning rate to 10^{-2}, momentum to 0.9 and mini-batch size to 32 for re-id backbone. The Retinex decomposition network is trained by Adam optimiser with 10^{-5}. The rest of framework are trained by SGD with learning rate 0.1. The input person image is resized to 288×144 and random horizontal flipping is performed.

4.3 Experimental results

In this section, we start with analyzing the effects of weak illumination on person re-identification performance. We see that a significant negative impact on re-id performance is caused by weak illumination. Then we evaluate the proposed bottom-up attention module. Both re-id performance and subjective visual effects are improved after adding this module. Last, re-id performance of our joint framework is compared with other light-enhancement and re-id methods. We also prove that our framework can be applied to current state-of-the-art re-id methods to further improve performance.

Table 1: Effect of weak illumination on Market-1501 dataset.

Re-ID methods	Illumination condition	Rank-1	mAP
ResNet50-IDE	Normal	82.5	62.9
ResNet50-IDE	Low-light	34.0	9.6
PCB	Normal	92.1	77.1
PCB	Low-light	48.7	16.8

Table 2: Effect of weak illumination on DukeMTMC-ReID dataset.

Re-ID methods	Illumination condition	Rank-1	mAP
ResNet50-IDE	Normal	73.6	54.0
ResNet50-IDE	Low-light	32.2	11.1
PCB	Normal	85.1	70.3
PCB	Low-light	48.6	21.1

Table 3: Performance of proposed bottom-up attention module.

Dataset	Model	Rank-1	mAP
MSMT17	ResNet50	57.3	29.7
	ResNet50-w-Enhance	53.0	27.4
	ResNet50-w-Enhance-w-BUatt	60.1	32.7
3DPeS	ResNet50	59.5	50.6
	ResNet50-w-Enhance	57.7	50.0
	ResNet50-w-Enhance-w-BUatt	60.4	51.6

Illumination-invariant features. To prove that our approach is able to learn illumination-invariant features, we compared the features of the last layer. As shown in Figure 6(a), baseline model produced different feature maps by inputting different illuminated images. The variation of feature maps results in the reduction of subsequent re-id performance. On the contrary, our model is able to produce similar feature maps under different illumination, as illustrated in Figure 6(b). This indicates that the our approach indeed learned the illumination-invariant features.

Effect of weak illumination. We use gamma correction to simulate the low-light conditions and synthesize low-light person datasets. To explore the effect of weak illumination on person re-id task, we evaluate the re-id performance by comparing the baseline results between the low-light datasets and their corresponding normal versions. We also analyze the

Table 4: Performance of the joint framework compared to other light enhancement + re-id schemes.

Dataset	Re-ID Method	Enhance Method	top-1	top-5	top-10	top-20	mAP
Low-light Market	ResNet50-IDE	-	34.0	48.0	53.6	59.1	9.6
	ResNet50-IDE	MSRCP [25]	38.1	56.8	**63.7**	**75.0**	**13.3**
	ResNet50-IDE	LIME [7]	29.6	45.6	52.8	58.8	8.1
	Our framework		**42.1**	**57.3**	**63.7**	69.8	12.9
Low-light Duke	ResNet50-IDE	-	32.2	46.3	51.5	57.9	11.1
	ResNet50-IDE	MSRCP	31.8	48.4	54.7	59.8	11.4
	ResNet50-IDE	LIME	34.0	49.7	56.1	61.9	12.9
	Our framework		**38.8**	**53.1**	**58.6**	**63.0**	**14.5**
MSMT17	ResNet50-IDE	-	57.3	72.7	78.6	**84.3**	**29.7**
	ResNet50-IDE	MSRCP	57.5	73.0	78.6	83.6	29.3
	ResNet50-IDE	LIME	55.9	71.6	77.5	82.4	28.1
	Our framework		**59.1**	**74.0**	**79.2**	83.8	28.6

Table 5: Performance of the joint framework in combination with the state-of-the-art method.

Dataset	Model	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Low-light Market	PCB	48.7	63.5	68.8	74.0	16.8
	Our framework + PCB	**52.6**	**68.5**	**73.7**	**78.9**	**18.8**
Low-light Duke	PCB	48.6	60.9	65.9	70.5	21.1
	Our framework + PCB	**49.5**	**62.1**	**66.6**	**70.9**	**22.1**

CNN features, including activation maps and feature vectors used for inference.

Using ResNet50-IDE [37] as the baseline model, the re-id performance of low-light and normal-light person datasets are shown in Table 1 and Table 2. We can see that rank-1 accuracy and mAP decrease by 48.5% and 53.3% on the Market dataset, and 41.4% and 42.9% on the Duke dataset, respectively. We then test the state-of-the-art PCB [28] method and the re-id performance still drops by a large margin.

Next, we analyze the features extracted under different lighting conditions by visualizing activation maps and feature vectors. Specifically, we visualize the activation maps before the last pooling layer of our baseline model. We found that under normal lighting conditions, the attention is mostly concentrated on the person area, as shown in Figure 7(a), while under low-light conditions, the attention maps become scattered and focus on those bright areas, as shown in Figure 7(b).

For feature embedding, we select 1,000 samples of Market-1501 dataset and extract 2048-dim feature vectors from them using the baseline model. To visualize the distribution of these feature vectors, we use the t-SNE method to reduce the dimension from 2048-dim to 2-dim, and the final result is shown in Figure 8. In normal-light situations, feature vectors belonging to the same identity tend to group closely, and those with different identities are separated. However, in low-light situations, the distribution of feature vectors becomes significantly scattered, which leads to overlapping decision regions, causing a negative impact on the performance of re-id task.

Evaluation of bottom-up attention. Bottom-up attention is proposed for suppressing the over-enhancement as described in Section 3.2. For those extremely dark areas, directly performing illumination decomposition (also referred to as enhancement in this paper) results in severe color distortion and noises magnification, as illustrated in Figure 10(b). Directly using such enhanced results inevitably lead to poor re-id performance. Table 3 shows the performance reduction on MSMT17 and 3DPeS dataset caused by over-enhancement.

After introducing our proposed bottom-up attention module, the illustration of enhanced images is shown in Figure 10(c). The distortion has been effectively eliminated shown in Figure 9, and the original properties of the dark region is well preserved. As can be seen in Table 3, using our bottom-up attention module can further improve the re-id performance.

Performance of the joint framework. To evaluate our joint framework for learning illumination-invariant feature, we have selected two recent image enhancement methods: LIME [7] and MSRCP [25], as competitors. We use these two methods to preprocess the person images and then feed them into ResNet50-IDE. This operation is performed on both training and testing stages.

Table 4 shows that our framework outperforms other light enhancement + re-id schemes and achieves consistently superior performances on all datasets. Specifically, the increment of rank-1 accuracy can reach +8.1%, +6.6% and +1.8% on

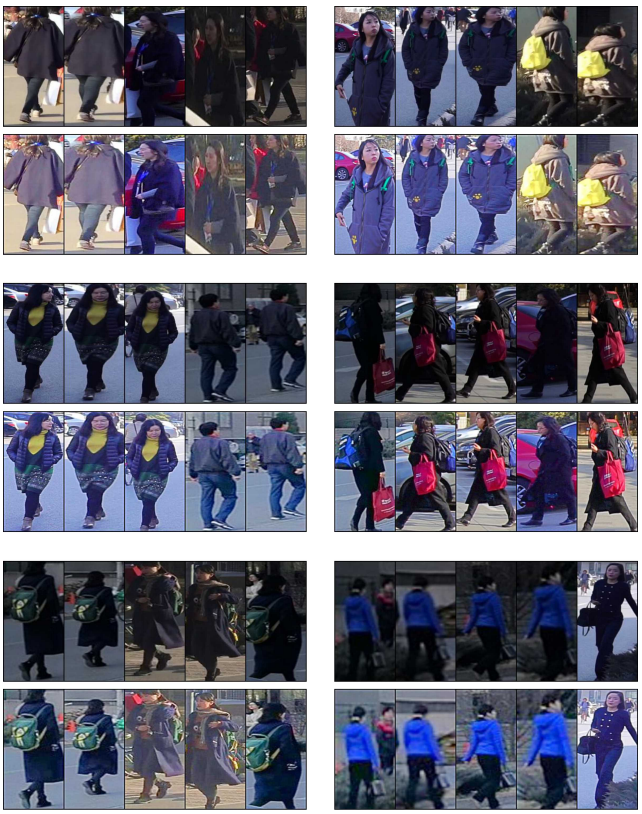


Figure 9: Illustration for effectiveness of our enhance module. Top: original images, bottom: enhanced images.

(a) Original images

(b) Enhanced images

(c) Enhanced images with Bottom-Up Attention

Figure 10: Illustration for effectiveness of our bottom-up attention module.

low-light Market, low-light Duke and MSMT17 datasets, respectively. The increment of mAP reached +3.4% on low-light Duke dataset.

Further evaluation. In the previous experiments, we have verified that the current state-of-the-art re-id methods can not avoid the large performance degradation caused by weak illumination. This is because these methods focus on solving the high-level semantic feature mismatch problem, such as pose changes and occlusion, without considering illumination changes. On the contrary, our framework is specifically designed to learn the illumination-invariant feature, which is able to benefit other state-of-the-art methods. To verify the idea, we add our proposed model with PCB [28]. As shown in Table 5, the framework brings further performance improvement to the PCB model, as expected. For low-light Market dataset, rank-1 accuracy and mAP are increased by +3.9% and +2.0%, respectively. For low-light Duke dataset, rank-1 accuracy and mAP are increased by +0.9% and +1.0%. As a result, the proposed framework is effective to tackle the problems of weak illumination and light changes and can be combined with existing state-of-the-art methods to further improve performance.

5 CONCLUSION

Low-light or mixed lighting conditions are common in real-world surveillance scenarios, while most existing re-id methods lack robustness to illumination changes. In this work, we first prove that weak illumination conditions have a negative impact on the person re-id task. Inspired by the Retinex theory, we then perform illumination decomposition on input person images to obtain reflectance maps as enhanced results. A bottom-up attention module is further introduced to suppress over-enhancement. To connect both low-level and high-level vision tasks, we propose a unified framework to learn the illumination invariant feature for person re-id. Extended experiments demonstrate the superiority of our framework by comparing to other light enhancement + re-id schemes. We also show that our framework can be directly combined with existing re-id methods and improves their robustness to weak illuminated images.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants 61622211 and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3908–3916.

[2] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. 2018. Domain Adaptation through Synthesis for Unsupervised Person Re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 189–205.

[3] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 2011. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM, 59–64.

[4] Di Chen, Zheng-Jun Zha, Jiawei Liu, Hongtao Xie, and Yongdong Zhang. 2018. Temporal-Contextual Attention Network for Video-Based Person Re-identification. In *Pacific Rim Conference on Multimedia*. Springer, 146–157.

[5] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision*. 2590–2600.

[6] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. 2018. FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification. In *Advances in Neural Information Processing Systems*. 1230–1241.

[7] Xiaojie Guo, Yu Li, and Haibin Ling. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Processing* 26, 2 (2017), 982–993.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[9] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. 2018. Deep low-resolution person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[11] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1062–1071.

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

[13] Edwin H Land. 1977. The retinex theory of color vision. *Scientific American* 237, 6 (1977), 108–129.

[14] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. 2017. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision*. 4770–4778.

[15] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.

[16] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive Transfer Network for Cross-Domain Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7202–7211.

[17] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 8.

[18] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the ACM Conference on Multimedia Conference*. ACM, 192–196.

[19] Jiawei Liu, Zheng-Jun Zha, Hongtao Xie, Zhiwei Xiong, and Yongdong Zhang. 2018. CA 3 Net: Contextual-Attentional Attribute-Appearance Network for Person Re-Identification. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 737–745.

[20] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*. 350–359.

[21] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition* 61 (2017), 650–662.

[22] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. [n. d.]. MBLLEN: Low-light Image/Video Enhancement Using CNNs. In *British Machine Vision Conference*.

[23] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[24] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. 2016. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1306–1315.

[25] Ana Belén Petro, Catalina Sbert, and Jean-Michel Morel. 2014. Multiscale Retinex. *Image Processing On Line* (2014), 71–88. https://doi.org/10.5201/ipol.2014.107

[26] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*. 17–35.

[27] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma. 2017. Msr-net: Low-light image enhancement using deep convolutional network. *arXiv preprint arXiv:1711.02488* (2017).

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496.

[29] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. 2018. Cascaded SR-GAN for Scale-Adaptive Low Resolution Person Re-identification.. In *IJCAI*. 3891–3897.

[30] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *BMVC*.

[31] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* (2018).

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

[33] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. 2017. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184* (2017).

[34] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1077–1085.

[35] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3219–3228.

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[37] Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person Re-identification: Past, Present and Future. *arXiv preprint arXiv:1610.02984* (2016).

[38] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*. 3754–3762.

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.