Illumination-Adaptive Person Re-identification

Zelong Zeng, Zhixiang Wang, *Student Member, IEEE,* Zheng Wang, *Member, IEEE,*
Yinqiang Zheng, Yung-Yu Chuang, *Member, IEEE,* Shin'ichi Satoh, *Member, IEEE*

***Abstract*—Most person re-identification (ReID) approaches assume that person images are captured under relatively similar illumination conditions. In reality, long-term person retrieval is common, and person images are often captured under different illumination conditions at different times across a day. In this situation, the performances of existing ReID models often degrade dramatically. This paper addresses the ReID problem with illumination variations and names it as *Illumination-Adaptive Person Re-identification (IA-ReID)*. We propose an Illumination-Identity Disentanglement (IID) network to dispel different scales of illuminations away while preserving individuals' identity information. To demonstrate the illumination issue and to evaluate our model, we construct two large-scale simulated datasets with a wide range of illumination variations. Experimental results on the simulated datasets and real-world images demonstrate the effectiveness of the proposed framework.**

***Index Terms*—Person Re-identification, Illumination-Adaptive, Feature Disentanglement**

I. Introduction

Person re-identification (ReID) is a cross-camera retrieval task. Given a query person-of-interest, it aims to retrieve the same person from a database of images collected from multiple cameras [1], [2], [3], [4], [5], [6], [7]. The key challenge of ReID lies in the person's appearance variations among different cameras. Most previous methods attempt to find a feature representation that is stable to the appearance variations. They have well investigated how to deal with variations in occlusions [8], resolutions [9], [10], poses [11], *etc*. However, the influence of ever-changing illumination conditions has been largely ignored. Most popular ReID datasets, such as Market1501 [12] and DukeMTMC-reID [13], have relatively uniform illumination conditions, as their images were captured under similar illumination at the same period of time.

In practice, long-term person retrieval is often required in video surveillance networks and criminal investigation applications. As Figure 1(a) shows, the images of a suspect could be taken under very different illumination conditions at different times across a day. He may appear in camera A with dim light at 6:00 a.m., then in camera B with normal light at 9:00 a.m., and finally in camera C with glare light at 12:00 p.m. Existing researches do not investigate this illumination-adaptive issue. We name the ReID task under different illumination conditions as *Illumination-Adaptive Person Re-identification (IA-ReID)*. In this task, given a probe image under one illumination, the goal is to match gallery images with several different illuminations. The images can be with normal illumination as existing ReID datasets (Market1501, DukeMTMC-reID), can be very bright if captured under dazzling sunshine, and also can be very dark if captured in the sunset or even during the night.

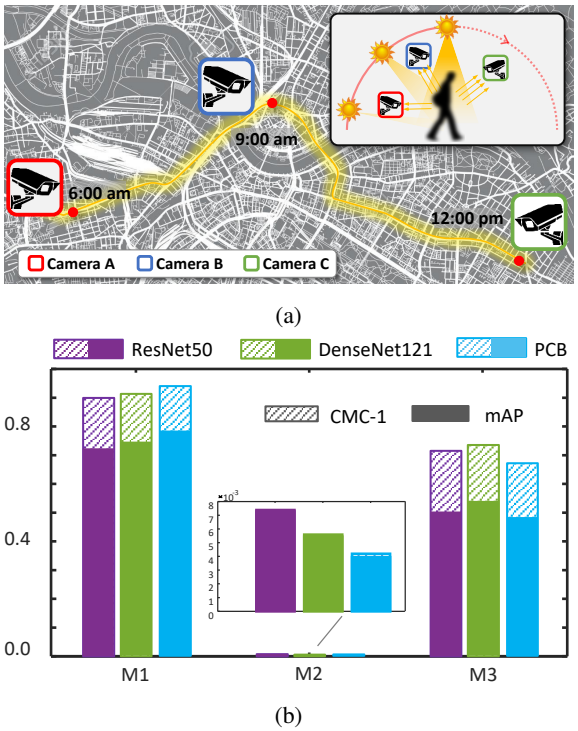


Figure 1: (a) A toy example to illustrate the real application scenario of ReID where images captured at different times could have quite different illumination conditions. (b) The results of preliminary experiments to show the impact of illumination-adaptive. The experiments were conducted on three kinds of networks, including ResNet50 [He *et al.*, 2016], DenseNet121 [Huang *et al.*, 2017], and PCB [Sun *et al.*, 2018]. "M1", "M2" and "M3" stand for the CMC-1 and mAP results for three different training and testing pairs. "M1" is obtained by training and testing on the Market-1501 dataset. "M2" is obtained by training on the Market-1501 and testing on the Market-1501++ dataset. "M3" is obtained by training and testing on the Market-1501++ dataset.

The illumination greatly affects the performance of re-identification. As we know, the re-identification performance heavily relies on the characteristics of the datasets. The model trained on one dataset often can not perform well on the

This work was supported in part by JST CREST under Grant JPMJCR1686, in part by Grant-in-Aid for JSPS Fellows under Grant 18F18378, in part by Honda R&D and in part by Microsoft Research Asia. Zelong Zeng and Zhixiang Wang are the co-first authors of this article. *(Corresponding author: Zheng Wang)*

Z. Zeng is with the Graduate School of Information Science and Technology, The University of Tokyo. (e-mail: zzlbz@nii.ac.jp). Z. Wang, Y. Zheng and S. Satoh are with the Digital Content and Media Sciences Research Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan. (e-mail: wangz@nii.ac.jp; yqzheng@nii.ac.jp; satoh@nii.ac.jp). Z. Wang and Y. Chuang are with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan. (e-mail: wangzx1994@gmail.com; cyy@csie.ntu.edu.tw).

other. Traditional models, although efficient and effective to re-identify gallery images with the same illumination, may suffer from a significant performance drop when the illumination conditions of gallery images vary greatly. We have conducted preliminary experiments for demonstrating the performance degradation using the Market-1501 dataset. We simulated different illumination conditions for images in the dataset. The resultant dataset with varying illumination conditions is named the Market-1501++ dataset. Three training-testing configurations were performed: M1 (both train and test on Market-1501), M2 (train on Market-1501 but test on Market-1501++), and M3 (both train and test on Market-1501++). As Figure 1(b) illustrates, 1) the learned models are not stable across datasets with different illuminations ("M2") and 2) even trained with images under different illumination conditions, the model cannot achieve satisfying performance ("M3"). That is to say, general ReID models lose their effectiveness in the situation with illumination variations.

As far as we know, some researches investigated the issue of different illuminations in ReID [14], [15]. However, they only consider a situation of two scales of illuminations. They assumed that the probe and the gallery images are respectively captured from two cameras, each with a corresponding illumination condition. In such a controlled setting, they proposed to learn the relationship between two scales of illuminations by brightness transfer functions [14] or the feature projection matrix [15]. IA-ReID is a more practical problem with multiple illuminations. Obviously, constructing the relationships among different scales of illuminations is not a practical solution for this problem. If there are ten different scales in the dataset, the method needs to construct ten different relationships, and it cannot be guaranteed the ten relationships work perfectly.

Removing the effect of illumination is another intuitive idea. One solution is to do image enhancement [16] for the low-illumination images and image reconstruction [17] for the images with high exposure. However, this kind of methods either cannot handle extreme illuminations [18], or are particularly designed for visualization and rely heavily on the data condition and training samples [16], [17].

Another solution is to disentangle the illumination information from the person feature. This idea is learned from the existing face recognition methods [19], where certain face attributes are separated from the feature vector. In this paper, we follow this thread of ideas and propose an Illumination-Identity Disentanglement (IID) network. Inspired by previous researches dealing with the image resolution issue [10], we construct two simulated illumination-adaptive datasets based on Market1501 [12] and DukeMTMC-reID [13]. Then, these two datasets are utilized to evaluate the effectiveness of the IID network. Our contributions can be summarized as follows:

- **A new and practical problem.** We raise a new and practical task, *i.e.*, IA-ReID. The task is practical for long-term person re-identification applications. We construct two simulated datasets with different illuminations to put forward this task. Most general ReID models are proved ineffective when evaluated on these two datasets.
- **A novel method.** We propose a novel Illumination-Identity Disentanglement (IID) network, which dispels the illumination information away from a person's appearance. The method achieves great performance improvement on our two datasets. We also evaluate our model on some real images, and it is capable of alleviating the effect of illumination discrepancy.
- **Simplicity and Effectiveness.** We construct the IID network based on a simple backbone. The network is easy to follow. Extensive experiments prove that IID is robust in long-term person re-identification applications. In this way, we set a benchmark for the new task.

II. Related work

A. Short-term ReID

In the short-term condition, existing researches paid attention to the challenges from resolution variations [10], pose changes [11] and occlusion [8]. Many methods were proposed to learn robust representations to overcome those challenges. They have achieved very high performances on the public datasets. Liao *et al.* [20] analyzed the horizontal occurrence of local features and maximized the occurrence to make a stable representation against viewpoint changes. Su *et al.* [21] leveraged the human part cues to alleviate the pose variations and learn robust feature representations from both the global image and different local parts. Zheng *et al.* [8] fused a local patch-level and a global part-based matching model to address the occlusion problem. However, when the re-identification task goes to the long-term situation, the illumination variations come to be the key issue.

We have claimed that public datasets, such as Market1501 [12] and DukeMTMC-reID [13], have relatively uniform illumination conditions. As evidence, for both Market1501 and DukeMTMC-reID, the file name of each image includes the frame index, which indicates the relative captured time of the corresponding image. The range of frame indexes of all images spans over 90,000 frames for Market1501, i.e., those images were captured in a time span of around one hour (25fps$\times$3600s). For DukeMTMC-reID, images were captured in around two hours. It supports our claim that existing ReID datasets were captured in a short period of time, and the images have relatively uniform illumination conditions.

B. Illumination Problem in ReID

Some researches start to investigate the illumination issue in ReID. Previous methods [22], [14], [15] consider the situation that the probe and the gallery images respectively captured from two cameras with two different illuminations. Kviatkovsky *et al.* [22] proposed an invariant feature exploiting a structure of color distributions, using different parts of the person. Bhuiyan *et al.* [14] learned robust brightness transfer functions to release the illumination change from one camera to the other. Wang *et al.* [15] designed a feature projection matrix to project image features of one camera to the feature space of another camera. Ma *et al.* [23] focused on the low illumination problem. They transformed all the images to a uniform low illumination and proposed metric learning methods to address the low illumination. There are also some researches related to more extreme illumination

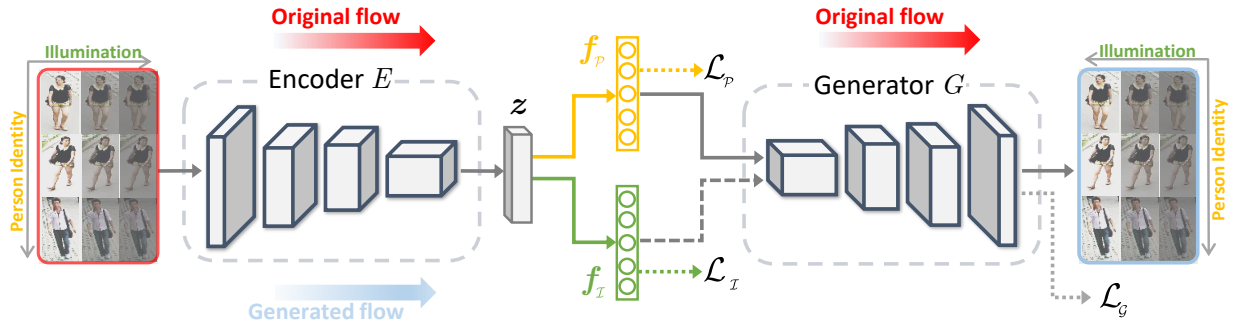


Figure 2: **The architecture of our proposed Illumination-Identity Disentanglement (IID) network.** It consists of an encoder E, a generator G, and two feature embedding layers $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{I}}$. The network is optimized through three loss functions. The embedded person (identity-relevant) feature $\boldsymbol{f}_{\mathcal{P}}$ is trained with the ReID loss $\mathcal{L}_{\mathcal{P}}$; the embedded identity-irrelevant feature $\boldsymbol{f}_{\mathcal{I}}$ is trained with the illumination loss $\mathcal{L}_{\mathcal{I}}$ and the generation loss $\mathcal{L}_{\mathcal{G}}$ is used to guide the optimization of the generator. The training process has two work flows. 1) The original flow is defined as $\boldsymbol{x} \xrightarrow{E} \boldsymbol{z} \xrightarrow{\mathcal{H}} (\boldsymbol{f}_{\mathcal{P}}, \boldsymbol{f}_{\mathcal{I}}) \xrightarrow{G} \hat{\boldsymbol{x}}$. The encoder E firstly encodes an image into a latent vector $\boldsymbol{z}$, and then disentangles into two components $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$. The generator G reconstructs the image $\hat{\boldsymbol{x}}$ fed with the combination of $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$. 2) The generated flow is defined as $\hat{\boldsymbol{x}} \xrightarrow{E} \hat{\boldsymbol{z}} \xrightarrow{\mathcal{H}} (\hat{\boldsymbol{f}}_{\mathcal{P}}, \hat{\boldsymbol{f}}_{\mathcal{I}})$. To enhance the identity-preserving of the reconstructed image $\hat{\boldsymbol{x}}$, we feed it into the network again. Similarly, we extract the disentangled representations $\hat{\boldsymbol{f}}_{\mathcal{P}}$ and $\hat{\boldsymbol{f}}_{\mathcal{I}}$.

conditions. They considered that when the imaging condition goes to late at night, the low-illumination image will transform into the infrared image. To this end, some methods [24], [25], [26], [27], [28], [29], [30] paid their attention to the infrared-visible re-identification, where one part of the images are from the visible camera and the other part is from the infrared camera. They designed networks to bridge the gap between infrared and visible images. In this paper, multiple illuminations are taken into consideration, including not only the low illumination images but also the high illumination ones. Our illumination-adaptive setting is more practical than previous methods for long-term ReID.

C. Disentangled Feature Learning

Some previous works tried to disentangle the representations in different kinds of recognition tasks [31], [32], [33], [34], for example pose-invariant recognition [19] and identity-preserving image editing [35]. They exploited attribute supervision and encoded each attribute as a separate element in the feature vector. Liu *et al.* [36] proposed to learn disentangled but complementary face features with face identification. Disentangled feature learning was investigated in image-to-image translation tasks as well. Lee *et al.* [37] exploited disentangled representation for producing diverse outputs, in particular, a domain-invariant content space capturing shared information across domains and a domain-specific attribute space was proposed. Gonzalez *et al.* [38] introduced the concept of cross-domain disentanglement and separated the internal representation into one shared part and two exclusive parts. There also exist some methods using disentangled representations to address the single-image deblurring tasks [39]. As far as we know, there is no method designed to disentangle features from individuals in ReID. Our method is the first to consider disentangling the illumination, one kind of identity-irrelevant information.

D. Encoder-Decoder Network

Our method exploits an encoder-decoder network. As we know, this framework is prevalent, *e.g.*, the Ladder Network [40] and U-Net [41]. Although our approach is similar to these two typical networks, there are key differences. 1) Different focuses: The Ladder Network focuses on the semi-supervised learning task, the U-Net focuses on the image segmentation task, while our method focuses on disentangling the person and illumination features. 2) Different structures: The Ladder Network consists of a supervised part and an unsupervised part, the U-net has no ID supervision and pays its attention to image generation, while our method introduces two supervised parts to extract two kinds of features. 3) Different outputs: The Ladder Network attempts to output a label for each input, the U-Net tries to obtain an image output, while our method attempts to extract features.

III. Our Method

A. Overview

Figure 2 depicts the overall architecture of our Illumination-Identity Disentanglement (IID) network. The encoder E takes ResNet-50 as the backbone network and encodes the input image $\boldsymbol{x}$ into a latent vector $\boldsymbol{z} = E(\boldsymbol{x})$, where $\boldsymbol{z} \in \mathbb{R}^{2048}$. Next, two independent fully-connected (FC) layers $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{I}}$, are employed to project the latent $\boldsymbol{z}$ into two different feature vectors, *i.e.*, the person feature $\boldsymbol{f}_{\mathcal{P}} = \mathcal{H}_{\mathcal{P}}(\boldsymbol{z})$ and the identity-irrelevant feature $\boldsymbol{f}_{\mathcal{I}} = \mathcal{H}_{\mathcal{I}}(\boldsymbol{z})$. Note that $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$ are expected to be respectively stable to illumination variations and irrelevant to person identity. To enforce the disentangled information fully represent the input image, the generator G is used to reconstruct the image $\hat{\boldsymbol{x}} = G(\boldsymbol{f}_{\mathcal{P}}, \boldsymbol{f}_{\mathcal{I}})$ to approximate the input image $\boldsymbol{x}$ from the disentangled feature vectors, $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$. Here, $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$ are concatenated, and act as the input of the generator G. The reason why we maintain the disentangled illumination-relevant feature rather than simply getting rid of it is to reduce the potential loss of discriminative information in the identity-relevant feature. Specifically, we

use the generator G and feed it with the disentangled identity-relevant and identity-irrelevant features to reconstruct the input with minimum loss of information. If we dump the disentangled illumination-relevant feature without guaranteeing the reconstruction, some useful information could slip away with the removed illumination-relevant feature.

B. Disentangled Feature Learning

Identity-relevant feature learning. Given the encoded latent vector $\boldsymbol{z}$ from the encoder E, the *FC* layer $\mathcal{H}_{\mathcal{P}}$ projects it to the person feature $\boldsymbol{f}_{\mathcal{P}}$, where $\boldsymbol{f}_{\mathcal{P}} = \mathcal{H}_{\mathcal{P}}(\boldsymbol{z})$. Since the feature $\boldsymbol{f}_{\mathcal{P}}$ is required to capture information relevant to person identity, we use the ReID loss $\mathcal{L}_{\mathcal{P}}$ and person identity information to supervise the training process. The ReID loss $\mathcal{L}_{\mathcal{P}}$ combines the triplet loss $\mathcal{L}_{\mathcal{P}}^{T}$ and the softmax loss $\mathcal{L}_{\mathcal{P}}^{S}$ and it can be written as

$$\mathcal{L}_{\mathcal{P}} = \lambda_1 \mathcal{L}_{\mathcal{P}}^{T} + \lambda_2 \mathcal{L}_{\mathcal{P}}^{S}\,, \tag{1}$$

where the weights λ_1 and λ_2 are used for balancing these two losses. The training strategy is the same as the one in the general ReID framework; hence the extracted feature can be used for the identification task directly.

The triplet loss is used for similarity learning and it can be formulated as

$$\mathcal{L}_{\mathcal{P}}^{T} = \sum_{\boldsymbol{f}_{\mathcal{P}}^{a}, \boldsymbol{f}_{\mathcal{P}}^{p}, \boldsymbol{f}_{\mathcal{P}}^{n} \in \mathcal{B}} \left[\mathcal{D}(\boldsymbol{f}_{\mathcal{P}}^{a}, \boldsymbol{f}_{\mathcal{P}}^{p}) - \mathcal{D}(\boldsymbol{f}_{\mathcal{P}}^{a}, \boldsymbol{f}_{\mathcal{P}}^{n}) + \xi\right]_{+}\,, \tag{2}$$

where $\mathcal{B}$ represents a mini-batch consisting of extracted person features $\boldsymbol{f}_{\mathcal{P}}$. For an anchor feature vector $\boldsymbol{f}_{\mathcal{P}}^{a}$, the positive sample $\boldsymbol{f}_{\mathcal{P}}^{p}$ and the negative sample $\boldsymbol{f}_{\mathcal{P}}^{n}$ respectively denotes a feature vector having the same identity with $\boldsymbol{f}_{\mathcal{P}}^{a}$ and one with different identity from $\boldsymbol{f}_{\mathcal{P}}^{a}$. Note that $\boldsymbol{f}_{\mathcal{P}}^{a} \neq \boldsymbol{f}_{\mathcal{P}}^{p}$. ξ is a margin parameter; $\mathcal{D}(\cdot)$ calculates the Euclidean distance; and $[\boldsymbol{d}]_{+} = \max(d, 0)$ truncates negative numbers to zero while keeping positive numbers the same. Note that we exploit the most primitive triplet loss function. While during the training process, we select hard samples for each triplet. For each anchor, we randomly select 16 samples with the same ID and 64 samples with different IDs, and compute the feature distance between the anchor and each selected sample. Then, we select the farthest positive sample as the hard positive sample, and the nearest negative sample as the hard negative sample. They together construct a triplet.

The softmax loss is employed for identity information learning, which is written as

$$\mathcal{L}_{\mathcal{P}}^{S} = -\frac{1}{N}\sum_{j=1}^{N} \log \hat{\boldsymbol{y}}_{\mathcal{P}}^{j}\,, \tag{3}$$

where N is the number of images in the mini-batch $\mathcal{B}$ and $\hat{\boldsymbol{y}}_{\mathcal{P}}$ is the predicted probability of the input belonging to the ground-truth class with $\boldsymbol{y}_{\mathcal{P}} = \texttt{softmax}(\mathbf{W}_{\mathcal{P}}\boldsymbol{f}_{\mathcal{P}} + \mathbf{b}_{\mathcal{P}})$, where $\mathbf{W}_{\mathcal{P}}$ and $\mathbf{b}_{\mathcal{P}}$ are the trainable weight and bias of $\mathcal{H}_{\mathcal{P}}$ respectively.

Identity-irrelevant feature learning. Given the encoded latent vector $\boldsymbol{z}$ from the encoder E, the *FC* layer $\mathcal{H}_{\mathcal{I}}$ projects it to the identity-irrelevant feature $\boldsymbol{f}_{\mathcal{I}}$, written as $\boldsymbol{f}_{\mathcal{I}} = \mathcal{H}_{\mathcal{I}}(\boldsymbol{z})$. To make the feature irrelevant to person identity, we need to feed the network with images taken under different illuminations. Thanks to our simulated dataset, each image is automatically assigned an illumination label, indicating the scale of relative illumination change. Note that the detail information of the datasets is described in Section IV-A. The illumination label is somehow coarse since we assumed that images in original datasets are captured in a relatively uniform illumination condition. To eliminate the reliance on this assumption, we make two necessary modifications. 1) We adopt the classifier problem to do regression. 2) We use a soft label strategy instead of a hard label. The soft label strategy is used because we would like to leave some room for tolerating slight changes, possibly caused by camera styles and viewpoints, occurring on this relatively uniform illumination condition. For the same purpose, we also transform the classification problem into a regression problem.

The regression loss is written as

$$\mathcal{L}_{\mathcal{I}} = \frac{1}{N}\sum_{j=1}^{N} \left\| \hat{c}_{\mathcal{I}}^{j} - (\mathbf{W}_{\mathcal{I}}\boldsymbol{f}_{\mathcal{I}}^{j} + \mathbf{b}_{\mathcal{I}}) \right\|_2^2\,, \tag{4}$$

where $\mathbf{W}_{\mathcal{I}}$ and $\mathbf{b}_{\mathcal{I}}$ are the trainable weight and bias respectively. The soft label $\hat{c}_{\mathcal{I}}$ is the summation of the ground truth label $c_{\mathcal{I}}$ and Gaussian noise ϵ.

$$\hat{c}_{\mathcal{I}} = c_{\mathcal{I}} + \epsilon\,, \quad \text{with} \quad \epsilon \in \mathcal{N}(0, 1)\,. \tag{5}$$

Note that there is no difference between the datasets of compared state-of-the-art ReID models and our network. Images with multiple illuminations are just used to separate the illumination information from the person feature.

C. Identity-preserving Image Generation

The image generator G is employed to ensure that the disentangled information has minimum information loss. It is fed with the combination of $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$ and generates the reconstructed image $\hat{\boldsymbol{x}} = G(\boldsymbol{f}_{\mathcal{P}}, \boldsymbol{f}_{\mathcal{I}})$. We use the *MSE* loss as the supervision information, which is defined as

$$\mathcal{L}_{\mathcal{G}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2\,. \tag{6}$$

It is worth mentioning that, in addition to guiding the reconstruction of the input and supervising the training process, G can be used to synthesize images with same identity but varying illuminations by altering the illumination-relevant feature vector $\boldsymbol{f}_{\mathcal{I}}$. For the network configure of generator G, we use six decode modules. Each decode module consists of a ReLU layer, a 2D transposed convolution layer, a batch normalization layer and a dropout layer. We concatenate $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$ as the input of the generator, and the output of generator is the generated image with size $128 \times 64 \times 3$.

To enforce the reconstructed image $\hat{\boldsymbol{x}}$ identity-preserving, we feed it into the network again. As Figure 2 shows, we name this process as generated flow, in contrast to the original flow. In the generated flow, the loss $\mathcal{L}_{\mathcal{P}}$ and $\mathcal{L}_{\mathcal{I}}$ are taken into account in the total loss function. However, the features $\boldsymbol{f}_{\mathcal{P}}$ and $\boldsymbol{f}_{\mathcal{I}}$ are not needed to utilize again to generate new images.

D. Training Process

The training process has three phases: feature disentanglement training, generator training, and joint training. Their details will be described as follows.

Phase I: Feature disentanglement training. In this phase, parameters of the disentangled feature learning module are updated. As mentioned above, the disentangled feature learning module consists of the encoder E, two feature embedding layers $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{I}}$, and the weight and bias $\mathbf{W}_{\mathcal{P}/\mathcal{I}}$, $\mathbf{b}_{\mathcal{P}/\mathcal{I}}$ for loss functions defined in Equation (3) and Equation (4). We denote the parameters of E by θ_E. The parameters of the FC layer $\mathcal{H}_{\mathcal{P}}$, the weights $\mathbf{W}_{\mathcal{P}}$ and the biases $\mathbf{b}_{\mathcal{P}}$ are denoted by $\theta_{\mathcal{P}}$. Similarly, parameters of $\mathcal{H}_{\mathcal{I}}$, $\mathbf{W}_{\mathcal{I}}$ and $\mathbf{b}_{\mathcal{I}}$ are denoted by $\theta_{\mathcal{I}}$. The object function for this phase is

$$\arg\min_{\theta_E,\theta_{\mathcal{P}},\theta_{\mathcal{I}}} \mathcal{L}_{\mathcal{P}} + \lambda_3 \mathcal{L}_{\mathcal{I}} \,. \tag{7}$$

We set $\lambda_3 = 1$. Note that $\mathcal{L}_{\mathcal{P}}$ is defined in Equation (1) with the hyperparameters $\lambda_1 = \lambda_2 = 0.5$. The hyperparameter ξ in Equation (2) is set to 0.3.

The encoder E is initialized with ImageNet pre-trained weights [42]. Other parts are initialized with He's method [43]. The optimizer utilizes SGD with momentum and weight decay set to 0.9 and 5×10^{-4}. The learning rate for E is set to 0.05 initially and divided by 10 after every 40 epochs. The learning rate for other parts is $\frac{1}{10}$ of the learning rate for E. Algorithm 1 depicts the detailed training procedure of this phase.

Algorithm 1: Phase I of the training procedure.

Input: Training data $\{\boldsymbol{x}_i\}$ along with the identity label $\boldsymbol{c}^i_{\mathcal{P}}$ and the illumination label $c^i_{\mathcal{I}}$. Initialized parameters $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{I}}$. Hyperparameters $\lambda_{1,2,3}$, ξ and learning rate μ^t. The number of iterations $t \leftarrow 0$.

Output: Parameters θ_E, $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{I}}$.

while *not converge* **do**
- $t \leftarrow t + 1$.
- Compute the joint loss by $\mathcal{L}^t = \mathcal{L}^t_{\mathcal{P}} + \lambda_3 \mathcal{L}^t_{\mathcal{I}}$.
- Compute the back-propagation error $\frac{\partial \mathcal{L}^t}{\partial \boldsymbol{x}^t_i}$ for each i by $\frac{\partial \mathcal{L}^t}{\partial \boldsymbol{x}^t_i} = \frac{\partial \mathcal{L}^t_{\mathcal{P}}}{\partial \boldsymbol{x}^t_i} + \lambda_3 \frac{\partial \mathcal{L}^t_{\mathcal{I}}}{\partial \boldsymbol{x}^t_i}$.
- Update the parameters θ_E, $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{I}}$

end

Phase II: Generator training. With the disentangled feature learning modules $\{E, \mathcal{H}_{\mathcal{P}/\mathcal{I}}, \mathbf{W}_{\mathcal{P}/\mathcal{I}}, \mathbf{b}_{\mathcal{P}/\mathcal{I}}\}$ fixed, we optimize the image generator G with the Adam optimizer in phase II. The learning rate is set to 0.01 and reduced to 0.1 of its previous value every 40 epochs. The training batchsize is 64. Our task is to reconstruct the input images, *i.e.*, minimize the *MSE* loss. The objective function is

$$\arg\min_{\theta_G} \mathcal{L}_{\mathcal{G}} \,, \tag{8}$$

where θ_G represents parameters of the generator G.

(a) Market-1501++ (b) DukeMTMC-reID++

Figure 3: **Sample images of the two simulated datasets.** Each row shows images of the same identity. For each identity, the images have different illuminations. (a) The Market-1501++ dataset. (b) The DukeMTMC-reID++ dataset.

Phase III: Joint training. Finally, we jointly train the entire network in an end-to-end manner and the overall objective function is expressed as

$$\arg\min_{\theta_E,\theta_{\mathcal{P}},\theta_{\mathcal{I}},\theta_G} \mathcal{L}_{\mathcal{P}} + \lambda_3 \mathcal{L}_{\mathcal{I}} + \lambda_4 \mathcal{L}_{\mathcal{G}} \,. \tag{9}$$

We use Adam optimizer for optimizing the overall objective function. The initial learning rate for the feature disentanglement part and the generator part is set to 1×10^{-4} and 1×10^{-3} respectively. It decreases to 0.1 of its previous value every 50 epochs. The hyperparameters are $\lambda_3 = 1$ and $\lambda_4 = 2$. The generation process gifts the network more ability to disentangle the illumination feature. Hence, a joint learning manner will better balance the network's abilities of re-identification and disentanglement.

IV. Experiments

A. Experimental Settings

Datasets. There are two widely used datasets under normal illumination, *i.e.*, Market-1501 [12] and DukeMTMC-reID [13]. For Market-1501, it consists of 12,936 images of 751 identities for training and 19,281 images of 750 identities in the gallery set for testing. For DukeMTMC-reID, it contains 16,522 training images with 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images.

Based on these two datasets, we constructed two simulated illumination-adaptive datasets. Considering that a slight illumination variation does not change the representation too much, and also that multiple scales of illuminations are required to simulate a wide range of illumination variation, we selected nine scales of illuminations. We adapt each image to a random one of nine illuminations. We apply a random gamma adjustment to each channel of the common images to produce the illumination-adaptive images, which is similar to [44]. The recorded value of an image captured by a camera is usually nonlinearly mapped from its corresponding scene radiance, and the nonlinearity often can be well approximated by a power function. The variance in the real-world illumination is then nonlinearly related to the image intensity. Therefore, we applied the nonlinear gamma transform to the

decomposed illumination for simulating images under different illumination conditions. To make the illumination change reasonable, we further add Poisson noise with peak value = 10 to illumination changed images. Finally, we constructed the simulated illumination-adaptive datasets and named them as Market-1501++ and DukeMTMC-reID++. Figure 3 gives examples of these two simulated datasets. We consider that the gamma transform is insufficient to model non-global illumination variation. However, although local illumination variation could affect face identification significantly, person ReID relies mostly on global information and is consequently less sensitive to local illumination variation. Thus, we only consider global illumination variation.

Evaluation metrics. To indicate the performance, the standard Cumulative Matching Characteristics (CMC) values and mean Average Precision (mAP) are adopted [12], since one person has multiple ground truths in the gallery set.

Real-world images. To prove the effectiveness of the proposed method on reducing the effect of illuminations, we have also collected some real-world images with different illuminations. Note that we collected real-world images indoors. Because of the controllable indoor environment, we can collect diverse real-world images easily. We will collect more real-world images outdoors (at morning, noon and nightfall) in the future. Through calculating the distances, we show the ability of the proposed IID network. Some examples are shown in Figure 5.

For analyzing illumination conditions of related datasets, we use the average luminance value of each image to represent its illumination and investigate the illumination distributions of the Market-1501 and DukeMTMC-reID datasets. We also made a comparison of the variance of the image illumination on related datasets. Figure 4 shows the results. We can find that the illumination variances of the simulated datasets and the real dataset are much more significant than those of the existing ReID datasets. Hence, existing ReID datasets were captured in a short period of time, and the images have relatively uniform illumination conditions. We introduce a person re-identification task that has more significant illumination variations and requires better illumination adaption.

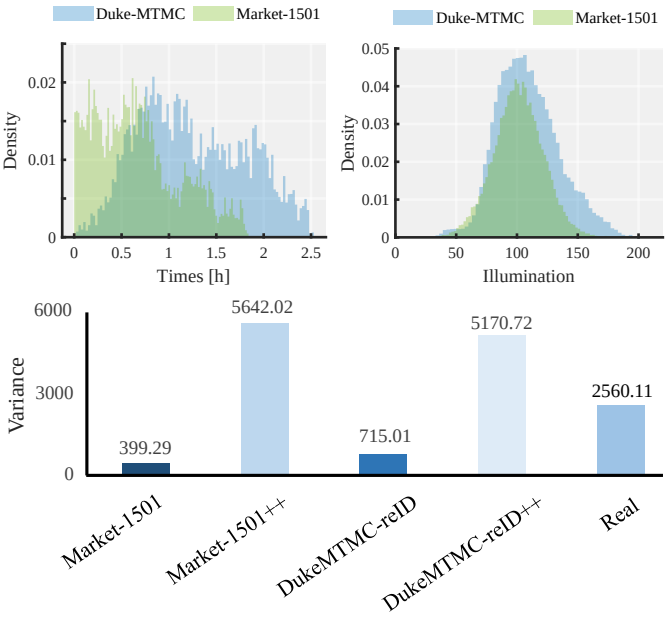


Figure 4: **Illumination analysis of related datasets.** We use the average value of the pixel luminance of each image to represent its illumination. The top two figures show the illumination distributions of Market-1501 and DukeMTMC-reID datasets. We can find that images were captured around one and two hours respectively, and their illuminations do not change too much from the mean value 100. The bottom figure shows the variance of the image illumination on related datasets. We can find that the illumination variances of the simulated datasets (Market-1501++ and DukeMTMC-reID++) and the real dataset are much larger than that of the existing ReID datasets (Market-1501 and DukeMTMC-reID).

B. Comparison with State-of-The-Arts

In this subsection, we make comparisons with the state-of-the-art methods. We exploit the Market-1501++ and the DukeMTMC-reID++ datasets to evaluate the methods. As IA-ReID is new, there are barely methods for comparisons. We selected DenseNet121 [45], PCB [46] and ResNet50 [47] as the comparison methods. DenseNet121 and ResNet50 are two popular baseline networks in ReID. PCB is one of the state-of-the-art ReID methods. Table I list their results. Note that the notation with a 'w/ Train' suffix means that the indicated model has been trained on the illumination-adaptive datasets before testing. From the table, we can find that the results of DenseNet121, PCB and ResNet50 drop dramatically when dealing with the IA-ReID datasets. However, all of these three deep networks can receive very high promotions when training on the illumination-adaptive data, which means that the deep learning network can somehow deal with part of the illumination-adaptive issue if given proper training data.

We took ResNet50 as our baseline. From Table I, we can find that comparing with the baseline (serving as our backbone network), our method improves the performance on both the simulated Market-1501++ and DukeMTMC-reID++ datasets. Thus, The improvement against the baseline better shows the effectiveness of the proposed method as they share the same backbone architecture. Although we do not report improvements against other methods, from Table I, it is clear that our method outperforms all methods. Note that our method uses a very basic baseline (ResNet50), and makes a considerable improvement (8.51% mAP on the Market-1501++ dataset and 6.57% mAP on the DukeMTMC-reID++ dataset). However, compared with PCB (designed to address the misalignment challenge in general ReID tasks), our method does not pay attention to the misalignment problem, which still exists in the ReID task, making our improvement less remarkable. We consider that if we design a new module for the misalignment problem as PCB does, our method will outperform PCB with a much larger gain.

C. Ablation Study

Our method consists of four kinds of losses. The triplet loss $\mathcal{L}_{\mathcal{P}}^{T}$ and the cross-entropy loss $\mathcal{L}_{\mathcal{P}}^{S}$ are responsible for extracting robust person features. The regression loss $\mathcal{L}_{\mathcal{I}}$ is responsible for predicting proper illumination features. The *MSE* loss $\mathcal{L}_{\mathcal{G}}$ is used for image reconstruction. Removing $\mathcal{L}_{\mathcal{P}}^{T}$

Table I: Comparison with the state-of-the-art methods on the Market-1501++ and DukeMTMC-reID++ datasets. CMC-1, CMC-5, CMC-10 (%) and mAP (%) are reported.

Method	**Market-1501++**				**DukeMTMC-reID++**			
	CMC-1	CMC-5	CMC-10	mAP	CMC-1	CMC-5	CMC-10	mAP
DenseNet121 [45]	0.74	2.29	3.53	0.73	1.21	2.74	4.13	0.80
DenseNet121 w/ Train	70.60	85.36	89.66	49.79	64.45	77.82	82.45	45.12
PCB [46]	0.56	1.69	2.91	0.54	0.72	2.15	3.23	0.49
PCB w/ Train	72.55	85.22	90.08	53.11	65.98	77.93	82.21	45.15
ResNet50 [47]	0.42	1.16	2.05	0.39	0.54	1.97	3.14	0.50
ResNet50 w/ Train (Baseline)	66.18	81.97	87.02	47.71	62.07	75.54	88.08	42.63
IID	**73.37**	**86.55**	**91.01**	**56.22**	**68.11**	**79.75**	**91.27**	**49.20**
Improvement over baseline	7.19↑	4.58↑	3.99↑	8.51↑	6.04↑	4.21↑	3.19↑	6.57↑

Table II: Ablation study on the Market1501++ dataset. CMC-1 (%) and mAP (%) are reported.

Method	**Components**				**Market-1501++**	
	$\mathcal{L}_{\mathcal{P}}^{T}$	$\mathcal{L}_{\mathcal{P}}^{S}$	$\mathcal{L}_{\mathcal{I}}$	$\mathcal{L}_{\mathcal{G}}$	CMC-1	mAP
Baseline	✓	✓	×	×	66.18	47.71
IID (no G)	✓	✓	✓	×	71.54	55.17
IID (no triplet for id)	×	✓	✓	✓	64.14	45.87
IID (no softmax for id)	✓	×	✓	✓	65.21	46.53
IID (no illum.)	✓	✓	×	✓	70.79	54.57
IID	✓	✓	✓	✓	73.37	56.22

Figure 5: **Sample images of the collected real-world images.** Each row shows images of one identity.

or $\mathcal{L}_{\mathcal{P}}^{S}$ will give less constraint to the person feature, and thus degrade the re-identification performance. $\mathcal{L}_{\mathcal{I}}$ and $\mathcal{L}_{\mathcal{G}}$ influence the effectiveness of illumination disentanglement. Removing either of them will degrade the re-identification performance indirectly.

Table III: Comparison with the baseline on the Market-1501 and DukeMTMC-reID datasets. CMC-1 (%) and mAP (%) are reported.

Method	**Market-1501**		**DukeMTMC-reID**	
	CMC-1	mAP	CMC-1	mAP
Baseline	88.84	71.49	79.71	61.77
IID	88.45	71.46	78.10	60.56

Here, we take the Market-1501++ dataset for the ablation study. When removing both of $\mathcal{L}_{\mathcal{I}}$ and $\mathcal{L}_{\mathcal{G}}$ (the baseline), the performance drops to 66.18% CMC-1 and 47.71% mAP, as it can only rely on the person feature without separating the illumination feature. When removing $\mathcal{L}_{\mathcal{G}}$, the performance does not drop so much as the baseline. The loss $\mathcal{L}_{\mathcal{I}}$ is useful for separating the illumination information and promotes the re-identification result. When removing $\mathcal{L}_{\mathcal{P}}^{T}$ or $\mathcal{L}_{\mathcal{P}}^{S}$, the performance drops dramatically to be even worse than the baseline. So the loss for re-identification is essential for the IA-ReID task. When removing $\mathcal{L}_{\mathcal{I}}$, the performance does not drop so much as the baseline. So even without the loss of illumination regression, the generation process can also benefit the illumination disentanglement.

D. Experiments on General ReID Datasets

Although we design a new network for the IA-ReID task, we do not expect the proposed IID network performing poorly on the general ReID dataset. As our network is proposed based on the baseline network Resnet50, we make a comparison with the baseline. The results are listed in Table III. We can find that the results do not change too much. Although the IID network is specially designed for the illumination-adaptive condition, it is still suitable for the general ReID task.

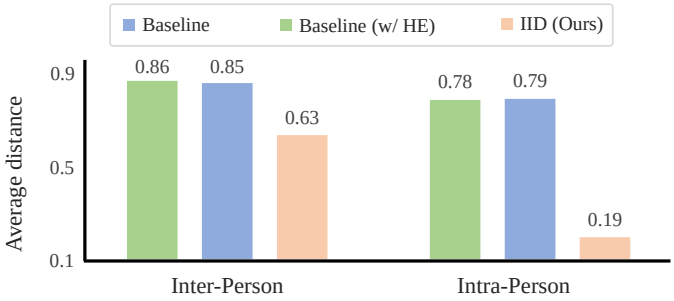


Figure 6: **The average values of intra-person distances and inter-person distance for real-world images.** The experiments are respectively conducted by the baseline method, the baseline method fed with image after histogram equalization (w/ HE) and our IID method.

E. Experiments on Illumination Prediction

As we know, the disentangled illumination feature can be used to predict the illumination scale of each image. We test its prediction accuracy on the Market-1501++ and DukeMTMC-reID++ datasets. The accuracy values are very high, which are 98.74% and 98.53% respectively for the Market-1501++ and DukeMTMC-reID++ datasets. It means that the disentangled illumination feature can well represent the illumination scale.

F. Experiments on Local Illumination Change

The synthetic datasets above are generated by the global illumination change. In this subsection, we generate datasets with local illumination changes and investigate the effectiveness of the proposed method.

We have tried to synthesize local illumination changes in two ways. 1) We exploite a person segmentation method [48] to obtain the foreground person of each image and then change its illumination by using gamma correction. We set seven different gamma values. We name this synthesis process as foreground-based illumination change. Some generated examples are shown in Figure 8(a) and Figure 8(b). 2) We randomly select a patch in each image and then change its illumination by using gamma correction. We use three different sizes of patches, *i.e.*, 32×32, 32×64 and 64×32. We also set seven different gamma values. We name this synthesis process as patch-based illumination change. Some generated examples are shown in Figure 8(c) and Figure 8(d). We exploit these two kinds of synthesis processes to generate new Market and Duke datasets, respectively.

Then, we conduct experiments on these four synthetic datasets with local illumination changes. We compare our methods with ResNet50, DenseNet, and PCB. Figure 9 shows the results, which demonstrates that the proposed method is more effective in the condition of local illumination changes.

G. Experiments on the Real-world Images

We recruited 15 volunteers, and for each volunteer, we collected ten images under different illumination conditions. Some examples are shown in Figure 5. We calculated the intra-person distances and inter-person distances respectively with the baseline model, the baseline model with histogram equalization, and the proposed IID network. Figure 6 shows the average intra-person and inter-person distances calculated by different methods. We can find that the proposed IID is more effective in reducing the intra-person distance, *i.e.*, to alleviate the effect of the illumination change.

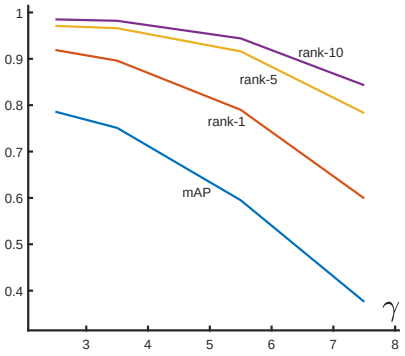


Figure 7: Results on low illumination conditions.

H. Experiments on Low Illumination Conditions

In this subsection, we report experiments on low illumination images. We created four low-illumination testing sets based on Market-1501 by using four gamma corrections. Our model was trained on the Market-1501++ dataset (random illuminations) and tested on these four testing sets. Figure 7 shows the results. We can find that the performances decrease as the illumination becomes lower. We consider that some useful information disappears and some noise is introduced in very low illumination conditions, and the proposed model can not extract enough useful information from the input image. However, in an appropriate low illumination range, our method can still work well.

I. Experiments on Different Parameters

To investigate the parameter λ_3 and λ_4 in the Equation (9), we conducted experiments on the Market-1501++ dataset with 1) fixed λ_4 and different λ_3 values, and 2) fixed λ_3 and different λ_4 values. Table IV reports the results. We can find that when $\lambda_3 = 1$ and $\lambda_3 = 2$, our model achieves the best result.

λ_3	λ_4	mAP(%)
0.1	1	44.78
0.5	1	50.44
1	1	54.31
2	1	53.81
5	1	52.75
1	0.1	53.19
1	0.5	53.31
1	**2**	**56.22**
1	5	48.47

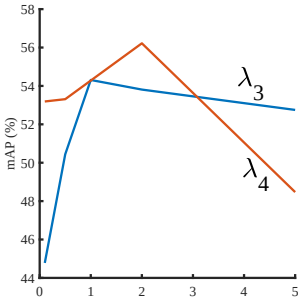
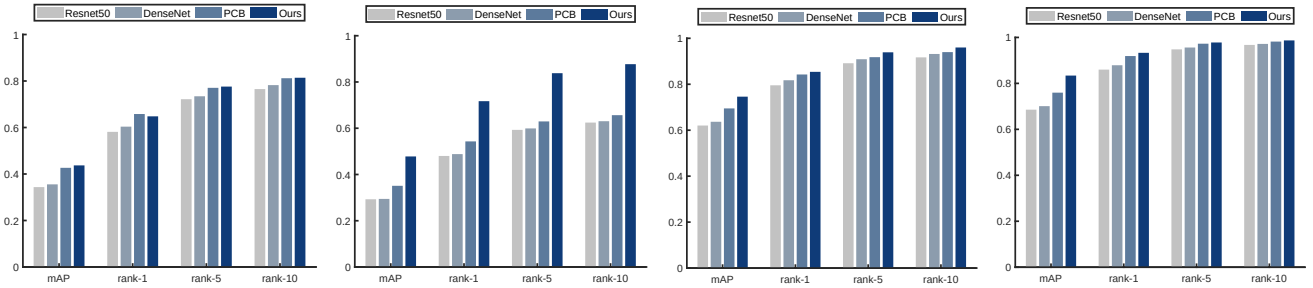


Table IV: Results on different parameters λ_3 and λ_4.

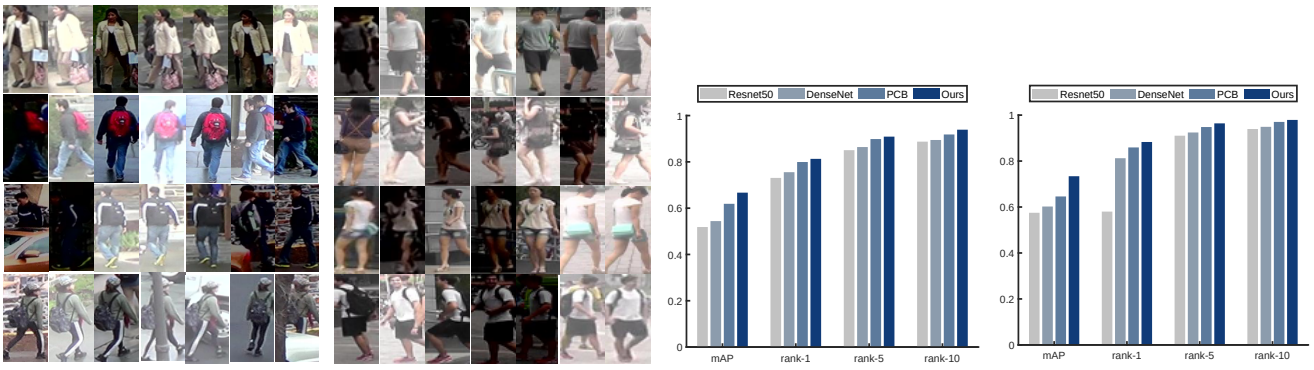
(a) Foreground Duke (b) Foreground Market (c) Patch Duke (d) Patch Market

Figure 8: Examples of synthetic datasets with local illumination changes. (a) and (b) show samples by the foreground-based illumination change on the Duke and Market datasets, respectively. The images are respectively the original image, the person segmentation result, and the corresponding generated image. (c) and (d) show samples by the patch-based illumination change on the Duke and Market datasets, respectively.



(a) Foreground Duke (b) Foreground Market (c) Patch Duke (d) Patch Market

Figure 9: ReID results on four synthetic datasets with local illumination changes.



(a) Duke Samples (b) Market Samples (c) Duke Results (d) Market Results

Figure 10: Examples and results of synthetic datasets with Y channel illumination changes. (a) and (b) show samples by the YUV illumination change on the Duke and Market datasets, respectively. (c) and (d) show the ReID results.

J. Experiments on Illumination-adaptive Datasets with Y Channel Changes

We exploit a non-parametric simulation method to generate images with different illuminations. We transfer the format of the image from RGB to YUV. The Y channel often reflects the illumination condition, and we change the illumination of each image by adjusting its Y value. In the experiments, we set seven different ΔY values and adjust Y by adding ΔY to Y. Some generated examples are shown in Figure 10(a) and Figure 10(b). We compare our methods with ResNet50, DenseNet, and PCB. Figure 10(c) and Figure 10(d) show the results, which demonstrate that our proposed method is more effective for illumination-adaptive datasets synthesizing by varying the Y values.

V. Conclusion

This paper raises a new issue, which has not been investigated before as far as we know. Traditional models for the

multi-illumination condition may not work well for this task. We propose to disentangle the illumination feature apart to address the new problem. Experimental results illustrate that the traditional model has a significant drop of performance when the illumination of gallery images are different and the scales vary unsteadily, and demonstrate the effectiveness of the proposed network. The idea of addressing the multi-illumination problem can be extended to related video surveillance applications, such as tracking [49] and activity analysis [50], [51], [52].

Disentangled representation learning is an effective mechanism to distill meaningful information from the mixed feature representation. As far as we know, we are the first to explore this mechanism to the novel illumination-adaptive person ReID problem. Our illumination-adaptive setting is more practical for long-term ReID. Compared with the work in the field of the face identification, the proposed model further predicts the scale of the illumination and strengthens the re-identification ability of identity features. By showing its effectiveness, we hope the method can also help address problems in other application domains. Also, our main contribution lies in pointing out an important problem for illumination-adaptive person ReID and demonstrating promising initial results. We believe that our paper could inspire more work in this direction and make the ReID techniques more practical.

In the future, we will improve our method in the following aspects: 1) Experiments show that our method does not work very well in very low-light conditions, because of the lose of person-related information. We will investigate how to recover person-related information under such a condition. 2) Although we have used a soft label strategy and adopted the classifier problem to do regression, the framework still tends to disentangle the illumination to different levels of scales. We will study how to estimate the illumination to a continuous value, and make our framework a genuine regression way for the illumination disentanglement. 3) Our method focuses on the illumination problem, while ignoring the other challenges, such as misalignment, occlusion, low-resolution. In the future, we will integrate with other modules to improve ReID performance as a whole.

REFERENCES

[1] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[2] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[3] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "Deeplist: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 513–524, 2016.

[4] Z. Wang, R. Hu, C. Chen, Y. Yu, J. Jiang, C. Liang, and S. Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 3006–3020, 2018.

[5] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2376–2386, 2019.

[6] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157–5166.

[7] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2013–2025, 2019.

[8] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4678–4686.

[9] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface." in *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2016, pp. 2669–2675.

[10] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded SR-GAN for scale-adaptive low resolution person re-identification," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2018, pp. 3891–3897.

[11] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "FD-GAN: Pose-guided feature distilling gan for robust person re-identification," in *Proceedings of the Advances in Neural Information Processing Systems*, 2018, pp. 1222–1233.

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[13] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[14] A. Bhuiyan, A. Perina, and V. Murino, "Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems," in *IEEE International Conference on Image Processing*, 2015, pp. 2329–2333.

[15] Y. Wang, R. Hu, C. Liang, C. Zhang, and Q. Leng, "Camera compensation using a feature projection matrix for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1350–1361, 2014.

[16] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3291–3300.

[17] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 178, 2017.

[18] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6306–6314.

[19] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[21] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960–3969.

[22] 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.

[23] F. Ma, X. Zhu, X. Zhang, L. Yang, M. Zuo, and X.-Y. Jing, "Low illumination person re-identification," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 337–362, 2019.

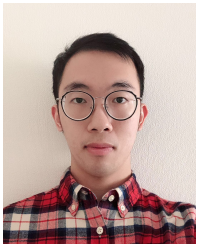
[24] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5380–5389.

[25] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2018, pp. 1092–1099.

[26] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2018, pp. 677–683.

[27] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, pp. 407–419, 2019.

[28] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 618–626.
[29] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, and S. Satoh, "Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification," *arXiv preprint arXiv:1905.10048*, 2019.
[30] K. Kansal, A. Subramanyam, Z. Wang, and S. Satoh, "Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2019.2963721
[31] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, and F. Wu, "Efficient parallel framework for hevc motion estimation on many-core processors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 12, pp. 2077–2089, 2014.
[32] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, and Q. Dai, "A fast uyghur text detector for complex background images," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3389–3398, 2018.
[33] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, and Q. Dai, "Cross-modality bridging and knowledge transferring for image understanding," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2675–2685, 2019.
[34] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE Transactions on Multimedia*, 2019. [Online]. Available: http://dx.doi.org/10.1109/TMM.2019.2924576
[35] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," *arXiv preprint arXiv:1610.05586*, 2016.
[36] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2080–2089.
[37] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 35–51.
[38] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Proceedings of the Advances in Neural Information Processing Systems*, 2018, pp. 1287–1298.
[39] B. Lu, J.-C. Chen, and R. Chellappa, "Unsupervised domain-specific deblurring via disentangled representations," *arXiv preprint arXiv:1903.01594*, 2019.
[40] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems*, 2015, pp. 3546–3554.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.
[44] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.
[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
[46] 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*, 2018, pp. 480–496.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[48] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 418–434.
[49] J. Cui, Y. Liu, Y. Xu, H. Zhao, and H. Zha, "Tracking generic human motion via fusion of low-and high-dimensional approaches," *IEEE transactions on systems, man, and cybernetics: systems*, vol. 43, no. 4, pp. 996–1002, 2013.
[50] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, "Action2activity: recognizing complex activities from sensor data," in *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 1617–1623.
[51] L. Liu, L. Cheng, Y. Liu, Y. Jia, and D. S. Rosenblum, "Recognizing complex activities by a probabilistic interval-based model," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1266–1272.
[52] H. Zhang, S. Feng, C. Liu, Y. Ding, Y. Zhu, Z. Zhou, W. Zhang, Y. Yu, H. Jin, and Z. Li, "Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario," in *The World Wide Web Conference*. ACM, 2019, pp. 3620–3624.



Zelong Zeng is currently a master student at the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, the University of Tokyo (UTokyo), Tokyo, Japan. His research interests on person re-identification and unsupervised learning.

Zhixiang Wang received his B.E. degree in automation from Nanjing Normal University in 2017. He is currently a master student at the Department of Computer Science and Information Engineering, Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan. His research interests include computer vision and deep learning.

Zheng Wang (M'19) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China, in 2017. He is currently a JSPS Fellowship Researcher at Shin'ichi Satoh's Lab, National Institute of Informatics, Japan. His research interests focus on person re-identification and instance search. He received the Best Paper Award at the 15th Pacific-Rim Conference on Multimedia (PCM 2014) and the 2017 ACM Wuhan Doctoral Dissertation Award.

Yinqiang Zheng received his Bachelor degree from the Department of Automation, Tianjin University, Tianjin, China, in 2006, Master degree of engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009, and Doctoral degree of engineering from the Department of Mechanical and Control Engineering, Tokyo Institute of Technology, Tokyo, Japan, in 2013. He is currently an associate professor in the National Institute of Informatics, Japan. He has served as program committee member for major international conferences, like CVPR, ICCV, ECCV, MICCAI, area chair for MVA2017, DICTA2018, ISAIR2019, guest editor for Pattern Recognition Letter, and workshop co-organizer for ICPR2018 and ICCV2019.

Yung-Yu Chuang (M'04) received his B.S. and M.S. from National Taiwan University in 1993 and 1995 respectively, and the Ph.D. from the University of Washington at Seattle in 2004, all in Computer Science. He is currently a professor with the Department of Computer Science and Information Engineering, National Taiwan University. His research interests include computational photography, computer vision and rendering.

Shin'ichi Satoh (M'04) received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. He has been a Full Professor with the National Institute of Informatics, Tokyo, Japan, since 2004. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997. His current research interests include image processing, video content analysis, and multimedia databases.