ER-PIQA: A TASK-GUIDED PEDESTRIAN IMAGE QUALITY ASSESSMENT VIA EMBEDDING RECONSTRUCTION

Yanzhe Zhong[⋆], *Huadong Pan*[⋆†], *Bangjie Tang*[⋆], *Zhonggeng Liu*[⋆], *Yiming Zhu*[⋆], *Jun Yin*[⋆]

[⋆] Advanced Research Institute of Zhejiang Dahua Technology Co.,Ltd., China
[†] Zhejiang Provincial Key Laboratory of Harmonized Application of Vision & Transmission, China

ABSTRACT

Image quality is an important factor for pedestrian recognition systems. Pedestrian image quality assessment aims at evaluating images in order to provide more reliable and stable images for the following analysis process. Previous work proposed supervised solutions that require artificially or manually labelled quality values. However, due to the lack of a clear quality definition and the variety of recognition tasks, there are still some difficulties in subjective labeling methods. In this paper, a novel task-guided method is proposed to measure pedestrain image quality based on embedding reconstruction without the involvement of subjective labels. Considering the various attention area in different tasks, the pedestrian image quality can be estimated by comparing the similarity between the semantics representation and tasks embedding after reconstruction. Experiments on both person re-identification and pedestrian attribute recognition show advantages of the proposed ER-PIQA. Meanwhile, our approach can be integrated into current recognition systems and adaptively modified for other tasks beyond pedestrian recognition.

Index Terms— pedestrian image quality assessment, task-guided, embedding reconstruction

1. INTRODUCTION

Nowadays, with the rapid development of information technology, more and more recognition systems have been applied in our daily life. As one of the most important components, pedestrian images are often captured, stored, identified and analyzed in numerous practical application systems. For example, Person reidentification [1–3], Pedestrian attribute recognition [4, 5], and Face verification [6, 7] have been widely used in computer vision applications such as video surveillance and financial identity authentication. Due to the influence of various unconstrained environments in practice, recognition accuracy will be significantly reduced and the system results will be unstable. In addition, the image is affected by various factors in the process of acquisition, compression, transmission and storage, which will also lead to the degradation of recognition performance.

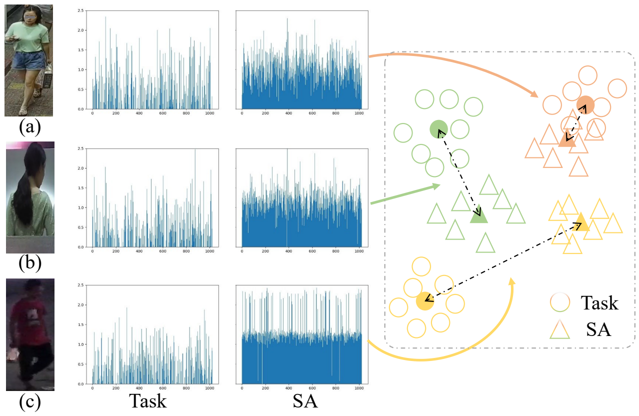


Fig. 1. Illustration of our motivation, best viewed in color. There are three images with large differences of quality in the left of the figure. In the middle are the features of the image under the recognition task(ReID) and self-attention(SA) mechanism. On the right are the distribution characteristics of these features on the hyperplane. Obviously, the sensitivity of task and self-attention to features is different, and there is a certain relationship between image quality and feature distribution distance.

Plenty of researchers try to consider various situations as much as possible in terms of recognition ability, strengthen generalized recognition ability to improve accuracy, but sometimes the performance is still not satisfied. In order to ensure the accuracy and stability of recognition system, the pedestrian image quality assessment is developed to select high-quality images and drop low-quality ones for better identification and analysis in recognition system.

Most of image quality assessment methods [8] adopt the subjective annotation to obtain the quality score for natural scene images, and apply supervised learning by using these labels. However, in the process of pedestrian image perception, it is particularly difficult to make an accurate definition of the image quality. For instance, when the system needs to recognize and analyze the whole body attributes of a person, not only the image should be clear enough, but the target in it should be complete and unique. What's more, tasks such as person re-identification, face verification and gait analysis [9]

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

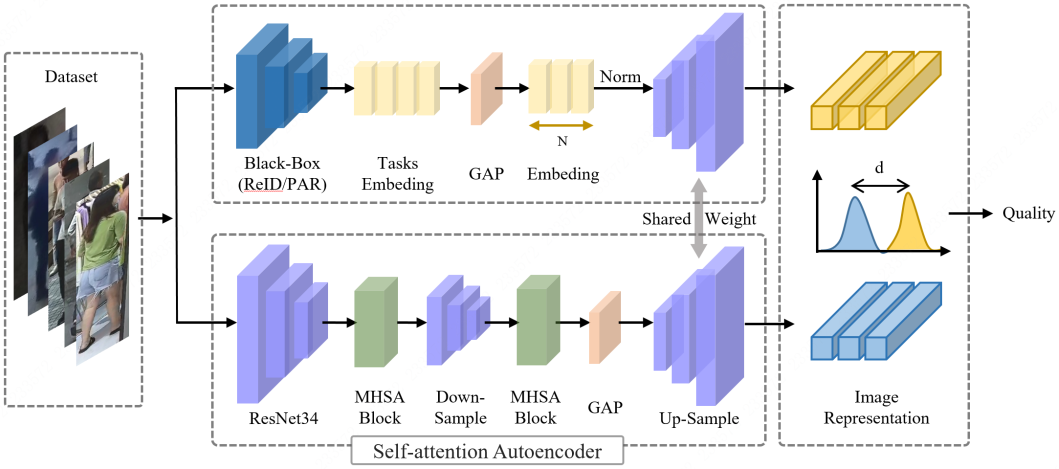


Fig. 2. The framework of ER-PIQA. Step 1:Train the self-attentional autoencoder below; Step 2:Generate embeddings from encoder and black-box model respectively; Step 3:Reconstruct embeddings and calculate the similarity as the image quality.

have totally different attention to different parts of body. It can be seen that the quality of pedestrian image depends on different recognition and analysis tasks. Therewise, it is necessary to find a reliable and task-guided method for pedestrian image quality assessment.

In this paper, we propose a novel unsupervised pedestrian image quality assessment method called ER-PIQA. Our solution attempts to explore the distance distribution between task focus and self-attention mechanism. The image embedding can be obtained through the task model, and then the generator will reconstruct it. By comparing the similarity between the generated image and the original one, the replicability of embedding is measured to determine the pedestrian quality. Experiments show that ER-PIQA can reasonably evaluate quality and perform robustness for different recognition tasks. The main idea of approach is shown in Fig. 1, and the major contributions of this paper are summarized as follows

1. An end-to-end task-guided pedestrian image quality assessment method is proposed to solve the image management problem in the application systems, which predict the quality adaptively according to the related recognition task.
2. The proposed ER-PIQA achieves point-to-point prediction in real time, rather than being limited to a sequence data set of a single target.
3. Our method can be regarded as a general framework for quality perception, and other similar tasks can also be applied, not just limited to pedestrians.

2. PROPOSED METHOD

In this section, we propose a task-guided pedestrian image quality assessment via Embedding Reconstruction(ER-PIQA). Through task embedding reconstruction, the replicability of features is measured to evaluate the quality of pedestrian images. The framework of the proposed ER-PIQA is shown in Fig. 2. Due to the wide variety of pedestrian recognition tasks, we take Person Re-identification(ReID) as an example for detailed explanation, mainly including self-attention autoencoder, task embedding generation, embedding reconstruction and quality evaluation.

2.1. Self-attention autoencoder

In order to better obtain image information represetation and have the ability to reconstruct embedding, a self-attention autoencoder is designed, which is composed of encoder and decoder as shown in Fig. 2. Since Resnet [10] is an effective feature extraction network, which is often used in image classification and regression tasks. the encoder refers to parts of Resnet34 to extract preliminary features. In addition, two multi-head self-attention transformer modules [11] are added at the appropriate position in model backbone, which effectively improve the representability of the attention area in images. The multi-head self-attention module is shown as Fig. 3, all2all attention is performed on a 2D featuremap with split relative position encodings R_h and R_w for height and width respectively. The attention logits are $qk^T + qr^T$ where q; k; r represent query, key and position encodings respectively. $\oplus$ and $\otimes$ represent element wise sum and matrix multiplication respectively, while 1×1 represents a pointwise convolution.

Given a set of pedestrian training set X with length N, the entire autoencoder is trained with dataset which composes of paired images. $D = \{(x_1, x_1), (x_2, x_2), \ldots, (x_n, x_n) | x_i \in X\}$. L2 norm is used as the loss function, which is defined as follows

$$loss(\{D^{(i)}\}; W_{en}, W_{de}) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\| \qquad (1)$$

where W_{en} and W_{de} denote encoder and decoder parameters, respectively. During the test step, when given a pedestrian image I, its semantic representation can be obtained in the following way

$$E_{SA} = W_{en}(I) \tag{2}$$

where E_{SA} represents the image embedding feature extracted by the encoder.

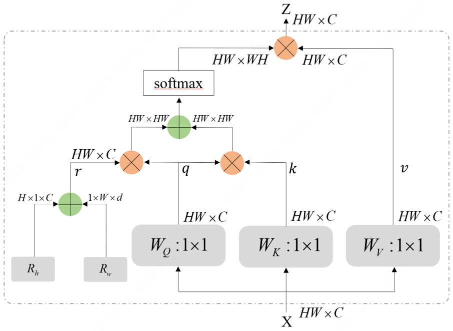


Fig. 3. The details in Multi-Head Self-Attention(MHSA).

2.2. Task Embedding Generation

A black-box is utilized to accommodate as many different recognition tasks as possible, because of various quality definitions in different tasks. For example, person reidentificationas model backbone is depolyed in the black-box. When the test image is input, the corresponding features can be obtained through the box. We flatten the features into embeddings and add a one-dimensional global average pooling(GAP) layer to ensure that the output lengths of different black-box models are the same.

In addition, different black-box models have different parameters settings during training. the values in the output embeddings will be quite different. So, the method of directly comparing the similarity between task embeddings and semantic represetation to determine the quality of the image is not reasonable enough. Therefore, it is necessary to normalize the output embedding to eliminate this effect. The parameters of the black-box model are expressed as W_{task}, and the generated task-guided embeddings are as follows

$$E_{Task} = norm(GAP(W_{task}(I))) \tag{3}$$

where E_{Task} represents the generate task embedding.

2.3. Embedding Reconstruction and Quality Evaluation

In order to reduce the disparities caused by different black-box models training settings, we introduce GAP and normalize data when acquiring task embeddings. At the same time, to enhance the quality differences of pedestrian images, task embeddings and image representations is reconstructed respectively through decoders. The process is shown as follows

$$I_{SA} = W_{de}(E_{SA}), I_{Task} = W_{de}(E_{Task}) \tag{4}$$

where I_{SA} and I_{Task} denote the result of forward inference with image representations and task embeddings, respectively. According to previous assumption, image quality is measured by the replicability of task-guided embeddings. Therefore, the pedestrian image quality is calculated by comparing the similarity of the two outputs after reconstruction.

$$Q(I) = F(I_{SA}, I_{Task}) \tag{5}$$

where $F(\bullet)$ is the mapping function from I_{SA} and I_{Task} to $Q(I)$. In the experiment, we find that the two output difference can effectively measure the image quality of pedestrians. Such discovery can also be explained intuitively. When the follow-up task considers the image quality to be poor, the extracted features are fuzzy and unstable, and it is difficult to restore; on the contrary, when the image quality is good, the extracted features are robust and stable, and the restored image is closer to the original image. Based on the above analysis, we propose to take advantage of the Wasserstein metric to measure image restoration situation as $q(I)$, which is expressed by

$$\begin{aligned} q(I) &= WD(I_{SA}||I_{Task}) \\ &= \inf_{\gamma\in\Pi(I_{SA},I_{Task})} E_{(I_{SA},I_{Task})\sim\gamma}[||I_{SA} - I_{Task}||] \end{aligned} \tag{6}$$

where $WD(\bullet)$ denotes Wasserstein distance,$\Pi(I_{SA}, I_{Task})$ denotes the set of all pixel whose marginals are respectively I_{SA} and I_{Task}. We define the pedestrian image quality

$$Q(I) = \frac{2}{1 + e^{2*q(I)}} \tag{7}$$

The embeddings of image I under the target task has good replicability and is considered as high sample quality Q.

3. EXPERIMENTS

3.1. Setup

Experiments are conducted on PRID2011 [12], iLiDS-VID [13], Market-1501 [14] and DukeMTMC-reID [15] datasets. PRID2011 dataset contains 1134 identities and each person has 5 to 675 images. iLIDS-VID dataset has 300 people, and each person has two sets also captured from different positions. The Market-1501 dataset includes 1501 pedestrians and 32668 detected pedestrian rectangles captured by 6 cameras. The DukeMTMC dataset provides more than 7,000 single-camera tracks and more than 2,700 independent characters.

Before training the model, the autoencoder network is pre-trained on ImageNet dataset. Adam optimization algorithm [16] is adopted in training step, and the initial learning rate is 0.01, which decreased each 20 epoches. In order to prevent overfitting, we utilize weight decay and add regularization to the loss function, The result is reported as the average of 10-fold cross validation.

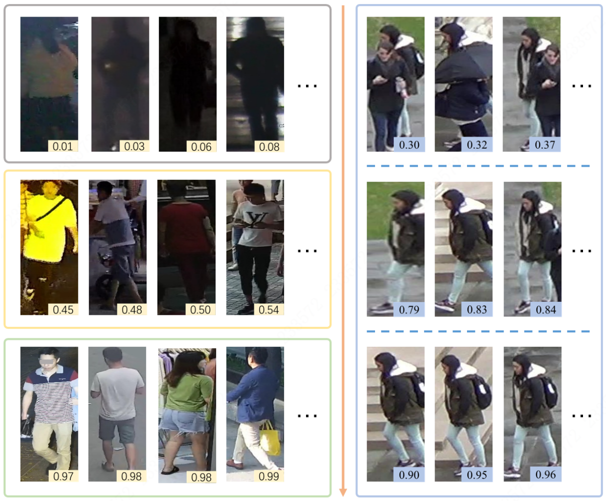
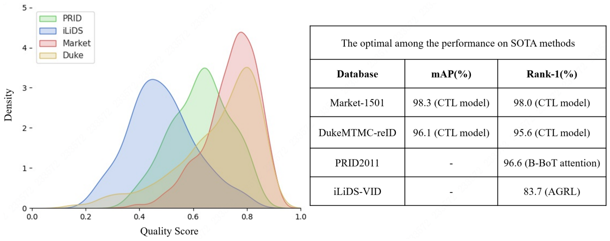


Fig. 4. Left side:performance between different ids; Right side:performance in single id. The pedestrian in the top, middle and bottom rows are sampled from the images with scores in the highest 10%, a 20% window centered at the median, and the lowest 10%, respectively.



The optimal among the performance on SOTA methods

Database	mAP(%)	Rank-1(%)
Market-1501	98.3 (CTL model)	98.0 (CTL model)
DukeMTMC-reID	96.1 (CTL model)	95.6 (CTL model)
PRID2011	-	96.6 (B-BoT attention)
iLiDS-VID	-	83.7 (AGRL)

Fig. 5. Quality score distribution of each dataset and the optimal among performance on state-of-the-art methods

3.2. Qualitative and Quantitative Analysis

In order to explore the meaning of quality, qualitative and quantitative analysis of ER-PIQA is made. At First, we visualize part of the images in the dataset and label them with quality prediction scores. Fig. 4 shows the corresponding results. ER-PIQA has a distinguished ablity of the pedestrian quality no matter within a set of single ID or multi-mixed IDs . What's more, the network is more inclined to high-resolution, complete and simple background images in ReID task.

Secondly, the image quality distribution curves is shown in Fig. 5. The red part denotes the density distribution of Market-1501. DukeMTMC-reID, PRID2011 and iLiDS-VID are represented by yellow, green and blue, respectively. It can be observed that the image quality of the Market-1501 is the best, followed by others. the table on the right shows the optimal indicators of the various methods in each dataset. Apparently, the quality ranking of four datasets is consistent with the corresponding indicator, which explains that ER-PIQA can effectively predict the quality by task guiding.

3.3. Person Re-id and Pedestrian Attribute Recognition

To understand the importance of different image quality measures for the task of person re-identification, we evaluate the deep learning method QAN [17], the traditional image assessment metrics Brisque [18] and Niqe [19](all represented as dotted lines). The performance on a series of testset filtered by different quality threshold is shown in Fig. 6. We ignore the indicator when the remaining gallery is insufficient. According to the phenomenon, ER-PIQA apparently superior to other three methods.

As for pedestrian attribute recognition(PAR), we collecte the Market-1501 and DukeMTMC, since [20] have been labeled with corresponding subjective attribute labels. we select five attributes shared by the two datasets and presented the results of the high and low quality parts in Table 1. Obviously, the performance of high-quality sequences is significantly improved under the promotion of ER-PIQA, as well as low-quality sequences.

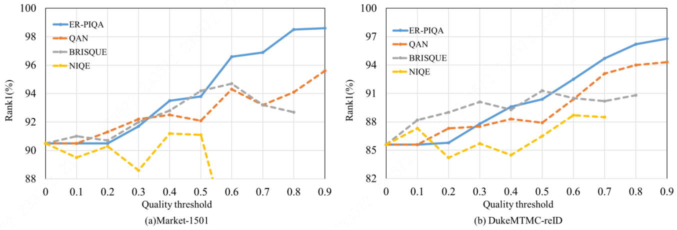


Fig. 6. Rank1 index curves under different quality thresholds on Market-1501 and DukeMTMC-reid.

Table 1. Performance compration on Market and Duke.

attributes	global		high quality part		low quality part	
	Market	Duke	Market	Duke	Market	Duke
Top color	74.57	80.06	78.65	81.14	72.05	78.14
bottom color	70.67	80.25	75.76	81.84	68.34	79.80
gender	92.82	90.54	93.06	91.07	90.75	91.07
Hat	97.79	93.99	99.23	95.30	93.76	93.38
Shoulder bag	83.28	88.41	83.78	88.86	79.32	85.81
average	83.83	86.65	86.10	87.64	80.84	85.64

4. CONCLUSION AND FUTURE WORK

Pedestrian image quality assessment aims at predicting the suitability of images for pedestrian recognition systems. In this paper we propose a novel unsupervised pedestrian image quality assessment by measure the replicability of image under task embedding. Reconstruct the embedding from the recognition model, compare the similarity between the generated and the original and thus, the sample's quality is determined. ER-PIQA showed good accuracy and generalization in experiments under different tasks. Considering this, our future work will expand to more recognition tasks.

5. REFERENCES

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[2] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[3] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, “Person re-identification by probabilistic relative distance comparison,” in *CVPR 2011*, 2011, pp. 649–656.

[4] Yubin DENG, Ping Luo, Chen Change Loy, and Xiaoou Tang, “Pedestrian attribute recognition at far distance,” in *Proceedings of the 22nd ACM International Conference on Multimedia*, New York, NY, USA, 2014, p. 789–792, Association for Computing Machinery.

[5] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang, “A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios,” *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575–1590, 2019.

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[8] Le Kang, Peng Ye, Yi Li, and David Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[9] Zongyi Liu and S. Sarkar, “Improved gait recognition by gait dynamics normalization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 863–876, 2006.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani, “Bottleneck transformers for visual recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16519–16529.

[12] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof, “Person re-identification by descriptive and discriminative classification,” in *Image Analysis*, Berlin, Heidelberg, 2011, pp. 91–102, Springer Berlin Heidelberg.

[13] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, “Person re-identification by video ranking,” in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 688–703, Springer International Publishing.

[14] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, “Mars: a video benchmark for large-scale person re-identification,” in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884.

[15] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, “Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5177–5186.

[16] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[17] Yu Liu, Junjie Yan, and Wanli Ouyang, “Quality aware network for set to set recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[19] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[20] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, “Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5177–5186.