SEMI-SUPERVISED RANKING FOR OBJECT IMAGE BLUR ASSESSMENT

Qiang Li*, Zhaoliang Yao*, Jingjing Wang, Ye Tian, Pengju Yang, Di Xie, Shiliang Pu

Hikvision Research Institute, Hangzhou, China

{liqiang23✉,yaozhaoliang,wangjingjing9,tianye10,yangpengju,xiedi,pushiliang.hri✉}@hikvision.com

ABSTRACT

Assessing the blurriness of an object image is fundamentally important to improve the performance for object recognition and retrieval. The main challenge lies in the lack of abundant images with reliable labels and effective learning strategies. Current datasets are labeled with limited and confused quality levels. To overcome this limitation, we propose to label the rank relationships between pairwise images rather their quality levels, since it is much easier for humans to label, and establish a large-scale realistic face image blur assessment dataset with reliable labels. Based on this dataset, we propose a method to obtain the blur scores only with the pairwise rank labels as supervision. Moreover, to further improve the performance, we propose a self-supervised method based on quadruplet ranking consistency to leverage the unlabeled data more effectively. The supervised and self-supervised methods constitute a final semi-supervised learning framework, which can be trained end-to-end. Experimental results demonstrate the effectiveness of our method. Source of labeled datasets: https://github.com/yzliangHIK2022/SSRanking-for-Object-BA

Index Terms— Object image blur assessment, Pairwise ranking, Quadruplet ranking, Semi-supervised learning

1. INTRODUCTION

In real-life object recognition systems (e.g. face recognition [1], vehicle re-identification [2], license plate recognition [3]), object images usually contain one centering object and appear with different qualities. Practically, it's very important to indicate the object image quality for it can significantly improve the application experience if we can filter out low-quality object image frames. Recently many works [4, 5, 6] turn to utilize the quality (e.g. occlusion, blur, illumination) of object images to improve the application performance.

As a key factor of the image quality [7, 8, 9], object image blur assessment (Object-BA) is not well studied yet and thus our work focuses on this issue. Object-BA can be defined as: for an image that only contains one object to be recognized, the blur level of foreground object requires to be assessed with reasonable blur scores by determining how well the object
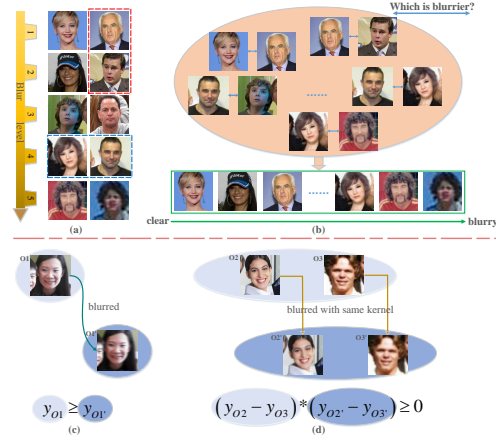
---

*Authors contribute equally to this research.

Fig. 1. Different dataset constructions (a *vs* b) and self-supervised learning (SSL) techniques (c *vs* d): (a) label with limited quality levels; (b) label with pairwise blur orders; (c) pairwise ranking consistency for SSL; (d) our proposed quadruplet ranking consistency for SSL.

image is suitable for recognition . Object-BA only evaluates the foreground object, and it is different from IQA problems since IQA attempts to assess the quality/blur of the entire image in which none or multiple objects may exist. Object-BA is also different with the object quality assessment [4], since it focuses on the global quality assessment which may lack of interpretation, while we aim to decompose the factors affecting object image quality, and solve one of the most important factors assessment, i.e. the object image blur assessment.

Although convolutional neural networks have achieved great success on many visual tasks, their application on quality assessment still suffers from two challenges: lack of abundant images with reliable labels and more effective learning strategy for fine-grained quality assessment. Current image quality assessment datasets either label the image with limited quality levels or synthesize images with different qualities using certain image degradation methods. For examples, BCNet [8] establishes a real-world dataset including 2000 images labeled with only blur or not, which is not sufficient for fine-grained quality assessment. To solve this problem, [10] labels the images with ten quality levels. However, labeling images with accurate quality levels is very
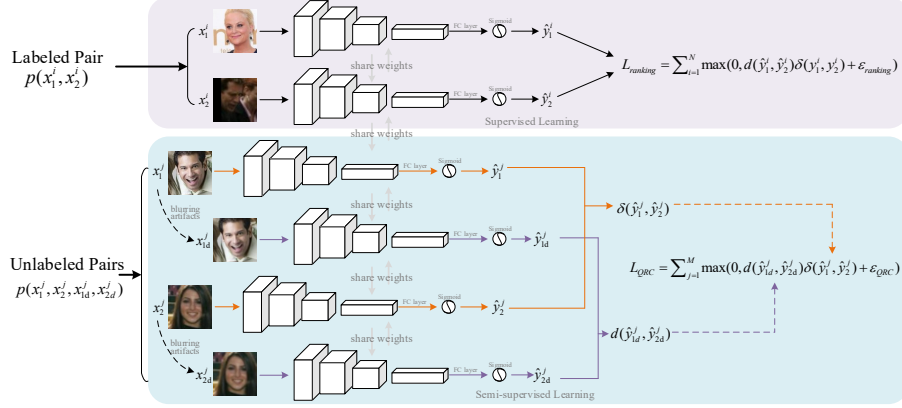
Fig. 2. Overview of our approach which is consisted of supervised learning branch and self-supervised learning branch.

difficult for humans, which may limit the number of quality levels, and lead to confusing labels between inter and intra levels. Overlaps of different blur levels are shown as red dotted boxes in Fig.1 (a), and even within the same level, the blur of two images can be distinct (blue dotted box in Fig.1 (a)). Others resort to synthesized images to alleviate this problem. LIVE [11] and TID2008 [12] synthesize large amount of distorted images from un-distorted images and the blur level is determined by different blur kernels. However there is a large domain gap between synthesized images and real ones, which may lead to severe performance degradation in real-world applications.

For more effective learning strategy, limited by current datasets, most existing methods [8, 9, 13] treat quality assessment as a classification task or directly regress the true quality levels. These methods have limited performance in real applications due to the aforementioned reasons. Some researchers [14, 15] convert it as a learn-to-rank problem according to their quality levels or based on synthesized images. However they are only based on pairwise relationships between the original and corresponding synthesized degraded images, which limits the representation ability of the learned features.

To solve the above limitations, we firstly proposed to label the object image qualities with pairwise ranks, as labeling which image is sharper than the other is much more easier than labeling blur levels. Therefore, we construct a new object-BA dataset (named FIB) by hiring crowdsourcing annotators to judge the blur ranking of all image pairs from two popular face datasets [16, 17]. In this way, we can get the total ranking of all images as shown in Fig.1 (b). The ranking labels are all from real images rather than the synthesized images [11, 12], which would benefit the applications in real world. Based on this well-labeled dataset, to get fine-grained quality scores, we propose to learn the scores only regularized by the ranking relationships. To further improve the performance, we propose a semi-supervised learning framework to leverage the large-scale unlabeled data [18, 19] (named as FIB-unlabeled). The framework is shown in Fig.2, which

consists of a supervised learning branch as mentioned above and a self-supervised learning branch. In the self-supervised learning branch, we propose to construct a quadruplet instead of an ordinary pair as previous methods [14, 15] to generate rank relationships in a self-supervised manner. In a quadruplet, two images are blurred to generate two corresponding synthesized images using a same blur kernel. Therefore, the rank relationships of the two synthesized images is consistent with the original ones as shown in Fig.1 (d). While in an ordinary pair, only one image is blurred and the rank relationship which can be leveraged is that the rank of the blurred image is lower than the original image as shown in Fig.1 (c). The two images in a quadruplet are different from each other, containing different contents, backgrounds, illumination, occlusion etc, while the only difference in an original pair is the blur degree. Therefore, learning with quadruplets can make the learned feature more robust against various changes, which can lead to better performance for real-world applications.

Our contributions are as follows: (1) We define a new object image blur assessment problem, which are valuable for interpretable object image quality assessment, and establish a large-scale realistic dataset using pairwise ranks for reliable labeling to assess the object image blurriness which would benefit the research of object image quality assessment in real-life applications. (2) We propose a new method to get absolute fine-grained blur scores only based on pairwise ranks, and a new semi-supervised framework based on quadruplet inputs which can leverage large amounts of unlabeled images more efficiently for better performance.

2. DATASET CONSTRUCTION

Different from previous datasets which label images with limited quality levels or use synthesized images, we aim to construct a large-scale realistic dataset with reliable label for fine-grained object image blurriness assessment. In this paper, we focus on face image blurriness assessment. To enrich the dataset, we sample images of different blur levels ran-
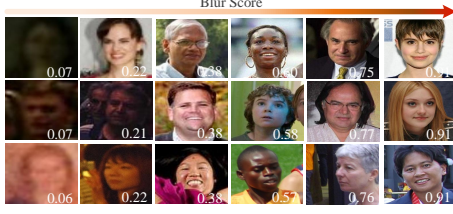
Fig. 3. Examples of prediction scores on Test1(first row), Test2(second row) and Test3(third row). Blur scores are written in white color under image and sorted from low to high.
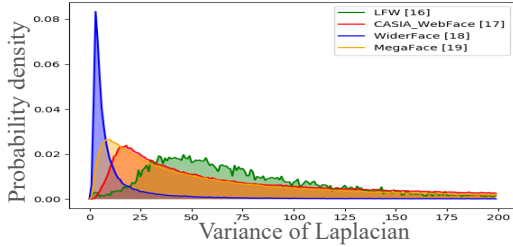

Fig. 4. Blur distributions of different face datasets. We use Laplacian operator to filter face images and take the variance of filtered images as their blur level roughly..

domly from LFW [16] and CASIA-WebFace [17] to construct our supervised training dataset FIB. The detailed labeling steps are as follows. Firstly, faces are cropped and resized to 256×256 resolution. Secondly, image pairs are randomly sampled for labeling. Due to the huge labeling cost for labeling all the image pairs, we propose to label randomly sampled image pairs to cover more images. Thirdly, image pairs are labeled by three annotators and the label is determined by the majority voting strategy. Specifically, letting x_1^i, x_2^i denote the two images in an image pair and y_1^i, y_2^i denote their corresponding blur scores, the relative ranking label of an image pair $p(x_1^i, x_2^i)$ is labeled as $\delta(y_1^i, y_2^i) = -1$ when the majority of annotators agree x_1^i is blurrier than x_2^i, and $\delta(y_1^i, y_2^i) = +1$ otherwise. In this way, we get a dataset with 10,000 labeled image pair, which covers 13,751 images and is named FIB-10K. To further investigate the effect of the label number, we randomly sample half of the labeled image pairs, and construct a sub-dataset named FIB-5K.

Although we construct the largest realistic dataset compared with other quality assessment datasets, the image number of the dataset is still limited. To alleviate this, we introduce a large-scale unlabeled data. To simulate the real scenarios where unlabeled data may come from a datset with different distribution from the labeled one, we randomly crop faces from WiderFace [18] and construct a unlabeled dataset FIB-unlabeled which covers 86,065 images.

To fully evaluate the effectiveness of the quality assessment algorithms, we construct three test datasets for evaluating them under different settings. The first one is the intra-dataset testing, in which test data is sampled exclusively from

Table 1. Statistics of training datasets FIB-labeled and FIB-unlabeled and testing datasets Test1, Test2 and Test3.

Dataset	Source	Labeled pairs No.	Images No.
FIB-5k	[16, 17]	5,000	8,093
FIB-10k	[16, 17]	10,000	13,751
FIB-unlabeled	[18]	-	86,065
Test1	[16, 17]	4,950	100
Test2	[18]	4,950	100
Test3	[19]	4,950	100

LFW [16] and CASIA-WebFace [17] as the labeled data. The second one is the half-inter-dataset testing, in which test data is sampled exclusively from WiderFace [18] as the unlabeled data. The third one is the inter-dataset testing, in which test data is sampled from a new dataset MegaFace [19] (the blur distributions of these datasets are shown in Fig.4). We denote the above datasets as Test1, Test2 and Test3 respectively. The labeling steps are similar as the supervised training dataset FIB. The only difference is that, we label all the image pairs to get the total ranking order in each dataset. Due to the huge labeling cost, each of the test datasets only has 100 images, however the number of the labeled pairs is 4,950. Some examples of testing datasets are shown in Fig.3. The statistics of our datasets are shown in Tab.1.

3. PROPOSED APPROACH

The framework is shown in Fig.2. It consists of a supervised learning branch which leverages the pairwise labeled data to predict blur scores, and a self-supervised branch which leverages the unlabeled data to improve the generalization performance. The details are described in following subsections.

3.1. Supervised Learning Based on Pairwise Ranking

The supervised learning branch takes an image pair $p(x_1^i, x_2^i)$ with relative ranking label $\delta(y_1^i, y_2^i)$ as input and predicts blur scores $\hat{y}_1^i, \hat{y}_2^i$. Each sub-network of this branch consists of a backbone network, followed by dense layers and finally a sigmoid to get the blur scores. As we don't have the true blur scores, we propose to use a pairwise margin ranking loss to regularize the learned blur scores, which is defined as:

$$L_{ranking} = \frac{1}{N} \sum_{i=1}^{N} max(0, d(\hat{y}_1^i, \hat{y}_2^i) * \delta(y_1^i, y_2^i) + \varepsilon_{ranking}) \quad (1)$$

where $\varepsilon_{ranking}$ is a controllable ranking margin; N is the number of image pairs $p(x_1^i, x_2^i)$; $d(\hat{y}_1^i, \hat{y}_2^i)$ denotes the distance of blur scores which is $d(\hat{y}_1^i, \hat{y}_2^i) = (-1) * (\hat{y}_1^i - \hat{y}_2^i)$.

During testing, we use one of the sub-network to predict the blur score of the input image. Some predicted results are shown in Fig.3, from which we can see our method can predict reasonable blur scores with only relative ranking labels.

Table 2. Experimental results.

Trainset	Method	Test1	Test2	Test3
FIB-10K	Baseline	0.9408	0.9643	0.9258
	Ours_RankIQA [14]	0.9286	0.9704	0.9275
	Ours_LSEP [22]	0.9274	0.9605	0.9210
	Ours	**0.9516**	**0.9734**	**0.9430**
FIB-5K	Baseline	0.9285	0.9572	0.9091
	Ours_RankIQA [14]	0.9240	0.9633	0.9125
	Ours_LSEP [22]	0.9180	0.9534	0.9005
	Ours	**0.9480**	**0.9723**	**0.9330**

3.2. Self-supervised learning Based on Quadruplet Ranking Consistency

Based on an observation that the relative ranking relationship of two images would not change after they have been subjected to the same image distortion attack, we propose to use a quadruplet instead of an ordinary pair to generate rank relationships in a self-supervised manner due to the reasons analyzed in the introduction section. Specifically, let x_{1d}^j, x_{2d}^j denote the corresponding blurred images of x_1^j, x_2^j with the same blur kernel. Then the relative ranking label $\delta(\hat{y}_{1d}^j, \hat{y}_{2d}^j)$ of x_{1d}^j, x_{2d}^j should be the same as the one $\delta(\hat{y}_1^j, \hat{y}_2^j)$ of x_1^j, x_2^j, and $x_1^j, x_2^j, x_{1d}^j, x_{2d}^j$ form a quadruplet. The self-supervised learning branch takes an image quadruplet as input, and we propose a Quadruplet Ranking Consistency (QRC) loss for training which is defined as:

$$L_{QRC} = \frac{1}{M} \sum_{j=1}^{M} max(0, d(\hat{y}_{1d}^j, \hat{y}_{2d}^j) * \delta(\hat{y}_1^j, \hat{y}_2^j) + \varepsilon_{QRC}). \quad (2)$$

where ε_{QRC} is a controllable ranking margin; M is the number of image quadruplets.

4. EXPERIMENTS

4.1. Implementation Details

We choose ResNet-18 [20] as our network backbone for balancing performance and efficiency. The input images are center-cropped from original images, and resized to 224×224. At training stage, batch size is set to 60, and SGD is used to optimize the model with weight decay 0.0005 and moment 0.9. Learning rate is initialized from 0.001 and scheduled by a CosineAnnealingLR policy [21]. We use *Spearman Rank Order Correlation Coefficient (SROCC)* as our evaluation protocol like others [14, 22, 15].

4.2. Experimental Results

Since previous methods are designed to regress the quality scores limited by existing datasets, while our dataset is labeled with pairwise rank relationships without the true quality scores. Therefore, directly comparing our method with them

on our dataset is impossible. We compare our methods with different variations to show the effectiveness of the proposed semi-supervised framework and the quadruplet based ranking consistency loss. We denote baseline as our method with only the supervised learning branch. We denote ours-RankIQA and ours-LSEP as our method with the quadruplet based ranking consistency loss in the self-supervised branch replaced by the rank loss proposed in RankIQA [14] and LSEP [22]. It is worth noting that both of the two rank losses use the pairwise rank relationships, i.e. the rank of the blurred image should be lower than the original image.

The experimental results are shown in Tab.2. It can be seen that Ours_RankIQA and Ours_LSEP achieve slightly worse performance on Test1 and comparable performance on Test2 and Test3. It shows that only leveraging the pairwise rank relationships in the self-supervised branch would hinder the performance under intra-dataset testing and would not improve the performance under half-inter-dataset testing and inter-dataset testing, since the image pairs only include blur changes which is easy to be learned by the network. Since the proposed quadruplet can introduce various changes during the feature learning, our method outperforms the others significantly under all of the three test settings, which verifies the effectiveness the proposed semi-supervised framework and the quadruplet based ranking consistency loss. Moreover, comparing the performances with FIB-5K and FIB-10K as the supervised dataset, the performance of other methods drop obviously with less labeled data, while the performance of our method is comparable under intra-dataset and half-inter-dataset testings, and only slightly worse under inter-dataset testing with only half of the labeled data. It shows that our method can alleviate the lack of labeled data by effectively leveraging the unlabeled data.

Furthermore, we illustrate the blur scores learned by our method in Fig.3, from which we can see our method can predict reasonable and discriminate score for blur assessment, although learned with only pairwise rank labels.

5. CONCLUSION

In this paper, we define a new object image blur assessment problem, and establish a large-scale realistic dataset with pairwise rank labels. Based on this dataset, we proposed a new semi-supervised learning method which consists of a supervised learning branch and a self-supervised learning branch. The supervised learning branch can learn to predict the fine-grained blur scores with only pairwise rank labels. The self-supervised learning branch can leverage the unlabeled data to improve the performance. To achieve this goal, a quadruplet ranking consistency loss is proposed to make the learned feature more robust against various changes. Extensive experiments under intra-dataset, half-inter-dataset and inter-dataset settings show the effectiveness the proposed semi-supervised framework and the quadruplet based ranking consistency loss.

6. REFERENCES

[1] Y. Taigman, Y. Ming, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[3] S. Zherzdev and A. Gruzdev, "Lprnet: License plate recognition via deep neural networks," 2018.

[4] J. Lu, B. Zou, Z. Cheng, S. Pu, S. Zhou, Y. Niu, and F. Wu, "Object-qa: Towards high reliable object quality assessment," 2020.

[5] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou, "Magface: A universal representation for face recognition and quality assessment," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Rushuai Liu and Weijun Tan, "Eqface: A simple explicit quality network for face recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[7] Luhong Liang, Jianhua Chen, Siwei Ma, Debin Zhao, and Wen Gao, "A no-reference perceptual blur metric using histogram of gradient profile sharpness," in *IEEE International Conference on Image Processing*, 2010.

[8] M. Fan, H. Rui, F. Wei, and J. Sun, "Image blur classification and blur usefulness assessment," in *IEEE International Conference on Multimedia and Expo Workshops*, 2017.

[9] R. Huang, W. Feng, M. Fan, L. Wan, and J. Sun, "Multiscale blur detection by learning discriminative deep features," *Neurocomputing*, pp. 154–166, 2018.

[10] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, pp. 1–1, 2017.

[11] H. R. Sheikh, "Live image quality assessment database," *http://live.ece.utexas.edu/research/quality*, 2003.

[12] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, and F. Battisti, "Tid2008 - a database for evaluation of full-reference visual quality assessment metrics," *adv modern radioelectron*, 2009.

[13] H. Heng, H. Ye, and R. Huang, "Defocus blur detection by fusing multiscale deep features with conv-lstm," *IEEE Access*, 2020.

[14] X. Liu, Jvd Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," *IEEE Computer Society*, 2017.

[15] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans Image Process*, pp. 3951–3964, 2017.

[16] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," *Month*, 2008.

[17] Y. Dong, L. Zhen, S. Liao, and S. Z. Li, "Learning face representation from scratch," *Computer Science*, 2014.

[18] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Adam Paszke, Sam Gross, and et.al. Massa, "Pytorch: An imperative style, high-performance deep learning library," in *NIPS*, pp. 8024–8035. 2019.

[22] Luojun Lin, Lingyu Liang, and Lianwen Jin, "Regression guided by relative ranking using convolutional neural network (r3cnn) for facial beauty prediction," *IEEE Transactions on Affective Computing*, 2019.