Learning from Mixed Datasets: A Monotonic Image Quality Assessment Model

Zhaopeng Feng,1 Keyang Zhang,2 Shuyue Jia,3 Baoliang Chen,1 and Shiqi Wang3

1School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, China
2Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR
3Department of Computer Science, City University of Hong Kong, Hong Kong SAR

Email: shiqiwang@cityu.edu.hk

Deep learning based image quality assessment (IQA) models usually learn to predict image quality from a single dataset, leading the model to overfit specific scenes. To account for this, mixed datasets training can be an effective way to enhance the generalization capability of the model. However, it is nontrivial to combine different IQA datasets, as their quality evaluation criteria, score ranges, view conditions, as well as subjects are usually not shared during the image quality annotation. In this paper, instead of aligning the annotations, we propose a monotonic neural network for IQA model learning with different datasets combined. In particular, our model consists of a dataset-shared quality regressor and several dataset-specific quality transformers. The quality regressor aims to obtain the perceptual qualities of each dataset while each quality transformer maps the perceptual qualities to the corresponding dataset annotations with their monotonicity maintained. The experimental results verify the effectiveness of the proposed learning strategy and our code is available at https://github.com/fzp0424/MonotonicIQA.

Introduction: Image quality assessment (IQA) model which aims to provide image quality objectively has been widely used as a key component in various computer vision tasks [1]. Popular IQA models can be classified into full-reference IQA (FR-IQA) models and no-reference (NR-IQA) models, depending on whether the reference image is available. FR-IQA models can be dated back to the 1970s when the signal fidelity measures are adopted for quality prediction, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [2], and Multi-scale SSIM (MS-SSIM) [3]. Compared with FR-IQA, NR-IQA is regarded as a more challenging task due to the lack of reference information. Typical NR-IQA models rely on the natural scene statistics (NSS) construction [4-6] with the assumption that the image quality can be estimated by measuring the destruction level of the NSS. In recent years, many pioneers resort to deep learning technology for both FR-IQA [7-11] and NR-IQA [8, 12-15], endowing the IQA models with a strong capability for quality-aware feature extraction. In particular, a shallow ConvNet was adopted for patch-based NR-IQA learning in [12]. Both the quality estimation and distortion identification were learned in a multi-task manner for NR-IQA in [16]. The rank information between image pairs was utilized in [17-19] which was able to enrich the training data efficiently.

Generally speaking, the performance of deep learning based methods heavily relies on the training data. An ideal training set should contain sufficiently diverse image contents, plenty of distortion types, and reliable annotations. However, it is an extremely laborious task to collect and construct such a dataset, leading to the constrained size of existing IQA datasets. The lack of training data usually brings the over-fitting problem and leads to the data-driven methods with poor generalization capability. An efficient way to enrich the training data is to combine the existing datasets together for training. As shown in Fig. 1, the image contents and distortion types in different IQA datasets can be diverse, implying the promising potential for generalized IQA model learning when they are mixed together. Despite the advantages, the quality annotations of existing datasets are usually not fully aligned. As shown in Fig. 1, the mean opinion scores (MOSs) range of the CSIQ dataset [20] is from 0 to 1 while the range is [0, 100] in the LIVE [21] dataset. Herein, one can re-scale the MOSs of each dataset non-linearity for training, while the re-scaling function is difficult to be acquired reliably. To account for this, Zhang et al. proposed to use the pairwise rank information in each dataset for mixed datasets training [22]. However, the number of rank pairs will increase explosively with the image number, causing the model convergence for a long-time. In [23], continual learning was adopted for diverse datasets learning while the catastrophic forgetting problem may not be completely avoided. Ma et al. proposed an incremental learning scheme for cross-task NR-IQA [24]. However, the pruning operator potentially causes performance degradation on the old task. In this paper, we learn the model from all the datasets simultaneously, achieving a better performance balance among different datasets.

Inspired by the four-parameter logistic function, Li et al. proposed a non-linear mapping layer to map the relative quality to the perceptual quality in [25], while the fitting capability of the non-linear function might be questioned. In this paper, we develop a new paradigm for mixed dataset training by introducing a monotonic neural network. In particular, we first adopt a dataset-shared network for perceptual quality regression of each image and subsequently map the perceptual quality monotonously to the annotated MOSs with the quality ranking in each dataset. The proposed training strategy enjoys two desired advantages. First, it gets rid of the laborious quality re-alignment to successfully combine different datasets for training. Second, the learnable monotonic neural network possesses a powerful non-linear fitting capability to map the image perceptual quality to the annotations. We evaluate our method on six IQA datasets that extensively cover both synthetic distortions and realistic distortions. Compared with state-of-the-art NR-IQA models, our method presents superior performance for image quality prediction, revealing the effectiveness of the proposed monotonic neural network.

Proposed Method: The overview framework of our method is presented in Fig. 2. As shown in Fig. 2, our method contains two main components: 1) the dataset-shared quality regressor which aims to obtain the perceptual quality of each image in the mixed datasets. 2) the dataset-specific quality transformer which monotonously maps the perceptual quality to the MOSs annotated in each dataset. The details of each component are elaborated.

Quality Regressor: For the perceptual quality regression, we first adopt the pre-trained ResNet-34 [27] to extract the quality-aware feature of each image. Then we flatten the spatial dimensions of the feature representation after the last convolution to obtain the feature \( \mathbf{F} \in \mathbb{R}^{c \times d} \), where \( c \) is the channel number and \( f \) is the spatial length. Analogous to [22], we perform the bilinear pooling on \( \mathbf{F} \), aiming to acquire the interactions between different channels,

\[
\mathbf{F} = \mathbf{F} \times \mathbf{F}^T, \\
\]  

where \( \mathbf{F} \in \mathbb{R}^{c \times c} \). Based upon \( \mathbf{F} \), two fully connected layers (each followed by a ReLU activation) with a final fully connected layer (the output
put dimension is one) are appended to utilize the $\overline{F}$ to the perceptual quality $Q^p$. The quality regressor is shared by different datasets. **Quality Transformer.** The connection between the perceptual quality $Q^p$ of each image and its MOS $Q^m$ is that their rank should be the same within the dataset. Following this principle, we propose a monotonic neural network to map the $Q^p$ to the $Q^m$. The architecture of our monotonic neural network is shown in Fig. 3. It consists of five constrained fully connected layers (CFCL) whose weights are all constrained to be positive. The positive weight ensures the gradient of the mapping function to be positive.a.k.a. higher $Q^p$ will be mapped to the higher $Q^m$ and vice versa. In addition, to maintain the network monotonicity and enhance the transformer non-linear fitting ability, we adopt the ELU [28] as the activation function. We provide a proof of the monotonicity of the quality transformer at Adopt the ELU [28] as the activation function. We provide a proof of the monotonicity of the quality transformer at

\[
\text{ReLU}(x) = \begin{cases} 
0 & \text{if } x < 0 \\
-x & \text{otherwise}
\end{cases}
\]

\[
\text{ELU}(x) = \begin{cases} 
0 & \text{if } x < 0 \\
x - \alpha e^{x} - 1 & \text{otherwise}
\end{cases}
\]

where $\alpha$ is a hyper-parameter and $\mu$ and $\sigma$ are the mean and standard deviation of the predicted quality scores and MOSs in a training batch. Our model is trained in an end-to-end manner.

**Experiments:** To verify the effectiveness of our method, six IQA databases are used and mixed for training, including LIVE [21], CSIQ [20], KADID-10k [30], BID [31], LIVE Challenge [32] and KonIQ-10k [26]. The synthetic distortions are involved in the first three datasets and the last three datasets contain the realistic distortions. We randomly split each database into three subsets for training, validation, and testing. There is no content overlap among the three subsets. The split proportions are 60%, 20%, and 20%, respectively. We apply Spearman’s rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) to evaluate the monotonicity and accuracy. The larger SRCC and PLCC indicate better quality prediction results. We linearly re-scale the MOS ranges of each dataset to $[0, 10]$. We implement our method by PyTorch [33] with Adam [34] as the optimizer. The learning rates of the quality regressor and quality transformer are $3 \times 10^{-5}$ and $3 \times 10^{-4}$, respectively. The learning rates of the pre-trained ResNet-34 and the rest of the model are initialized to $3 \times 10^{-5}$ and $3 \times 10^{-4}$, respectively. In the training phase, we set the batch size by 32 and the $\lambda$ in Eqn. (2) is 1.0. We first re-scale the short length of images in the training sets and test sets to 512 and 768 respectively, with their aspect ratios maintained. Then patches with a size of $384 \times 384 \times 3$ are cropped from each training sample for data augmentation during training.

**Performance Comparisons.** We compare our method with six state-of-the-art NR-IQA models, including NIQE [35], ILNIQE [36], PQR [37], DB-CNN [38], Koncept512 [26] and UNIQUE [22]. The linear re-scaling strategy proposed in [22] is also included and denoted as UNIQUE-Linear [22] in Table 1. The median SRCC and PLCC of ten splitting results are reported. For better comparison, the image number weighted SRCC and PLCC of all six datasets are also presented. As shown in Table 1, we can observe that the NSS-based methods including NIQE [35] and ILNIQE [36] achieve better performance on synthetic datasets while are less effective on realistic datasets. The performance of deep learning based methods is determined by the training set. For example, when the DB-CNN [38] is trained on the CSIQ [20] dataset, higher SRCC can be achieved on the LIVE [21] and KADID-10k [30] datasets than it on the LIVE Challenge [32], BID [31], and KonIQ-10k [26] datasets. On the contrary, the performance on synthetic datasets will drop dramatically when the KonIQ-10k [26] is used for training. Compared with the models that are trained on a single dataset, significant performance improvement can be observed by UNIQUE [22] and our method, revealing the effectiveness of the mixed dataset training. Moreover, our method achieves the best performance in terms of both the weighted SRCC and PLCC, demonstrating the superiority of the monotonic neural network based learning strategy when compared with the linear re-scaling strategy (UNIQUE-Linear) or the rank-based strategy (UNIQUE).

As shown in Table 2, we further conduct extra experiments to study the best number of CFCL in our proposed monotonic neural network. From Table 2, we observe the best layer number is five. Less or more layers will cause the performance to drop to some extent. The reason may lie in that fewer layers will bring a negative effect on the non-linear fitting ability and more layers will cause the overfitting problem. To verify our model efficiency, we further compare the floating point operations (FLOPs) that our method consumed during inference with the FLOPs of the other three deep-learning based methods, including the Koncept512, DB-CNN, and UNIQUE. The results show that our method is comparable with the UNIQUE (3.95G FLOPS) and presents a significant gain compared with DB-CNN (16.53G) and Koncept512 (13.22G). It reveals that our method achieves better quality prediction performance without much sacrifice of model complexity.

**Conclusion:** In this paper, we concentrate on the NR-IQA model learning from mixed datasets. To avoid the laborious quality annotation alignment, a monotonic neural network is proposed to map the regression quality to the MOS values. Experimental results demonstrate that the proposed method is more effective than both the unreliable linear re-scaling strategy and pair rank strategy. We believe our method is able to shed the light on the exploration of more generalized IQA models and inspire more works for learning monotonic neural networks.

**References**

1. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2018).
2. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2018).
Table 1. The median SRCC and PLCC of ten splitting results. The databases used for training models are included in the bracket.

| Database      | LIVE | CSIQ | KADID-10k | BD | LIVE Challenge | KonIQ-10k | Weighted | LIVE | CSIQ | KADID-10k | BD | LIVE Challenge | KonIQ-10k | Weighted |
|---------------|------|------|-----------|----|----------------|-----------|----------|------|------|-----------|----|----------------|-----------|----------|
| NIQE          | 0.906| 0.632| 0.374     | 0.448| 0.464         | 0.521     | 0.471    | 0.908| 0.726| 0.434     | 0.541| 0.515          | 0.528     | 0.563    |
| ILNIQE        | 0.907| 0.632| 0.531     | 0.516| 0.469         | 0.507     | 0.528    | 0.912| 0.673| 0.573     | 0.533| 0.576          | 0.574     | 0.576    |
| FLR3 (PRO)    | 0.663| 0.522| 0.121     | 0.481| 0.414         | 0.414     | 0.414    | 0.673| 0.622| 0.461     | 0.633| 0.560          | 0.633     | 0.633    |
| PQR (KADID-10k)| 0.916| 0.803| 0.385     | 0.419| 0.405         | 0.827     | 0.970    | 0.845| 0.505| 0.501     | 0.492| 0.492          | 0.492     | 0.492    |
| DB-CN (CSIQ)  | 0.851| 0.651| 0.379     | 0.451| 0.499         | 0.851     | 0.361    | 0.851| 0.485| 0.363     | 0.742| 0.515          | 0.515     | 0.515    |
| DB-CN (LIVE Challenge) | 0.723| 0.491| 0.690       | 0.488 | 0.700 | 0.754 | 0.685 | 0.539 | 0.832 | 0.825 |
| KonIQ-10k (KADID-10k) | 0.776| 0.439| 0.640     | 0.781| 0.672         | 0.927     | 0.880    | 0.917| 0.663| 0.568     | 0.818| 0.864          | 0.852     | 0.727    |
| UNIQUE-Uni (All databases) | 0.935| 0.921| 0.870     | 0.889| 0.809         | 0.896     | 0.909    | 0.905| 0.906| 0.927     | 0.976| 0.940          | 0.940     | 0.940    |
| UNIQUE (All databases) | 0.909| 0.901| 0.870       | 0.658 | 0.854 | 0.486 | 0.906 | 0.927 | 0.876 | 0.937 | 0.900 | 0.901          | 0.892     | 0.892    |
| Proposed (All databases) | 0.908| 0.908| 0.887     | 0.847| 0.904         | 0.902     | 0.902    | 0.902| 0.902| 0.926     | 0.958| 0.958          | 0.958     | 0.958    |

Table 2. The median SRCC of ten splitting results using different structure of the proposed monotonic neural network.

| Model                     | LIVE | CSIQ | KADID-10k | BD | LIVE Challenge | KonIQ-10k | Weighted |
|---------------------------|------|------|-----------|----|----------------|-----------|----------|
| Original CFCLs and ELUs   | 0.863| 0.502| 0.496     | 0.512| 0.507         | 0.903     | 0.957    |
| Three CFCLs and Two ELUs  | 0.958| 0.903| 0.882     | 0.853| 0.854         | 0.902     | 0.902    |
| Five CFCLs and Four ELUs (Proposed) | 0.958| 0.903| 0.882     | 0.853| 0.854         | 0.902     | 0.902    |
| Seven CFCLs and Six ELUs  | 0.942| 0.932| 0.934     | 0.868| 0.857         | 0.902     | 0.902    |

with unsupervised domain adaptation. IEEE Transactions on Image Processing 30, 5463–5476 (2021)

20. Larson, E.C., Chandlar, D.M.: Most apparent distortion: Full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging 19(1), 101106:1–101106:19 (2010)

21. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing 15(11), 3440–3451 (2006)

22. Zhang, W., et al.: Uncertainty-aware blind image quality assessment in the laboratory and wild. IEEE Transactions on Image Processing 30, 3474–3486 (2021)

23. Zhang, W., et al.: Continual learning for blind image quality assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

24. Ma, R., et al.: Remember and reuse: Cross-task blind image quality assessment via relevance-aware incremental learning. In: ACM International Conference on Multimedia, pp. 5248–5256. (2021)

25. Li, D., Jiang, T.; Jiang, M.: Unified quality assessment of in-the-wild videos with mixed datasets training. International Journal of Computer Vision 129(4), 1238–1257 (2021)

26. Hou, V., et al.: KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing 29, 4041–4056

27. He, K., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. (2016)

28. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)

29. Li, D., Jiang, T., Jiang, M.: Norm-in-norm loss with faster convergence and better performance for image quality assessment. In: ACM International Conference on Multimedia, pp. 789–797. (2020)

30. Lin, H., Hou, V., Sapoe, D.: KADID-10k: A large-scale artificially distorted IQA database. In: IEEE International Conference on Quality of Multimedia Experience, pp. 1–3. (2019)

31. Ciancio, A., et al.: No-reference blur assessment of digital pictures based on multifeature classifiers. IEEE Transactions on Image Processing 20(1), 64–75 (2010)

32. Gadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing 25(1), 372–387 (2015)

33. Paszke, A., et al.: PyTorch: An imperative style, high-performance deep learning library. In: Neural Information Processing Systems, pp. 8026–8037. (2019)

34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations, pp. 1–15. (2015)

35. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. IEEE Signal Processing Letters 20(3), 209–212 (2012)

36. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing 24(8), 2579–2591 (2015)

37. Zeng, H., Zhang, L., Bovik, A.C.: Blind image quality assessment with a probabilistic quality representation. In: IEEE International Conference on Image Processing, pp. 609–613. (2018)

38. Zhang, W., et al.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology 30(1), 36–47 (2020)