A Soft-ranked Index Fusion Framework with Saliency Weighting for Image Quality Assessment

Liangwei Yu*, Zhao Wang*, Yan Ye*, Lingyu Zhu[†], Shiqi Wang[†]
*Alibaba Group [†]City University of Hong Kong
Beijing, 100102, China Hong Kong, China

Abstract

The compression technique is widely adopted for efficient data storage and transmission. Accurate image quality assessment (IQA) measures are urgently desired to evaluate the compression performance. To obtain a more robust evaluation, we propose a soft-ranked index fusion framework for the perceptual preference prediction task, with a combination of different quality measures. The derived soft-ranked indices are fully leveraged to provide the strong discriminability of ranking information. Furthermore, a saliency weighting approach is utilized to investigate the impact of visual attention on our framework. Experimental results indicate that our method achieves a promising prediction accuracy compared with the state-of-the-art quality measures.

1. Introduction

Image and video compression play the fundamental role due to the large volume of image and video data acquired, transmitted and stored. Central to the image and video compression is achieving a good trade-off between bit-rate and distortion. Considering that subjective quality assessment is strenuous and inconvenient, objective quality assessment is essential for practical use.

The last decade has witnessed the boom in objective quality assessment based on the five principles: error visibility, structure similarity, information-theoretic, learning-based, and fusion-based methods. Error visibility reflects the pixel level error, e.g., mean squared error (MSE) and peak signal-to-noise ratio (PSNR). Structural similarity (SSIM) [1] and its variants (e.g., MS-SSIM [2] and FSIM [3]) consider image degradation as a perceived change in structural information. The prototypical information-theoretic method is VIF [4], which exploits natural scene statistics and the notion of image information extracted by the human visual system. More recently, learning-based measures (e.g., LPIPS [5] and DISTS [6]) utilize the convolutional neural network (CNN) to transform the reference and distorted images to a high-dimensional representation and learn the quality from distortion-aware features. Video Multimethod Fusion Approach (VMAF) [7] is formulated by Netflix to estimate the quality by computing multiple QA measures and fusing them using the machine learning technique.

There are many "in-capture" artifacts before compression, such as blur, noise, underexposure. These imperfect images are sub-sequentially compressed by different algorithms, further introducing distortions. To estimate the perceptual preference of compressed images, it is challenging to rely on a single quality measure. Since different quality measures have their own characteristics, the measurements may have excellent performance on specific distortion types but perform poorly on more complicated distortions. Inspired by previous work [7,8], we propose a soft-ranked index fusion framework for the perceptual preference prediction task, combining different quality measures. The intuition behind our proposed method is to leverage characteristics from different quality measures that could complement each other.

Given the fusion strategy, we take advantage of the ranking information provided by multi-measurements to boost the estimation performance, which is substantial to perform preference prediction. In particular, the predicted scores are first converted to score differences according to the elaborately selected image pairs. Then, the score differences are adaptively rescaled according to the image contents and processed using the Bradley-Terry model [9] to generate soft rank indices, which could indicate the rank relationship of these image pairs. Besides, as analyzed in [10] that saliency information is beneficial for further improvement of IQA tasks, we further integrate saliency weighting into our framework to produce a robust prediction.

In this paper, we take initial steps towards quality assessment for learned image compression based on fusion strategy. Our main contributions are three-fold as follows:

1). A fusion-based quality assessment framework is proposed for the perceptual preference prediction task. Nine quality measures including an enhanced measurement are chosen for fusion.

2). Preference ranking information and saliency information are further utilized to boost prediction accuracy.

3). Our fusion-based framework outperforms other quality measures, and achieves a promising prediction accuracy.

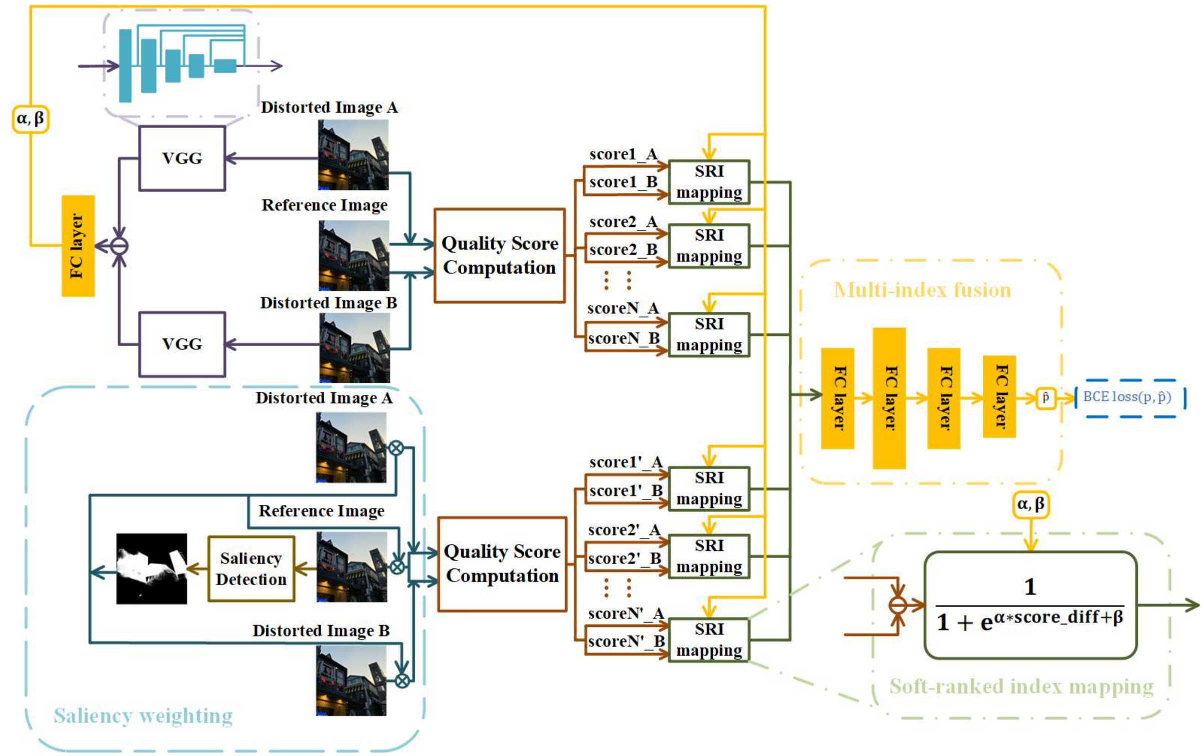

Figure 1. The framework of our proposed soft-ranked index fusion with saliency weighting.

2. Proposed Model

As shown in Figure 1, our proposed fusion-based framework contains three parts: a soft-ranked index (SRI) mapping module, a saliency weighting module and a multi-index fusion module. SRI mapping module maps predicted scores of different quality measures into soft-ranked indexes, indicating distorted image pairs' preference ranking relationship. The saliency weighting module predicts a saliency mask, weights the input images, and feeds the weighted images into multiple quality measures. The index fusion module fuses a series of soft-ranked indices and predicts the probability of preferring image B.

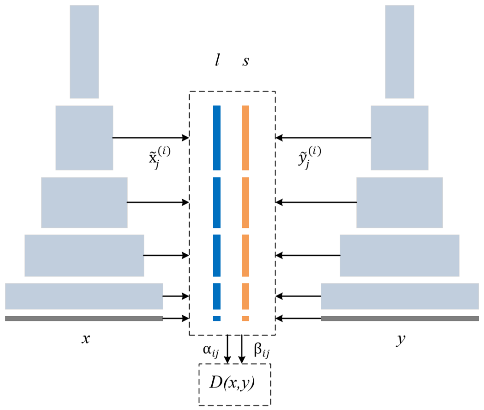

Figure 2. Structure of DISTS_resnet.

2.1. Enhanced DISTS

DISTS [6] has built-in tolerances for texture resampling and texture similarity, providing good perceptual evaluation predictions. Due to its effectiveness, DISTS and variant measures were also studied in the CLIC2021 competition [11].

Original DISTS utilizes VGG [12] as its backbone model and concatenates the multi-stages' output features for further evaluation. In this paper, we further exploit the influence of different backbone models. The evaluation is performed on the CLIC-V dataset which will be additionally introduced in Section 3.1, and the evaluation results are listed in Table I. As can be seen, DISTS with ResNet50 [13] as its backbone model can achieve higher

Table I. The accuracy of DISTS with different backbones on CLIC-V dataset.

Backbone model	Accuracy
VGG16 [12]	0.749
ResNet50 [13]	**0.777**
ResNet101 [13]	0.775
WideResNet [14]	0.774
EfficientNet-b0 [15]	0.661
EfficientNet-b7 [15]	0.736
DenseNet [16]	0.748

accuracy in our experiment. Therefore, we modify the original DISTS model to DISTS_resnet and subsequently integrate the enhanced DISTS_renset into our soft-ranked index fusion framework.

2.2. SRI mapping

In MMFN [8], adaptively rescaled scores from different quality measures are fused to produce a final score. Then a score2prob layer is integrated to learn the rank ability for the preference prediction task. However, the utilized score2prob layer can only learn the rank ability from two fused scores, while the ranking relationship predicted by multiple quality measures is neglected. In this paper, an SRI mapping module is proposed to fully exploit the rank information from numerous quality measures. As shown in the right-bottom of Fig. 1, we first calculate the score differences of image pairs from different quality measures. Then the score differences are adaptively rescaled according to the distorted-aware features from the VGG network. Finally, the rescaled score differences are mapped to soft-ranked indexes by the Bradley-Terry model and fed into the index fusion model to predict the quality. The mapping function can be formulated as follows:

$$soft\ rank\ index = \frac{1}{1 + e^{\alpha * score\_diff + \beta}}, \qquad (1)$$

where *score_diff* indicates the score differences of image pairs generated by different quality measures. α and β are the weights and biases estimated from the distortion-aware features, which are extracted from the pre-trained VGG network.

2.3. Saliency weighting

Following [10], a saliency weighting module is utilized in our framework to boost prediction accuracy. We use a pre-trained salient object detection network TRACER [17] to predict a saliency mask. Then, the saliency mask is unified and followed by a morphological closing operation to capture more complete regions. Finally, the reference image and the distorted image pairs are weighted by the saliency mask and fed into multiple quality measures to produce a series of soft-ranked indexes with saliency information involved.

2.4. Multi-index fusion

To perform a more stable prediction, nine common quality measures with different characteristics have been chosen in our index fusion module, including PSNR, SSIM [1], MS-SSIM [2], GMSD [18], FSIM [3], VIF [4], VSI [19], LPIPS [5] and DISTS_resnet. It is worth mentioning that DISTS_resnet utilizes pretrained Resnet50 instead of VGG as the backbone.

In addition to the nine soft-ranked indexes generated from these chosen quality measures, another corresponding nine soft-ranked indices generated with saliency weighting are also utilized for final fusion. The multi-index fusion module comprises four fully connected layers, and produces a probability of preferring image B.

2.5. Loss function

Two loss functions have been adopted since the enhanced DISTS_resnet and the soft-ranked index fusion framework are trained separately,

In analogous to the loss function mentioned in [6], we adopt a mixed loss function to optimize the DISTS_resnet network:

$$loss_{DISTS\_resnet} = \frac{1}{N}\left(\widehat{D}(ref, A) - D(ref, A)\right)^2 + \lambda\frac{1}{N}D(B_1, B_2)^2, \qquad (2)$$

where $\widehat{D}(*,*)$ and $D(*,*)$ represent prediction scores and ground-truth quality score respectively; *ref* and *A* represent the reference image and distorted image, respectively; B_1 and B_2 denote randomly cropped image patches from the same distorted image; *N* stands for the mini-batches, and λ governs the trade-off between the two terms, which is set as 1 in our experiment.

To train the whole multi-index fusion framework, binary cross entropy loss is utilized to guide the optimization process:

$$loss_{fusion} = -\frac{1}{N}\sum_{i=1}^{N}\left[p_i log\widehat{p_i} + (1 - p_i)\, log(1 - \widehat{p_i})\right], (3)$$

where p_i and $\widehat{p_i}$ represent the ground-truth preference and predicted perceptual probability.

3. Experiment

3.1. Datasets

In this paper, an enhanced DISTS_resnet and a soft-ranked index fusion framework have been introduced, which are trained with different tasks. DISTS_resnet is trained to predict a perceptual score between the reference image and the distorted image. At the same time, the soft-ranked index fusion framework aims at predicting the probability of preferring image B over image A. In our experiment, three datasets are involved while training our whole framework.

PIPAL: PIPAL [20] dataset contains 200 reference, 40 distortion types and 23,000 distorted images. In addition to classical noise types, PIPAL includes learning-based noises. The MOS value is provided for each distorted image. In addition to PIPAL, other two datasets have also been chosen to train DISTS_resnet for evaluation, and the experimental results are shown in Table II. As can be seen, DISTS_resnet trained on PIPAL performs the best. Since the MOS values in PIPAL dataset are derived from a

Table II. Accuracy of DISTS_resnet trained on different dataset.

dataset	Accuracy on CLIC-V
KADID [21]	0.767
LICQA [22]	0.770
PIPAL [20]	**0.777**

Table III. Accuracy of different quality measures on CLIC-V.

Quality measure	Accuracy
PSNR	0.572
SSIM	0.627
MS-SSIM	0.612
VIF	0.554
FSIM	0.640
GMSD	0.649
VSI	0.635
LPIPS	0.736
DISTS	0.749
DISTS_resnet	0.777
ours	**0.792**

ranking model, a hidden ranking relationship between these distorted images can be further exploited during training, which greatly benefits our task.

CLIC_T: It is the training set provided by the CLIC2022 competition [12]. There is a total of 122,107 records, including images pairs and preference labels. Since the records are not filtered, which contain many noise labels, the records with different labels or records that none of our selected nine indexes can predict correctly are removed. In total, 96036 pairs are chosen as our training set, and 24,009 pairs are selected as our validation set.

CLIC_V: It is the validation set provided by the CLIC2022 competition [12]. There is a total of 5,220 images pairs with preference labels in this database. We examine the final performance of our framework on this dataset.

3.2. Implementation details

Our framework is trained on the Pytorch framework with NVIDIA V100 GPUs. In the training process, we set the mini-batch size to 64, and choose the Adam optimizer with an initial learning rate of 0.0001 to optimize our model. Firstly, the enhanced DISTS_resnet is trained on the PIPAL dataset. The training procedure finishes when no more accuracy improvement is observed on our validation set. Secondly, the soft-ranked index fusion framework is trained on the CLIC-T dataset. Finally, the performance of our framework is tested on the CLIC-V dataset.

Table IV. Performance of ablation studies, including our final framework, our method without SRI module, and our method without saliency weighting module.

Model	Accuracy
ours w/o SRI module	0.776
Ours w/o saliency weighting module	0.788
ours	**0.792**

3.3. Results

We compare the performance of our framework and other quality measures on the CLIC-V dataset in Table III. As can see in Table III, our framework outperforms other quality measures.

3.4. Ablation study

In this subsection, we conduct ablation experiments to verify the effectiveness of critical modules. “Ours without the SRI module” means that the rescaled predicted scores are directly fed into the fusion module without ranking information. “Ours without the saliency weighting module” means that only nine soft-ranked indices between original references and distorted images are provided in the fusion module, without utilizing saliency information. The experimental result is shown in Table IV.

As shown in Table III, without ranking information being exploited, the fusion framework cannot precisely learn the rank relationship for the distorted image pairs, leading to a poor performance. Similarly, without the saliency weighting module, the framework loses the guidance of saliency prior knowledge and has a relatively lower performance than our final framework.

4. Conclusion

In this paper, we propose a soft-ranked index fusion framework with saliency weighting for the preference prediction task. Instead of directly fusing scores of multiple quality measures, score differences of image pairs are mapped into soft-ranked indexes in our designed SRI module to fully exploit the preference ranking information, which is more substantial for the perceptual preference prediction task. Besides, a saliency weighting module is utilized in our framework, which investigates the impact of visual attention into our method. The proposed framework investigates the perceptual information from both signal and feature domain, which improves the effectiveness and robustness of quality assessment. Experimental results on the CLIC-V dataset demonstrates the superiority of our method. We may continue the study on more effective quality measure fusion approaches and more elegant integration methods to introduce saliency information into IQA

References

[1] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on image processing*, 13(4):600–612, 2004.

[2] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. *In Proceedings of the Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, 2003.

[3] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378– 2386, 2011.

[4] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing, 15(2):430–444*, 2006.

[5] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *In Proceedings of Computer Vision and Pattern Recognition*, 2018.

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. CoRR, abs/2004.07728, 2020.

[7] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

[8] Peng Y, Xu J, Luo Z, et al. Multi-Metric Fusion Network for Image Quality Assessment. *In Proceedings of Computer Vision and Pattern Recognition,* 1857-1860, 2021.

[9] Hunter D R. MM algorithms for generalized Bradley-Terry models. *The annals of statistics*, 32(1): 384-406, 2004.

[10] Zhang, Wei, et al. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE transactions on neural networks and learning systems,* 27(6): 1266-1278, 2015.

[11] The workshop and Challenge on Learned Image Compression. *http://compression.cc/.*

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *In Proceedings of the International Conference on Learning Representations*, 2015.

[13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *In Proceedings of Computer Vision and Pattern Recognition,* 770-778, 2016.

[14] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv*:1605.07146, 2016.

[15] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *In Proceedings of the International Conference on Machine Learning*, 6105-6114, 2019.

[16] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. *In Proceedings of Computer Vision and Pattern Recognition*, 4700-4708, 2017.

[17] Lee, Min Seok and Shin, WooSeok and Han, Sung Won. TRACER: Extreme Attention Guided Salient Object Tracing Network. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[18] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014.

[19] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[20] Gu J, Cai H, Chen H, et al. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. *arXiv preprint arXiv*:2007.12142, 2020.

[21] Lin H, Hosu V, Saupe D. KADID-10k: A large-scale artificially distorted IQA database. *In Proceedings of the Conference on Quality of Multimedia Experience*, 2019.

[22] Yang Li, Shiqi Wang, et al. Quality Assessment of End-to-End Learned Image Compression: The Benchmark and Objective Measure. *In Proceedings of the 29th ACM International Conference on Multimedia*, 2021.