Multi-Metric Fusion Network for Image Quality Assessment

Yanding Peng, Jiahua Xu, Ziyuan Luo, Wei Zhou, Zhibo Chen
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China
{pyd, xujiahua, luozi01, weichou} @mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

With the fast proliferation of multimedia applications, the reliable prediction of image/video quality is urgently needed. Many quality assessment metrics have been proposed in the past decades with various complexity and consistency with human ratings. The metrics are designed from different aspects, e.g., pixel level fidelity, structural similarity, information theory and data-driven. In this paper, we design a Multi-Metric Fusion Network (MMFN) for aggregating the quality scores predicted by diverse metrics to generate more accurate results. To be specific, we utilize the image features extracted from the pretrained network to adaptively rescale the predicted quality from different metrics, and leverage the fully-connected layers to regress a single scalar as the final score. Pairwise images can be further integrated into the training procedure by adding a Score2Prob layer. Experimental results on the validation and test sets demonstrate that our proposed MMFN achieves better prediction accuracy compared with other metrics.

1. Introduction

Nowadays, multimedia data has been applied in various applications, e.g., entertainment, education, medical examination and electronic retailing, which are significant sources for acquiring information in everyday life [10]. High quality images/videos can promise the integrity and accuracy of the perceived visual information. However, more or less distortions are inevitably introduced during the processing chain [12], e.g., acquisition, compression, transmission and reconstruction as shown in Figure 1. Thus, to guarantee the quality of experience for end users, image/video quality assessment (IQA/VQA) plays a crucial role to guide the current image processing and video coding systems. The quality assessment can be roughly divided into two categories according to human engagement, namely subjective quality assessment and objective quality assessment [8]. Subjective quality assessment can provide the most accurate quality

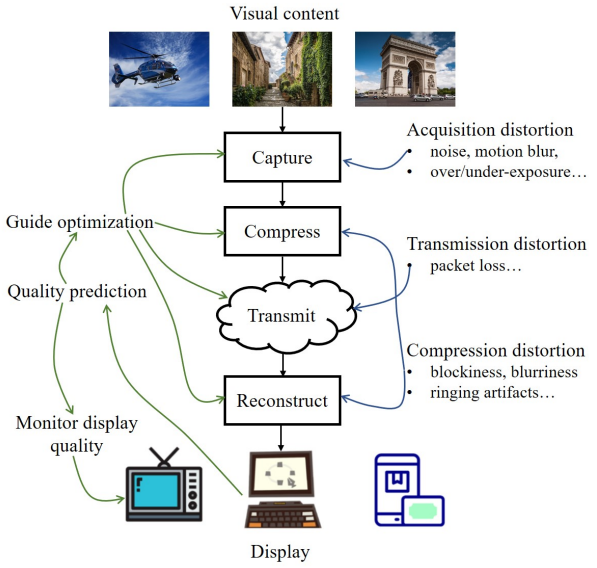

Figure 1. Distortions introduced during the processing chain, and typical usages of quality assessment metrics.

labels since human being is the final receiver. But the subjective experiment is labor-intensive and time-consuming, which is unsuitable for real-time scenarios. Therefore, objective quality assessment models are deeply researched to achieve automation.

To precisely evaluate the perceptual quality of images, the full-reference (FR) IQA methods can be broadly classified into five categories [1], namely error visibility, structure similarity, information-theoretic, learning-based and fusion-based methods. Error visibility methods measure the pixel level error and the representative is mean squared error (MSE). Structure similarity (SSIM) [12] methods consider the human vision system (HVS) and utilize the local structure similarity to evaluate image quality. SSIM and its variants (e.g. MS-SSIM [14], IW-SSIM[13], FSIM [17]) show better correlation with human perception than simple error visibility methods. Information-theoretic methods measure the mutual information between the reference and distorted images. The prototypical example is the VIF mea-

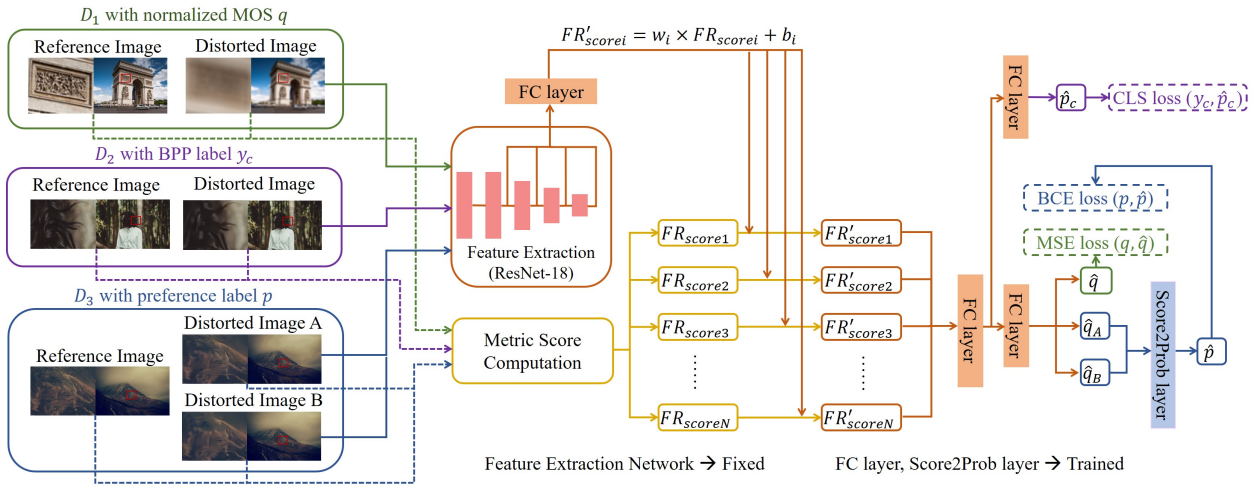

Figure 2. Pipeline of Muiti-Metric Fusion Network (MMFN), D_1, D_2 and D_3 are databases with different quality labels.

sure [9]. Learning-based methods benefit from leveraging images with human ratings to train deep neural networks (DNN) [18]. DNNs can automatically extract distortion-aware features for quality regression with supervised learning [2]. Fusion-based methods aim to combine existing IQA metrics to form a better model by considering the diversity and complementarity of different metrics. The famous VMAF [5] is one example of fusion-based metrics.

Although numerous metrics have been proposed in the last decades, there is no single quality measure that significantly outperforms others [7, 5]. Different metrics have their own characteristics, thus the metrics may have excellent performance on one distortion type but perform poorly on another. When the image content is diverse and the distortion is complex, relying on a single metric to give accurate quality predictions is a challenging task. As a result, the fusion-based methods can complement existing works to build a better general-purpose evaluator. Moreover, we can further boost the performance by incorporating newly proposed powerful models.

Existing fusion-based methods determine the weights for each metric either empirically or learned from data [1]. They neglect that the weights may change under different image contents and distortions. Therefore, we propose the Multi-Metric Fusion Network (MMFN) that adopts the image content and distortion related features extracted from distorted images to adaptively rescale the quality scores given by typical quality metrics. Then, the weights are adjusted according to the content and distortion related features. Besides, we can incorporate training data in the format of pairwise preference via adding a Score2Prob layer. Experimental results on the validation and test sets demonstrate that our proposed MMFN can outperform other quality metrics in terms of accuracy.

2. Proposed Model

2.1. Adaptive Rescaling

As different image contents and distortions may influence the performances of IQA metrics, we resort to use image content and distortion related features to guide the fusion process of different metrics. Considering the powerful feature extraction capabilities of convolutional neural networks, we utilize a ResNet-18 [3] backbone (pretrained on 2D IQA database KonIQ [6]) to extract the content-related and distortion-aware features from the input distorted image. Note that the features adopted in MMFN are extracted from different spatial scales, e.g., global average pooling for features of various basic blocks. Then, these features are sent through a fully-connected layer to generate the weights and biases for rescaling. The predicted scores of FR metrics will be adaptively rescaled as follows:

$$FR'_{scorei} = w_i \times FR_{scorei} + b_i, \tag{1}$$

where FR_{scorei} and FR'_{scorei} denote the i-th metric before and after adaptive rescaling. w_i and b_i are the weight and bias for the i-th metric rescaling. With this process, the model obtains the ability to adjust the weights of metrics according to the image content and distortion, which will further exploit the strengths of different metrics in the scene that they are good at.

2.2. Regression and classification

In MMFN, we adopt ten representative metrics due to their different mechanisms, high quality prediction accuracy and differentiability, including PSNR, SSIM [12], MS-SSIM [14], GMSD [15], FSIM [17], VSI [16], NLPD [11], VIF [9], LPIPS [18], DISTS [2]. All five categories except fusion-based metrics are contained to improve the expression ability of the regression input. After the rescaling

process, the ten rescaled metrics scores are sent to the final regression network to get a single scalar as our perceptual quality score.

To cater with the training on databases with preference labels, we use a Socre2Prob layer to predict perceptual judgement from the pair of two distorted images. The Score2Prob layer is composed of three fully connected layers, which accepts five inputs (q_A, q_B, $q_A - q_B$, $q_A/(q_B+\epsilon)$, $q_B/(q_A+\epsilon)$), and gives a probability of preferring image B. q_A and q_B refer to the predicted scores of image A and image B, ϵ is a smooth constant.

The classification task is also introduced to increase the generalization performance of MMFN. Therefore, we add the side way of a fully-connected layer for our model to predict the bit per pixel (BPP) class of the distorted image.

2.3. Loss function

While different databases have different data formats, we need to design suitable loss functions to train our model. Figure 2 shows the pipeline of our MMFN. Three kinds of labels are considered in our work, namely the MOS value, the preference label, and the BPP classification type. For the MOS value, we adopt MSE loss formulated as Eq. 2 to optimize our model:

$$loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i)^2, \tag{2}$$

where q_i and $\hat{q}_i$ refer to the ground-truth label and the predicted score of the i-th image in a mini-batch, N denotes the batch size. For the preference label, we regard it as a classification task and use binary cross entropy loss formulated as Eq. 3 to guide the optimization process:

$$loss_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[p_i \log \hat{p}_i + (1-p_i)\log(1-\hat{p}_i)], \tag{3}$$

where p_i and $\hat{p}_i$ represent the ground-truth and predicted probability for preferring image B over image A of the i-th pair in the mini-batch. For the BPP classification label, we use cross entropy loss to optimize our model:

$$loss_{CLS} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} y_{ic} \log \hat{p}_{ic}, \tag{4}$$

where y_{ic} and $\hat{p}_{ic}$ denote the class label and the predicted probability of i-th sample being class c (0.075BPP, 0.15BPP and 0.3BPP).

When training our MMFN on different databases, we can apply different loss functions or combine them with the trade-off coefficients λ_1 and λ_2:

$$loss = loss_{MSE} + \lambda_1 loss_{BCE} + \lambda_2 loss_{CLS}. \tag{5}$$

We can adjust the coefficients λ_1 and λ_2 according to the importance or effectiveness of different databases to our task.

3. Experiment

3.1. Databases

Four different databases are used in our experiment, named PIPAL [4], BAPPS [18], CLIC-V, and Sub-T.

PIPAL [4]: It contains 200 reference, 40 distortion types and 23,000 distortion images. Especially, it includes the outputs of GAN-based algorithms as typical distortion types, which are beneficial to our task. The MOS value is provided for each distorted image.

BAPPS [18]: It contains 6 coarse distortion classes, namely traditional distortions, CNN-based distortions, super resolution artifacts, frame interpolation artifacts, video deblurring artifacts, and colorization artifacts. It provides preference labels of 161,000 patch pairs, and each pair includes two distorted images and one reference image.

CLIC-V: It is the validation set provided by the CLIC2021 competition. There are totally 5,220 images pairs with preference labels in this database.

Sub-T: It is a database built by ourselves. 626 reference images provided by CLIC compression track are compressed using seven methods under three different bit rates. Therefore, there are totally 13,146 distorted images with BPP labels (Sub-T-BPP). Meanwhile, we select 408 typical pairs for the subjective experiment, and each pair originates from the same reference image and bite rate (Sub-T-Prefer). For the 408 pairs, 10 subjects are asked to judge which is more similar to the reference image.

Among all these databases, PIPAL, BAPPS and Sub-T-BPP are used for training MMFN, CLIC-V is used for validation and Sub-T-Prefer is used for testing.

3.2. Implementation details

The MMFN is implemented based on Pytorch framework with a NVIDIA 1080Ti GPU. In the training process, we set the mini-batch size as 64, and choose the Adam optimizer with an initial learning rate as 0.01 to optimize our model, The learning rate will decay by a factor valued 0.5 every 100 epochs. Firstly, the model is trained on PIPAL to match the predicted score into the range [0,1]. Secondly, the BAPPS and Sub-T-BPP database is further introduced to train the Score2Prob layer and finetune the entire model. Here we set the trade-off coefficient λ as 0.01 when we jointly train MMFN on these two databases. Finally, the model is validated on CLIC-V and tested on Sub-T-Prefer.

3.3. Results

We compare the performance of MMFN and other FR metrics in Table 1. Except for MMFN, we also train an-

Table 1. Accuracy of different metrics on CLIC-V and Sub-T-Prefer databases.

Metric	CLIC-V	Sub-T-Prefer
PSNR	0.573	0.409
SSIM [12]	0.571	0.449
MS-SSIM [14]	0.614	0.341
GMSD [15]	0.647	0.623
FSIM [17]	0.640	0.466
VSI [16]	0.627	0.449
NLPD [11]	0.591	0.380
VIF [9]	0.605	0.404
LPIPS [18]	0.744	0.799
DISTS [2]	0.756	0.689
MMFN-FC	0.771	0.781
MMFN	0.795	0.787

other model which simply fuses ten metrics with the fully-connected layer (MMFN-FC) to study the importance of adaptive rescaling. As we can see in Table 1, MMFN outperforms other metrics which proves the superiority of our method. Besides, the performance of MMFN is higher than MMFN-FC, which verifies the effectiveness of adaptive rescaling.

4. Conclusion

In this paper, we propose the Multi-Metric Fusion Network as a full reference image quality assessment model which can give the absolute perceptual quality and preference judgement between two images. We fuse ten metrics to build a more accurate and robust model. To guide the fusion process, we introduce the adaptive rescaling operation by utilizing a ResNet-18 backbone for extracting the content-related and distortion-aware features from the input distorted image. Moreover, we deal with multi-databases training problem by designing different loss functions and introducing a Score2Prob layer. Experimental results on the CLIC-V and Sub-T databases demonstrate the superiority of our method. Feature level fusion and distortion specific analysis will be done to continue improving the power and interpretability of the model in our future work.

References

[1] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of image quality models for optimization of image processing systems. *arXiv*, 2020. 1, 2

[2] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. 2, 4

[3] 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*, pages 770–778, 2016. 2

[4] Haoyu Chen Xiaoxing Ye Jimmy Ren Chao Dong Jinjin Gu, Haoming Cai. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision (ECCV) 2020*, pages 633–651, Cham, 2020. Springer International Publishing. 3

[5] 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. 2

[6] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Koniq-10k: Towards an ecologically valid and large-scale iqa database. *arXiv preprint arXiv:1803.08489*, 2018. 2

[7] Tsung Jung Liu, Weisi Lin, and C. C. Jay Kuo. Image quality assessment using multi-method fusion. *IEEE Transactions on Image Processing*, 22(5):1793–1807, 2012. 2

[8] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010. 1

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

[10] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 1

[11] L. Valero, B. Johannes, B. Alexander, and E. P. Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016(16):1–6, 2016. 2, 4

[12] 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. 1, 2, 4

[13] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2010. 1

[14] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 1, 2, 4

[15] 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. 2, 4

[16] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, Oct 2014. 2, 4

[17] 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. 1, 2, 4

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 3, 4