Focused Feature Differentiation Network for Image Quality Assessment

Gang He[1,2], Yong Wang[1,*], Li Xu[1], Wenli Zhang[1], Ming Sun[2], Xing Wen[2]
[1] Xidian University, Xi'an, China
[2] Kuaishou Technology, Beijing, China

Abstract

Image quality assessment (IQA) intended to assess the perceptual quality of images has been an essential problem in both human and machine vision. Recently, with the help of deep neural network (DNN), IQA algorithms can extract more valuable differences between the distorted and reference images than the traditional algorithms, and thus the performance of DNN-based algorithms is more satisfactory than that of previous algorithms. However, the accuracy for different distorted images preference rating of the existing DNN-based quality assessment methods will be decreased when multiple distorted images are quite similar to each other or to the reference image. To tackle this problem, we propose a focused feature differentiation network (FFDN) to highlight the feature maps with greater distorted and reference differentiation. Furthermore, we use the multi-scale feature fusion module to fuse the focused differentiation features at different scale receptive fields. To further improve the accuracy of our method, we predict the mean opinion score and differentiation score by stages and combine them with different self-learning weights. Finally, we convert the weighted score into different image preference degrees. Experimental results on the validation dataset of CLIC2022 and test dataset of CLIC2021 show that the accuracy of our model FFDN is higher than other excellent quality assessment methods.

1. Introduction

In recent years, due to the rapid development of multimedia, people increasingly rely on images to obtain information. Image quality assessment (IQA) plays a vital role [12] in various scenarios owing to the existence of quality degradations in various stages of image acquisition, compression, transmission and display. Image quality assessment can be divided into subjective quality assessment and objective quality assessment [6]. Subjective quality evaluation is the intuitive evaluation of image quality by observers and pays more attention to people's instinctive feelings.

According to different scenarios, objective quality assessment can be divided into full-reference methods (FR-IQA), reduced-reference methods (NN-IQA) and no-reference methods(NR-IQA) [5]. For FR-IQA, many classic methods, such as MSE [9], SSIM [9] and MS-SSIM [10], are widely used in many fields. Inspired by them, FSIM [14], SR-SIM [13], and GMSD [11] are proposed. These traditional methods require manual extraction of the distorted image and the reference image features difference. FR-IQA methods based on deep learning have better assessment than traditional methods. Zhang et al. [15] proposed LPIPS to evaluate deep features across different architectures and tasks and compare them with classic metrics. Importantly, they found that features extracted from deep architectures outperform hand-crafted features. Ding et al. [2] proposed DISTS method by using the structure and texture similarity of shallow and deep feature maps extracted from reference and distorted images, which had an excellent assessment effect on common IQA datasets.

However, when multiple distorted images are quite similar to each other or to the reference image, the assessment effect of these methods is relatively decreased. Moreover, these methods lack quality preference for comparing different similar distorted images. In view of these existing problems, we propose a focused feature differentiation network (FFDN) for image quality assessment, which focuses on the feature maps with greater distorted and reference differentiation, and fuses multiple scores into different image preference degrees.

FFDN obtains preference degree for different images by learning image quality scores from common datasets and converting multiple scores with different weights into probabilities. One score is the mean opinion score (MOS) predicted by DISTS [2] training on the common IQA datasets such as LIVE [7] and KADID-10k [4], and another is the preference score of focused differential network training on datasets with preference labels. We introduce the channel attention module to make the feature maps focus on greater distorted and reference differentiation. Moreover, we devise the multi-scale differentiation feature fusion module to

*Corresponding author.

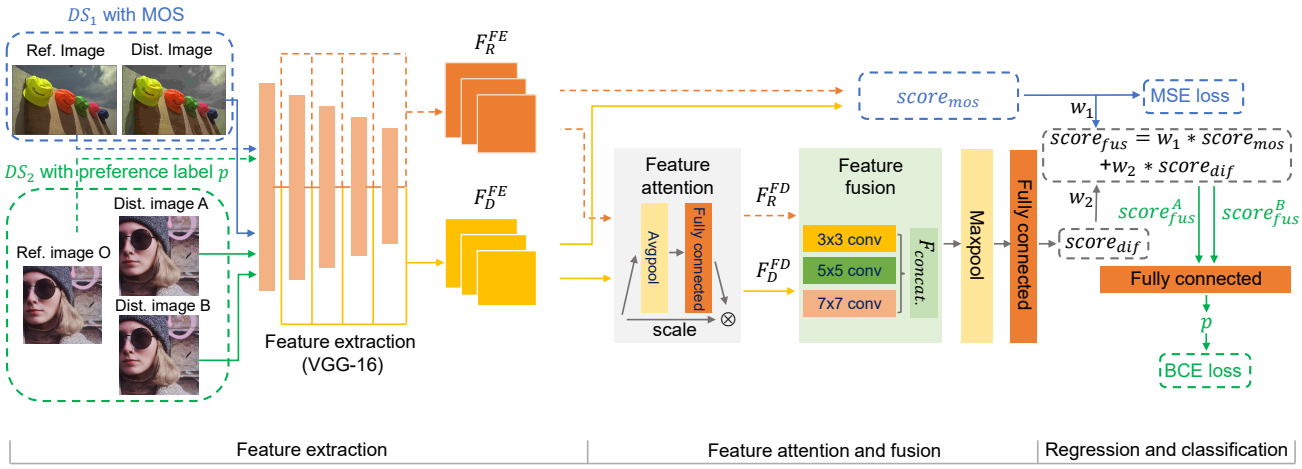

Figure 1. The architecture of the proposed approach focused feature differentiation network (FFDN). Dist. image A and B with different degrees of corruptions are quite similar with each other. DS_1 and DS_2 are the datasets with different quality labels. F_R^{FE} and F_D^{FE} are the reference image feature maps and distorted image feature maps output by VGG-16. F_R^{FD} and F_D^{FD} are focused differential reference image features and distorted image features respectively. $score_{mos}$ is the MOS score learned by DS_1 in the first phase and $score_{dif}$ is the differentiation score of the distorted image and reference image. $score_{fus}$ is the fused score learned by DS_2 in the second phase. $score_{fus}^A$ and $score_{fus}^B$ is the quality score of distorted image A and image B with reference image O. p is the preference degree of distorted images.

fuse the features of the focused differentiation distortion and reference feature maps under different scale receptive fields and purify the differentiation features. Finally, we assign each score a weight that can be self-learned by the differences between distorted image features and reference image features. Experimental results on the validation and test sets being provided by CLIC 2022, demonstrate that our proposed method FFDN can outperform DISTS and other quality assessment method in terms of accuracy.

2. Proposed architecture

2.1. Feature extraction

Considering the excellent performance of DISTS [2] which has been empirically proven sensitive to structural distortions and robust to texture substitutions [1], we use it as a fraction of FFDN to learn the MOS score. We use VGG-16 as the backbone to extract shallow and deep features of images. First, the reference images and distorted images are input into VGG-16 respectively. Then, the 64, 128, 256, 512, 512 reference and distorted feature maps are obtained from the five phases of the network. DISTS defines separate quality measurements for the texture $l(F_{R_j}^{(i)}, F_{D_j}^{(i)})$ (using the global means) and the structure $s(F_{R_j}^{(i)}, F_{D_j}^{(i)})$ (using the global correlations) of each pair of corresponding feature maps, where $F_{R_j}^{(i)}$ and $F_{D_j}^{(i)}$ are the i-th reference image feature map and distorted image feature map in the j-th phase of VGG-16. Finally, DISTS combines the quality measurements from different convolution layers using a weighted sum to get the distorted image MOS $score_{mos}$:

$$score_{mos} = 1 - \sum_{i=0}^{m} \sum_{j=1}^{n_i} (\alpha_{ij} * l(F_{R_j}^{(i)}, F_{D_j}^{(i)}) + \beta_{ij} * s(F_{R_j}^{(i)}, F_{D_j}^{(i)})), \quad (1)$$

where m = 5 is the number of convolution layers, n_i is the number of feature maps in the i-th convolution layer, and α_{ij}, β_{ij} are positive learnable weights, satisfying $\sum_{i=0}^{m} \sum_{j=1}^{n_i} (\alpha_{ij} + \beta_{ij}) = 1$. We obtain image quality MOS score in the first phase of FFDN.

2.2. Feature attention

As we described before, previous algorithms that concentrated on predicting image quality MOS of distorted images with different degradation ignored subtle differential features information, as Figure 2 (a) and (b) show. When they are used to assess those distorted indistinguishable images, the accuracy of the existing DNN-based quality assessment methods will decrease, because of the reduction of the extracted differential features. The differential features mean that the feature map contains more information that reflects the difference between the distorted image and reference image. To compensate for the loss of accuracy of $score_{mos}$, we use channel attention and multi-scale feature fusion to maximize the extraction of differentiated features in the second phase of FFDN. VGG-16 extracts a large

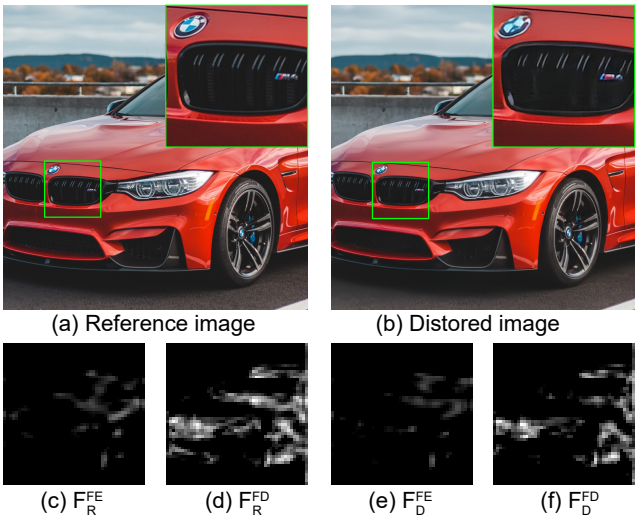

Figure 2. A visualization of the differential features. (c) F_R^{FE} and (e) F_D^{FE} are the reference image feature maps and distorted image feature maps by VGG-16. (d) F_R^{FD} and (f) F_D^{FD} are focused differential features of reference image and distorted image features respectively. Compared with (c) and (e), the focused differential features (d) and (f) highlight more valuable information that can reflect the difference between the distorted and reference images.

number of shallow and deep feature maps of reference and distorted images. At the feature attention stage, inspired by SENet [3], we first need to squeeze features by carrying out global average pooling of these feature maps to obtain features of 1×1×C size, C denotes the total number of feature maps. Then, we assign the weights to the 1×1×C features by using fully connected layers to get the excited features. Finally, the scale is multiplied by the input feature maps and features after weight distribution. We obtain focused differential reference image features F_R^{FD} and distorted image features F_D^{FD}, as shown in Figure 2 (d) and (f) , which contain more differential information than (c) and (e).

2.3. Feature fusion

We need fuse reference image features F_D^{FD} and distorted image features F_R^{FD} in order to maximize the use of differentiated features. Hence, we utilize multi-scale differentiation feature fusion (MDFF) module, which is composed of three convolution layers with difference receptive fields (3×3, 5×5, 7×7) and a concatenation layer. MDFF can perform feature fusion on focused differentiation distorted feature maps and reference feature maps under different scale receptive fields and extract differentiation features. Then, the image quality differentiation $score_{dif}$ is obtained through putting the fused features into the maximum pooling layer and the fully connected layer.

2.4. Regression and classification

To avoid the limitations of single method's quality score, our final image quality score $score_{fus}$ which will eventually be used during the predict time, consists of $score_{mos}$, and $score_{dif}$ in the second phase of FFDN, assigned different weights as follows:

$$score_{fus} = w_1 * score_{mos} + w_2 * score_{dif} \quad (2)$$

where w_1 and w_2 perform automatic optimization learning by means of backward gradient propagation according to the differences between different distorted images and reference images.

In order to obtain the preference for different distorted images, we convert the predicted $score_{fus}$ through a fully connected (FC) layer into the preference degree for a particular image. Distorted image A and image B yield $score_{fus}^A$ and $score_{fus}^B$ respectively, which are used for input to the FC layer with another two inputs $(score_{fus}^A/(score_{fus}^A + score_{fus}^B), score_{fus}^B/(score_{fus}^A + score_{fus}^B))$. Finally we end up with a preference degree p for distorted images in the second phase of FFDN. The larger p indicates a preference for distorted image B over image A.

2.5. Loss function

We use different datasets required different loss functions for different training phases. As shown in Figure 1 pipeline of our model FFDN, in the first phase of FFDN, we utilize LIVE [7] and KADID-10K [4] datasets with MOS to train the $score_{mos}$ by using mean squared error (MSE) loss function as the following equation:

$$loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i)^2 \quad (3)$$

where q_i and $\hat{q}_i$ refer to the predicted $score_{mos}$ and the ground-truth MOS label of the i-th image in a mini-batch, N denotes the batch size. For the datasets with preference label, we utilize binary cross entropy (BCE) loss formulated as following Eq. 4 to train the second phase of FFDN:

$$loss_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[p_i log\hat{p}_i + (1-p_i)log(1-\hat{p}_i)] \quad (4)$$

where p_i and $\hat{p}_i$ represent the predicted probability p and the ground-truth label for preferring distorted image B over image A of the i-th pair in the mini-batch.

The final loss function as following Eq. 5 combines $loss_{MSE}$ and $loss_{BCE}$ with the trade-off coefficients λ_1 and λ_2 to train FFDN with different datasets:

$$loss = \lambda_1 * loss_{MSE} + \lambda_2 * loss_{BCE} \quad (5)$$

We can adjust the coefficients λ_1 and λ_2 according to the significance of different datasets to adapt the challenge task.

3. Experiment

3.1. Datasets

We utilize four datasets named LIVE [7], KADID-10K [4], CLIC-V, and CLIC-T to train and test our model FFDN.

LIVE [7]: It contains 29 reference images and 779 distortion images, including JPEG2000, JPEG, white noise, Gaussian blur, and fast fading Rayleigh distortion.

KADID-10K [4]: It contains 81 reference images, each of which is distorted by 25 distortion types at 5 distortion levels and total 10,206 distorted images.

CLIC-V: It is the validation set provided by the CLIC2022 competition. It contains 5,220 pairs images, each pair containing a reference image and two distorted images with quality preference label.

CLIC-T: It is the test set provided by the CLIC2021 competition. It contains 122,107 pairs image data, which contain some error labels in the perceptual quality assessment. When it is used as the training set, we filter its data and remove the wrong label data.

3.2. Training and testing details

Our method FFDN is implemented based on Pytorch framework. We utilize LIVE [7] and KADID-10K [4] to train $score_{mos}$ and $score_{dif}$ with $loss_{MSE}$, and separately use CLIC-V and CLIC-T to train the preference score with $loss_{BCE}$. It should be noted that we swap CLIC-V and CLIC-T as testing sets and training sets. In the training phase, we set the mini-batch size as 8 and utilize the Adam optimizer with an initial learning rate as 0.001 to optimize our model FFDN. We totally train 50 epochs and the learning rate decay by a factor valued 0.1 every 20 epochs. We perform data enhancement during the training phase. During the test phase, we test FFDN model with ensemble ($\times 4$) testing, i.e., flipping in four ways (none, horizontally, vertically, both horizontally and vertically) and averaging these outputs to obtain robust higher accuracy.

3.3. Comparison with previous methods

We compare the performance of our model FFDN with some current excellent FR-IQA methods, such as PSNR, FSIM [14], GMSD [11], LPIPS [15], DISTS [2], and RADN [8] on CLIC-V and CLIC-T datasets. The experimental results are listed in the following Table 1. As we can see in Table 1, the accuracy of traditional methods PSNR, FSIM [14], and GMSD [11] based on manual feature extraction is far lower than that of deep learning methods LPIPS [15], DISTS [2], and RADN [8]. Compared with other deep learning methods, our method FFDN further focuses on image feature differentiation and can predict more accurate scores than other FR-IQA models on these datasets.

Table 1. Accuracy of different methods on CLIC-V and CLIC-T datasets. CLIC-T without any processing is used as a test set.

Methods	CLIC-V	CLIC-T
PSNR	0.572	0.507
SSIM [9]	0.571	-
FSIM [14]	0.640	-
GMSD [11]	0.647	-
LPIPS [15]	0.740	0.682
DISTS [2]	0.742	0.725
RADN [8]	0.741	0.710
FFDN (ours)	**0.762**	**0.744**

3.4. Ablation study

To further investigate the effectiveness of our proposed components, we conduct ablation studies on CLIC-V and CLIC-T datasets. As shown in Table 2, $score_{dif}$, which obtained more differential features by using attention, performs better than $score_{mos}$. Moreover, the accuracy of the fused $score_{fus}$ is higher than that of both $score_{mos}$ and $score_{dif}$. It demonstrates that both mean opinion scores and differentiation scores contribute significantly to our method. The accuracy is further improved by using preference classification after scores fusion.

Table 2. Ablation study on CLIC-V and CLIC-T datasets.

Components	CLIC-V	CLIC-T
$score_{mos}$	0.742	0.725
$score_{dif}$	0.753	0.734
$score_{fus}$	0.759	0.740
$score_{fus} + classification$	0.762	0.744

4. Conclusion

In this paper, we propose a full-reference image quality assessment approach called focused feature differentiation network (FFDN), which can give objective quality score and preference degree of images. Based on DISTS, we use VGG-16 as a backbone to extract image shallow and deep features. We utilize channel attention to focus our attention on the more differentiated feature maps between the distorted image and reference image, and use multi-scale convolution feature fusion module to perform feature fusion at different scale receptive fields and extract differentiated features. Finally, we fuse the two predicted scores with assigning different weights and convert them into preference degree for images. Experimental results on the CLIC-V and CLIC-T datasets demonstrate the superiority and higher accuracy of our method.

Acknowledgement. This work was supported by Key Laboratory of MIIT for Intelligent Products Testing and Reliability 2021 Key Laboratory Open Project Fund (No. CEPREI2022-01).

References

[1] Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. Locally adaptive structure and texture similarity for image quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2483–2491, 2021. 2

[2] K Ding, K Ma, S Wang, and EP Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 4

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[4] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 1, 3, 4

[5] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 1

[6] 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

[7] Hamid R Sheikh. Image and video quality assessment research at live. *http://live. ece. utexas. edu/research/quality*, 2003. 1, 3, 4

[8] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021. 4

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

[10] 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

[11] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013. 1, 4

[12] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63(11):1–52, 2020. 1

[13] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE international conference on image processing*, pages 1473–1476. IEEE, 2012. 1

[14] 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, 4

[15] 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 the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 4