Image Quality Assessment with Transformers and Multi-Metric Fusion Modules

Wei Jiang[1], Litian Li[1], Yi Ma[1], Yongqi Zhai[1], Zheng Yang[1], Ronggang Wang[1,2*]
[1]Shenzhen Graduate School, Peking University, China
[2]Peng Cheng Laboratory, Shenzhen, China
{jiangwei,lilitian,zhaiyongqi,zyang}@stu.pku.edu.cn {mayi}@pku.edu.cn {rgwang}@pkusz.edu.cn

Abstract

Image quality assessment is crucial for low-level vision tasks such as compression, super-resolution, denoising and etc. It guides researchers how to design networks, design loss functions, and decide the optimization direction of networks. A good quality assessment metric should comform to people's subjective feelings as much as possible. Traditional PSNR and MS-SSIM have more and more obvious shortcomings in quality evaluation with the popularity of GANs. Inspired by metrics such as LPIPS, IQT, etc., we decided to design a metric that is learned by the network itself. In this paper, we use a ConvNeXt-Tiny network to extract features and calculate nonlinear residuals between reference images and distorted images. We feed residuals into transformers to compare the degree of distortion. In addition, we use multi-metric fusion to improve the performance of our network. Our model achieves 0.780 accuracy on CLIC validation set. Our code is available at https://github.com/JiangWeibeta/IQA-TMFM.

1. Introduction

In the era of rapid development of multimedia technology, a large amount of image or video data is generated in our daily life. In order to reduce the storage cost and bandwidth brought by these data, many traditional and learning-based lossy compression methods have been proposed. However, the distortion introduced by these methods is difficult to measure, and obtaining the Mean Opinion Score (MOS) using manual methods is expensive and difficult to get immediate feedback in the production environment. Therefore, in order to meet the growing visual demands of the industry and people, we hope to find an accurate and efficient image quality assessment metric that is close to subjective quality assessment and can be easily employed in compression and other low-level vision tasks.

The IQA task aims to predict the subjective opinions of human viewers, and existing IQA methods are mainly divided into 3 categories: full-reference (FR), reduced-reference (RR), and no-reference (NR) IQA methods. NR models, such as NIQE[13], are very flexible in practical applications, but the absence of reference pictures makes it difficult to predict feelings of human raters. FR models mainly focuses on the differences in texture and structure between the reference image and distorted images, which are still widely used in various visual reconstruction tasks.

The most classic FR reference metrics are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM)[17], which focus on pixel differences and structural differences between two images, respectively. They are often selected as optimization targets due to their excellent performance on previous tasks and low computational complexity. In addition, fusion-based metrics like VMAF[19] are also popularized in video quality assessment.

However, with the continuous development of deep learning technology, especially the application of GANs[5] in image compression, restoration and other fields, the reconstructed images contain unrealistic generation artifacts, sharper edges and noise similar to real textures, which brings new challenge to the IQA task. Traditional metrics can't assess the quality of these images well. In this respect, deep learning-based perceptual quality assessment metrics[4][21] have better performance in the IQA task. Because of the strong expressive ability of the transformers, Cheon et al[2]. proposed to use a transformer[16] to deal with the fake pictures generated by GANs in the PIPAL dataset[7]. They used a transformer to fuse the features of reference images and distorted images, and finally obtained satisfactory performance.

In this paper, the CLIC perceptual quality assessment task is modeled as a binary classification task. Similar to IQT[3], we first use a pre-trained model to extract multi-scale features from the reference image and distorted images, and then use a transformer to obtain the probability of preference for distorted images. Finally, multiple metrics are fused to further improve our prediction accuracy.

*Corresponding author.

2. Method

Figure 1 provides an overview of our proposed method. In the following chapters, we will introduce the network architecture, Feature Extraction Modules, Nonlinear Residuals Modules, Transformer Blocks and Multi-Metric Fusion Module.

In order to match the training dataset given by CLIC, the input of the network is a reference image and its corresponding two distorted images. To reduce computational overhead, full-resolution images are cropped to several patches before fed into the neural network. We calculate the nonlinear residuals between the distorted patches and reference patches and concatenate them across channel. The calculation results on the feature maps of different resolutions are inputs of different transformer modules. The average value of the output of the transformer modules serves as our main reference metric and other reference metrics will be used as biases. Therefore, the method can be formulated as:

$$f_a, f_b, f_r = FE(i_a, i_b, i_r) \tag{1}$$

$$r_{a,r}, r_{b,r} = NR(f_a, f_r), NR(f_b, f_r) \tag{2}$$

$$P_A = MFM\{TF[(r_{a,r}||r_{b,r}), f_r], bias\} \tag{3}$$

FE means Feature Extraction Module, f_a, f_b and f_r are features extracted by FE. NR means Nonlinear Residuals Module, $r_{a,r}$ and $r_{b,r}$ are residuals between features of distorted images and features of reference images. TF is our transformer[16] block. $bias$ is other metrics. P_A means the probability of distorted image A better than distorted image B.

At first, we wanted to make our model like a discriminator, but it didn't work. Making it like a discriminator means that we directly input feature maps of distorted images in a batch and concatenate them with feature maps of reference images, using reference images as conditions. However, there is a situation where A has better quality than B, and B has better quality than C, when B and C is in a batch, the model should choose B, when A and B is in a batch, the model should choose A. Interactions within batches are not possible, and it may make our model confused.

2.1. Feature Extraction Module

Considering that when people assess image quality, they pay more attention to the semantic information of the image, and the effect of pixel differences on image quality is not so obvious, we plan to measure the quality of the image according to the semantic information which is extracted from a pre-trained model. Since image quality assessment is a very subjective task, we do not intend to use the encoder of the image compression track or the network for other low-level vision tasks as our backbone, here we use a ConvneXt-Tiny[10] which is pre-trained on imagenet to extract features.

ConvneXt[10] follows the previous swin-transformer[9], using a multi-stage design, it has four stages, each stage outputs feature maps of different resolutions. Convolutions with large kernel size, deep-wise Convolutions and GELU activation functions are used. A ConvNeXt block is shown in Figure 3. The number of ConvNeXt[10] block of every stage in ConvNeXt-Tiny is $\{3, 3, 9, 3\}$. Feature maps of different resolutions represent different information extracted from the image. We only extract feature maps from the middle two stages to avoid too much computation. In order to make the final result more convincing and reduce randomness, before inputting an image, we crop it to several patches.

2.2. Nonlinear Residual Module

In this section, we will describe how we calculate the residuals between distorted features and reference features. Our Nonlinear Residual Module is shown in Figure 2.

To calculate residuals between features, we first concatenate them across channels and compute the linear residuals between them using 1×1 convolutions. The formula is as follows:

$$Residual_{linear} = Conv1 \times 1(f_d||f_r) \tag{4}$$

$||$ means concatenate operation.

We implement the non-linearity of the residuals by using an activation function. Same as the backbone we use, we choose GELU as our activation function.

$$Residual_{non-linear} = GELU(Residual_{linear}) \tag{5}$$

Some previous work directly used subtraction to obtain residuals, which we think is too simplistic. The feature maps mapped to the high-dimensional space through the backbone may not satisfy the linear relationship, because the input image has undergone a lot of nonlinear transformations, so we prefer to let the network learn to calculate residuals.

2.3. Transformer Block

Our transformer[16] module is very similar to IQT's[3]. It has N encoders and decoders. In the experiment, we set N to 1 to reduce the computational cost. Different from IQT[3], our input to the encoder is the nonlinear residual between the two distortion maps and the reference image. The core of the transformer[16] is self-attention, which is calculated according to the following formula.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

Where queries Q, keys K, its dimension d_k, and values V are needed. We use a linear layer to get Q, K, V from input.

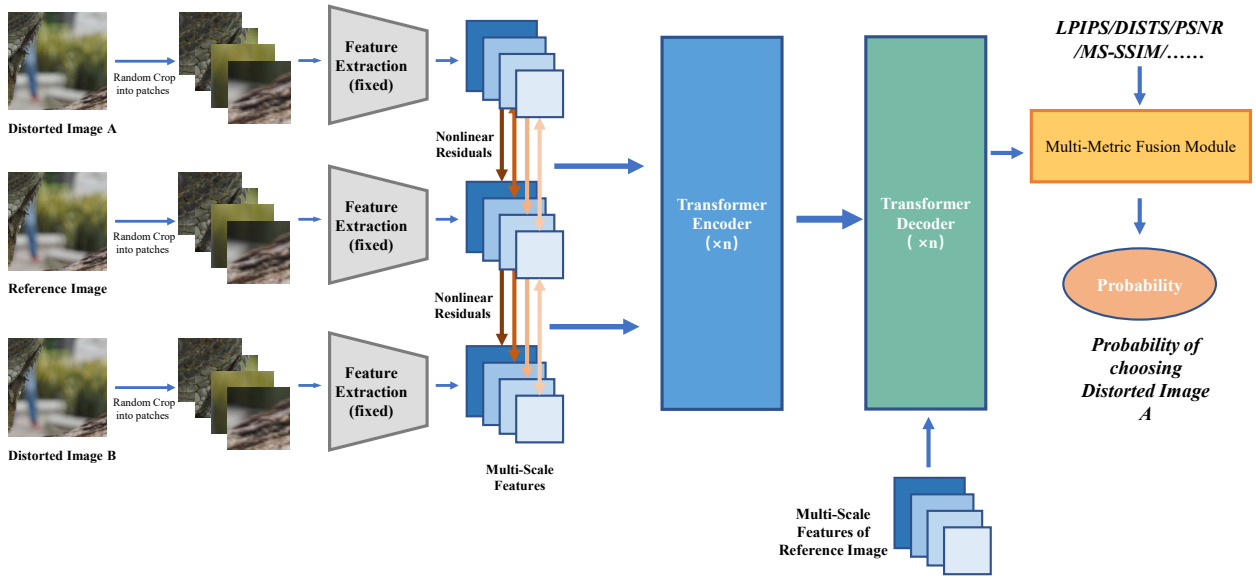

Figure 1. Overall architecture of the proposed image quality assessment method.

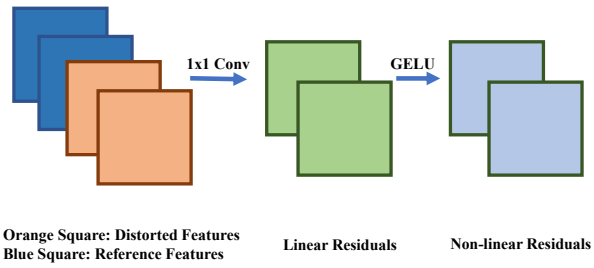

Figure 2. Nonlinear Residuals Module

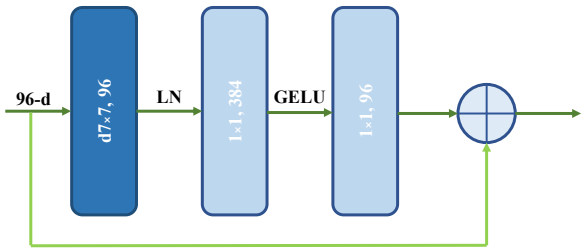

Figure 3. ConvNeXt Block

It should be noted that the K and V of the decoder come from the encoder.

Encoder. When the encoder receives the input, it first uses 1×1 convolution to reduce the dimension of all residuals to the same dimension. We can get $N = H \times W$ patches, H and W are the width and height of residuals, by flattening the residuals in the spatial dimension. We use trainable position embedding to emphasize spatial order between patches.

Decoder. We input the feature map of the reference image into the decoder, and input the feature map of the reference image to emphasize that the comparison is based on the reference image. The output of the encoder is another input of the decoder.

Head. At the end, we use a MLP head to compute quality predicition. The MLP head consists of two fully connected (FC) layers, and the first FC layer is used followed by the ReLU activation. The second FC layer has one channel to predict a single logit.

2.4. Multi-Metric Fusion Module

We use Multi-Metric Fusion Module to improve the generalization of our model. Multi-Metric Fusion takes outputs

of decoder and other metrics as input and first normalizes them before input. The Multi-Metric Fusion Module consists of 7 hidden layers. The size of them is {192, 64, 64, 32, 16, 8, 4}. We send the output to the sigmoid function to get the final probability.

We also tried to use decision tree to fuse multiple metrics, and the final result is not much different from using MLP.

2.5. Loss Function

Different from previous work[3, 6], we treat quality assessment as a classification task rather than a regression task according to the characteristics of the dataset. The classification here includes two labels, the quality of distorted image A is better then the quality of distorted image B (positive, label is 1) and the quality of distorted iamge B is better then the quality of distorted image A (negative, label is 0).

We use cross entropy as our loss function.

$$Loss = -\frac{1}{N}[Label \times log(P_A) + (1 - Label) \times log(1 - P_A)] \tag{7}$$

where P_A represents the probability that the quality of distorted image A is better than the quality of distorted image B and we choose distorted image A.

Actually, we also tried MSE as a loss function like LSGAN[12], but it didn't work well.

3. Experiments

Training. Our experiments are conducted on the datasets provided by CLIC, which includes 122107 triples of images. We use 80% of them as our training set and 20% of them as our testing set. Our model is implemented based on Pytorch[14] framework with a NVIDIA Tesla V100 GPU. We train the model in two stages. In the first stage we train the model without Multi-Metric Fusion Module to initialize the parameters, and the model with Multi-Metric Fusion Module are trained to improve accuracy in the second stage. During training, distorted images and reference images are cropped into to four 224×224 patches. We use Adam[8] optimizer with the initial learning rate of 1e-4, and use cosine annealing scheduler[11] to adjust the learning rate dynamically. The first stage takes 5 hours for 3 epochs and the second stage takes about 30 minutes.

Results. Table 1 demonstrates the accuracy on CLIC validation set, which includes 5220 triples of images. When testing, we crop every images to thirty-two 224×224 patches. The results shows that our model has higher accuracy on the CLIC validation set than LPIPS[21], MSE, MS-SSIM[18].

The results of the ablation experiments are shown in Table 2. When using MSE as the loss function like LSGAN[12], there is a very large drop in performance. The gain brought by the nonlinear residual layer is very considerable, which is in line with our previous assumptions. The gain of multi-metric fusion module is very small, we think it is because the metrics such as LPIPS[21], DISTS[4] and our metric are very similar. When using the cheng2020-anchor[2] pretrained model provided by CompressAI[1] as our backbone, the performance is poor. We think the reason is that the model we selected is optimized for MSE, but the MSE metric itself cannot reflect the quality of the image well, as shown in Table 1, which leads to the fact that the feature map extracted by cheng2020-anchor[2] is not as capable of reflecting semantics as ConvNeXt's[10] .

Table 1. Accuracy on CLIC validation set.

	Accuracy↑
LPIPS(Vgg)[21]	0.744
LPIPS(Alex)[21]	0.737
LPIPS(Squeeze)[21]	0.739
DISTS[4]	0.756
MSE	0.573
MS-SSIM[18]	0.613
GMSD[20]	0.646
VIF[15]	0.605
Ours	**0.780**

Table 2. Ablation experiments on CLIC Validation Set.

	Accuracy↑
w/o MFM	0.779
w/o NRM	0.745
Ours(MSE)	0.484
Ours(Cheng2020-anchor[2])	0.502
Ours	**0.780**

4. Conclusion

In this paper, we study how to design an image quality assessment metric that conforms to human subjective perception. Although we finally outperformed metrics such as LPIPS[21], MS-SSIM[18], DISTS[4] on the CLIC validation set, it is still difficult for our metric to make a very reasonable evaluation of image quality, which is one of the contents of our future research.

Acknowledgements. This work is supported by National Natural Science Foundation of China U21B2012, 62072013 and 61902008, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RCJC20200714114435057, Shenzhen Research Projects of JCYJ20180503182128089 and 201806080921419290, Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003).

References

[1] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020. 4

[2] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7948, 2020. 1, 4

[3] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2021. 1, 2, 4

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 1, 4

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[6] Haiyang Guo, Yi Bin, Yuqing Hou, Qing Zhang, and Hengliang Luo. Iqma network: Image quality multi-scale assessment network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 443–452, 2021. 4

[7] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 2, 4

[11] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[12] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[13] A. Mittal, Fellow, IEEE, R. Soundararajan, and A. C. Bovik. Making a 'completely blind' image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 1

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

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

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[17] 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

[18] 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. 4

[19] A. B. Watson. Toward a perceptual video-quality metric. *Proceedings of SPIE - The International Society for Optical Engineering*, 3299, 1998. 1

[20] 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. 4

[21] 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