Multi-Scale Features and Parallel Transformers Based Image Quality Assessment

Abhisek Keshari*
IIT Jammu
2018ume0126@iitjammu.ac.in

Komal*
IIT Jammu
2020pcs2024@iitjammu.ac.in

Sadbhawna
IIT Jammu
2018rcs0013@iitjammu.ac.in

Badri Subudhi
IIT Jammu
subudhi.badri@iitjammu.ac.in

Abstract

With the increase in multimedia content, the type of distortions associated with multimedia is also increasing. This problem of image quality assessment is expanded well in the PIPAL dataset, which is still an open problem to solve for researchers. Although, recently proposed transformers networks have already been used in the literature for image quality assessment. At the same time, we notice that multi-scale feature extraction has proven to be a promising approach for image quality assessment. However, the way transformer networks are used for image quality assessment until now lacks these properties of multi-scale feature extraction. We utilized this fact in our approach and proposed a new architecture by integrating these two promising quality assessment techniques of images. Our experimentation on various datasets, including the PIPAL dataset, demonstrates that the proposed integration technique outperforms existing algorithms. The source code of the proposed algorithm is available online: https://github.com/KomalPal9610/IQA.

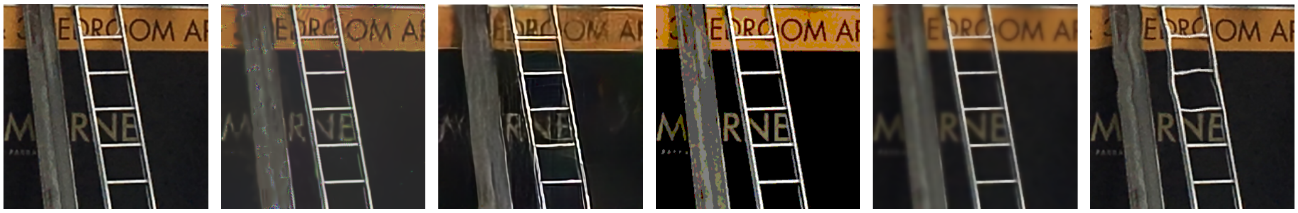
1. Introduction

[1] In recent years, IQA(Image Quality Assessment) gained a lot of attention because image quality is the key factor for various image-based applications such as Image Restoration(IR), Quality Benchmarking [27, 56]. To calculate the perceptual quality of an image, there is a requirement of an automatic method that can be directly linked with the human perception. Full-Reference (FR), No-Reference (NR), and Reduced-Reference (RR) algorithms are the three types of IQA algorithms. In FR, the quality of an image is predicted by comparing the properties or features of target image with its reference image. While in RR and NR algorithms some and no information about the reference image is available. In general, FR algorithms are performing better than the NR images but NR algorithms are preferred in real-time scenario.

Over the years, several IQA metrics have been proposed by different researchers. The most well-known and traditional IQA metrics are mean-squared error (MSE) [57], peak signal-to-noise ratio (PSNR), and SSIM [45]. SSIM tries to anticipate the perceptual quality score based upon the structure similarity between the reference and distorted images.
A few researchers have used natural scene statistics (NSS) such as MSCN coefficients, image entropy, features based on Benford's law and energy subband ratio for the purpose of quality assessment [31, 32, 34, 39]. BRISQUE (dubbed blind/referenceless image spatial quality evaluator) [31] IQA method only uses the pixel information of an image to extract the features. BRISQUE uses the normalized luminance coefficients and pairwise products of these coefficients of the spatial natural scene statistics (NSS) model in the spatial domain.

NIQE [32] is also a most popular machine learning-based algorithm for IQA. Without any display of distorted images and any training on distorted images with human opinion scores, NIQE [32] mainly uses recognizable deviations from statistical regularities observed in natural images. Several researchers have also used gradient information for the purpose of quality assessment such as: [28, 48]. To calculate the change in contrast and structure of the image, in [28] authors have proposed a gradient based method. These gradients are then pooled using component and spatial pooling. Gradient Magnitude Similarity Deviation (GMSD) [48] is based upon predicting the local quality map using the global variation in gradients of reference and distorted images. Further, a global pooling is proposed us-

[1] * indicates that the authors have an equal contribution in the work.



(a) Reference Image (b) Distorted Image 1 MOS: 0.431 (c) Distorted Image 2 MOS: 0.521 (d) Distorted Image 3 MOS: 0.702 (e) Distorted Image 4 MOS: 0.449 (f) Distorted Image 5 MOS: 0.320

Figure 1. Example reference and distorted Images from PIPAL Dataset. [17]

ing this gradient map to calculate the final quality score.

With the development of large datasets (such as TID [35], KADID [26], PIPAL [17]), CNN-based IQA methods have recently attracted significant attention since convolution neural network(CNN) based state-of-the-art methods are used in many image processing and computer vision applications [49] [38]. In [49], the quality assessment is done by using Siamese architecture in such a way that cross-dataset performance is not suffered. And by adding low-level quality cues such as, sharpness, tone and colourfulness, etc. Sebastian Bosse *et al.* [5] proposed a CNN-based image quality assessment method that can be used for both FR and NR image quality assessment. This method does not depend on any hand-crafted features or images statistics. An end-to-end neural network with a number of convolutional and pooling layers is used for feature extraction. By cross-database evaluation of this method, the learned features are extremely robust. Deep similarity for image quality assessment (DeepSim) [14] measures the local similarities of features of distorted and reference images. To determine the overall quality score, the local quality indices are moderately pooled together. Deep Image Quality Assessment (DeepQA) [23], is designed without using a human-oriented perspective, and it learns data distribution by various IQA datasets. Feature Pooling Networks(FPN) constructs high-level semantic feature maps at various scales. FPN adds upsampling procedures to the featurized image pyramid [1] architecture to extract semantically stronger features. Different fields have used these kinds of network designs like Multi-scale multi-view feature aggregation (MSMVFA) [21]. It also combines mid-level attribute features, deep visual features, and high-level semantic features into a single representation for food recognition.

In recent years, transformer networks [43] have received a lot of attention due to it's better performance as compared to conventional CNN models. The transformers has also been used in the field of evaluating image quality and shown cutting-edge performance. For example, You *et al.* proposed the use of vision transformer [13] for No-Reference IQA [50], in which features are extracted using the ResNet-50. Ke *et al.* [22] also used transformer for the purpose of No-reference quality assessment of images. In this paper, authors have used the images at different scales and aspect ratios as the input to the transformer networks and named this network as MUSIQ (Multi-Scale Image Quality Transformer). MUSIQ has the ability to capture the image features at different granularities which makes this network work. In [8], authors have utilised the transformer networks and contrastive loss to catch the features which are quality-aware for the purpose of no-reference quality assessment of images. The winner of the NTIRE 2021 challenge winners have also used the transformer in Full-Reference IQA [9]. In this paper they have extracted the perceptual features from a CNN backbone. Then these features of reference and distorted images are fed into the encoder and decoder of the transformer for the purpose of evaluating image quality.

In this paper, we proposed a MultiScale transformer-based IQA which is an Full-Reference IQA approach. We named our proposed method as Multi-Scale Features and Parallel Transformers(MSFPT) based quality assessment of images. MSFPT is specially designed to capture GAN-based distortions which are introduced by PIPAL dataset [17]. Some examples of reference and distorted images in PIPAL dataset are shown through Fig. 1. Inspired by multi-scale image approaches, we extract the image's features in four different scales by the CNN model. Then these multi-scale features are fed into individual transformers at each scale. The transformer architecture and parameters for all scales are identical. The proposed transformer-based model is then trained for all scales to reliably predict perceptual quality.
To summarize, the following are our key contributions:

- We proposed a new architecture by integrating multi-scale feature extraction and parallel transformers for quality assessment of images.
- Our method significantly outperforms previous existing methods on benchmark datasets LIVE [41], TID2013 [35], and KADID-10k [26]. Also, proposed MSFPT has comparable performance on PIPAL dataset [17] when evaluated as part of NTIRE 2022 IQA Challenge.

The remaining paper is organised as: the proposed MSFPT IQA method is elaborated in Section 2, a detailed comparison is conducted on various IQA datasets in Section 3 followed by concluding remarks in Section 4.

2. Proposed Method

In this section, we proposed a Multi-Scale Features and Parallel Transformer(MSFPT) network based on NTIRE 2021 challenge winner i.e FR Reference IQA with transformers [9]. The MSFPT network takes pairs of image patches as input. Our proposed method follows multi-scale image quality assesment, via traning four independent model for four different scales of PIPAL dataset images, Scale 1(original image), Scale 2(down-scaled image by factor of 2), Scale 3(down-scaled image by factor of 3) and Scale 0.5(up-scaled image by factor of 2). Multi-scaling is used to analyse the image's GAN-based distortion at different scales. It captures GAN-based texture level noises; hence the multi-scale analysis is critical for image quality assessment [20].

Our proposed model consist of four components, Feature extraction block, Interpolation block, Transformer Block, and Averaging Block. *Algorithm 1* is the brief psuedo-code of the proposed algorithm. We have also shown the architecture of proposed algorithm through Fig 2 and Fig 3.

Algorithm 1 MultiScale Transformer based IQA

Input: A pair of reference R_{img} and distored D_{img} image
Output: A predicted IQA score

Denotes feature extraction as FE,
enc_inp_emb $=\{x_{ij}$, where i$\in$ $\{1$ ... BatchSize$\}$, j$\in$ $\{1$ ... SequenceLength$\}$, $x_{ij}=1$ $\}$,
dec_inp_emb $=\{x_{ij}$, where i$\in$ $\{1$ ... BatchSize$\}$, j$\in$ $\{1$ ... SequenceLength$\}$, $x_{ij}=1$ $\}$,
Denotes Transformer block as TB

for $j \leftarrow 1$ **to** 4 **do**
  f_{ref_j}, f_{diff_j} := FE(R_{img}, D_{img}, Scale=j)
  $f^i_{ref_j}$:= Interpolate(f_{ref_j})
  $f^i_{diff_j}$:= Interpolate(f_{diff_j})
  S_j := TB($f^i_{ref_j}$, enc_inp_emb, $f^i_{diff_j}$, dec_inp_emb)
end for
Final Score := Avg(S_1,S_2,S_3,S_4)

2.1. Feature Extraction block

Similar to [9], InceptionNet-V2 CNN model [42], pre-trained on Image-Net [11], is used as a backbone to extract features. Pre-trained weights are imported and frozen. Intermediate layers, namely *block35_2, block35_4, block35_6, block35_8, block35_10, and mixed_5b* are used as a feature map [20]. These blocks are of the same shape for respective scale values, i.e. $320\times a_i \times a_i$, where $a_i \in \{33, 21, 15, 9\}$ for scale values i$\in$$\{$0.5,1,2,3$\}$ respectively. The output of these six feature blocks of the CNN model is concatenated and used as a feature map for the transformer. Pair of Reference and the distorted image is fed to the backbone model via a bilateral branch [3]. It gives two feature maps as an output, f_{ref} and f_{diff}, where f_{ref} is the feature vector extracted from the reference image and f_{diff} is acquired from the difference information between reference and distorted images i.e.

$$f_{diff} = f_{ref} - f_{dist} \tag{1}$$

2.2. Interpolation Block

Feature volumes extracted from the above method have a different shape for respective scale values. To process these feature volumes into the transformer, we need a constant shape of 1920×21×21. Using the bilateral interpolation method, we translate the features from different scales (that are 33×33 for Scale 0.5, 15×15 for Scale 2 and 9×9 for Scale 3) to match 21×21.

2.3. Transformer

The features extracted from the previous stage are fed into the transformer block. A transformer is a ubiquitous, and recently popular deep learning architecture which works on the principle of self-attention mechanism, weighing the importance of each part of the input data in a deferential manner. The transformers has been successfully used to determine the quality of an image. Many researchers [13, 43, 50] have reported the use of transformer for image quality assessment. The attention is the fundamental concept that help in improving the performance of neural machine translation applications in a transformer block. Transformers are primarily developed to operate sequential input data. The transformer's attention layer has access to all past states and weighs them according to a learnt measure of relevance, providing relevant information about tokens that are far away. The Transformer employs the architecture with an encoder and decoder network by using the concepts of attention mechanisms and improving parallelization. The output of a transformer is calculated using a weighted average of the values, where the weights for each value are determined by the query's compatibility function with the relevant key. In the proposed scheme we have used the parallel transformers, corresponding to the multi-scale features obtained from at each scale. **Transformer encoder**, The difference feature embeddings $F_d \in \mathbb{R}^{N\times D}$, N is number of patches and D is the transformer input dimension, is used as the transformer encoder's input. We begin by reducing the vector F_d dimension to D using 1×1 convoluion layer, followed by flattening of dimensions. The number of patches

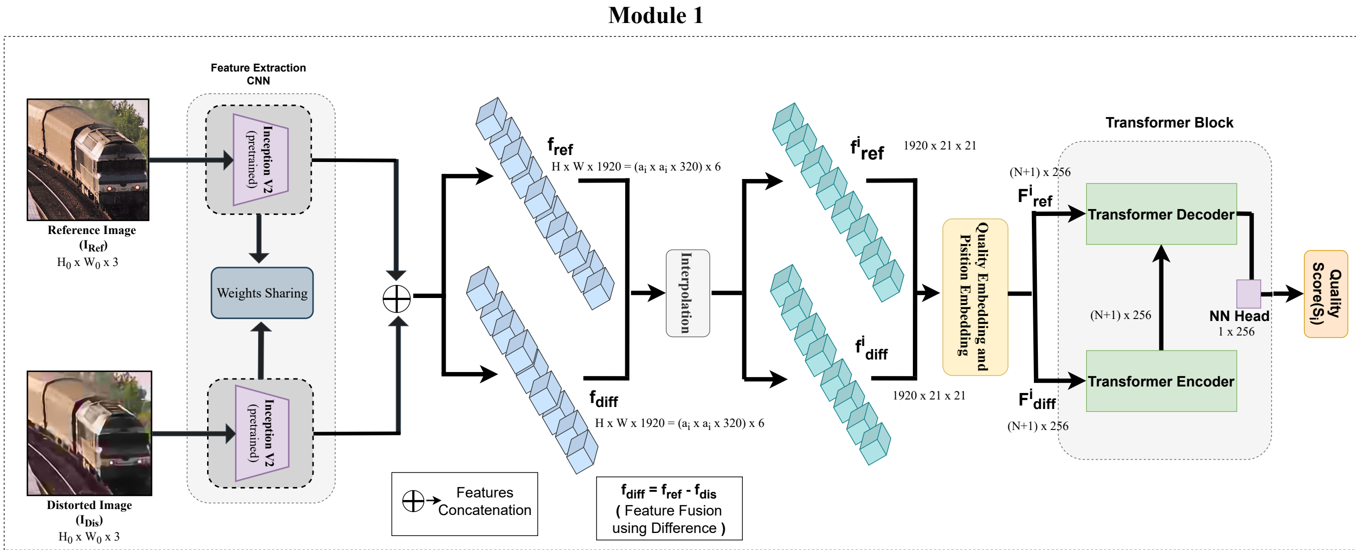


Figure 2. Workflow Diagram of the proposed Module 1.

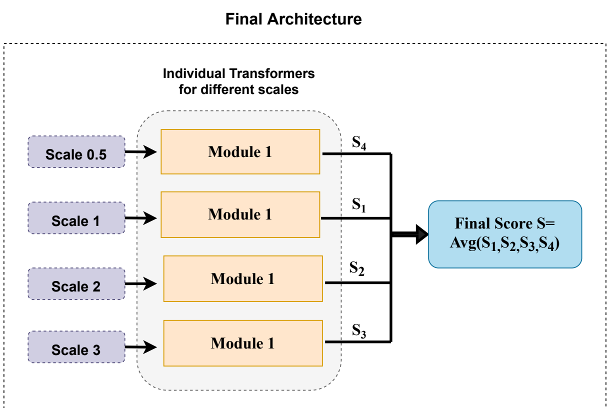


Figure 3. Workflow Diagram of the proposed overall model.

is determined as $N = W \times H$. We append F_{d_0} to the starting of the input feature embedding to add extra quality embedding as used in others vision transformer models [13,50]. In order to keep the positional information, the trainable position embedding $P_d \in \mathbb{R}^{(1+N)\times D}$ is also incorporated. The encoder's calculation can be expressed as shown below:

$$y_0 = \{F_{di} + R_{di}, i \in \{0, 1, ..., N\}\}, \tag{2}$$

and

$$q_i = k_i = v_i = y_{i-1}, \tag{3}$$

and

$$y_i^{'} = LN(MHA(q_i, k_i, v_i) + y_{i-1}), \tag{4}$$

where

$$y_i = LN(MLP(y_i^{'}) + y_i^{'}), i \in \{1, 2, ..., L\}, \tag{5}$$

and

$$\{F_{Ei}, i \in \{1, 2, ..., N\}\} = y_L, \tag{6}$$

where L is the number of encoder layers. The input feature embeddings and output has the same size $F_e \in \mathbb{R}^{(1+N)\times D}$.

Transformer decoder The decoder takes three components as input the output of encoder F_E, the reference feature embeddings $F_r \in \mathbb{R}^{(1+N)\times D}$, obtained through reduction followed by flattening, extra quality embeddings and position embedding. F_E is utilised as key-value in second Multi head attention layer. The calculation of decoder can be formed as:

$$y_0 = \{F_i + P_i, \forall i \in \{1, 2, ..., N\}\}, \tag{7}$$

and

$$v_i = q_i = k_i = z_{i-1}, \tag{8}$$

and

$$y_i^{'} = LN\left(MLA\left(q_i, k_i, v_i\right) + z_{i-1}\right), \tag{9}$$

where

$$k_i^{'} = v_i^{'} = y_L, \tag{10}$$

,

$$q_i^{'} = z_i^{'}, \tag{11}$$

$$z_i^{"} = LN(MHA(q_i^{'}, k_i^{'}, v_i^{'}) + z_i^{'}), \tag{12}$$

and

$$z_i^{'} = LN(MLP(z_i^{"}) + z_i^{"}), i \in \{1, 2, ..., L\}, \tag{13}$$

and

$$\{F_{Di}, i \in \{1, 2, ..., N\}\} = z_L, \tag{14}$$

Table 1. A tabulated summary of the datasets used for the performance comparison.

Database	Reference Images	Distorted Images	Distortion Types	Ratings	Rating Type	Distortion Type	Environment
LIVE [41]	29	779	5	25k	MOS	traditional	lab
TID2013 [35]	25	3000	25	524k	MOS	traditional	lab
KADID-10k [26]	81	10.1k	25	30.4k	MOS	traditional	crowdsourcing
PIPAL [17]	250	29k	40	1.13m	MOS	trad. + algo outputs	crowdsourcing

where L is the number of encoder layers. The input feature embeddings and output has the same size $F_E \in R^{(1+N)\times D}$.

Head. The Neural Network block calculates the final quality score. The NN Block receives the first vector of the decoder output, $F_{D_0} \in R^{1\times D}$ in Eq. 2, which carries the quality information. The Block is made up of two completely connected (FC) layers, with the ReLU activation occurring after the first FC layer. A single score is predicted by the second FC layer, which contains one channel.

2.4. Averaging Module

Transformer Block T_i predicts the quality score for scale i (S_i). The final quality score (S) is calculated by averaging the estimated quality score for each scale:

$$FinalQualityScore(S) = \frac{\sum_{i=1}^{4} S_i}{4}. \tag{15}$$

3. Experiments

3.1. Datasets

Our experiments are conducted on four benchmark Image quality datasets, LIVE [41], TID2013 [35], KADID-10k [26] and PIPAL [17]. The LIVE dataset contains 29 reference images; from these images, using five different traditional distortion types, 779 distorted images are created. TID2013 contains 25 reference images as well as 3,000 distorted images generated by 24 different distortions, with five levels for each distortion type. KADID-10k includes 81 reference images and 10.1k distorted images generated by 25 distortions. PIPAL contains 250 reference images and 29k distorted images generated by 40 kinds of distortions. This dataset has traditional and algorithmic outputs, i.e. GAN-based distortions produced by different GAN based algorithms. The validation set of the PIPAl dataset contains 25 reference images and 1650 distortion images, whereas the testing set of the PIPAl dataset contains 25 reference images and 1,650 distortion images.

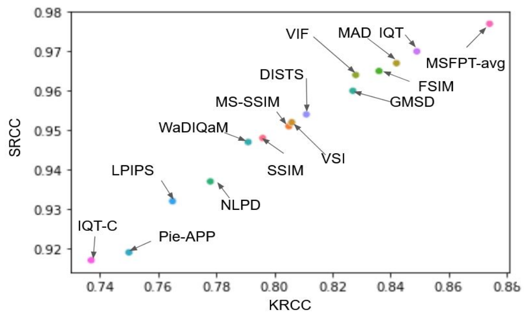
3.2. Implementation Details

In the training phase, a given image is cropped to obtain the desired patch of size HxWxC. For PIPAL dataset we have H = W = 192, C = 3 and for LIVE [41], KADID-10k [26] and TID2013 [35] H = W = 256, C = 3. The feature volume of MSFPT has N = 442 patches. In testing phase, same number of patches are obtained from the image pair given. We extract M overlapping patches where M is the number of ensembles used and use an average of M individual patch quality ratings to predict the final quality score. The Adam optimizer was used with weight-decay $\alpha = 1e^{-5}$, β_1 = 0.9 and β_2 = 0.999 with L1 as a loss function since it is more resilient to outliers than MSE loss. We have set the learning rate to $2e^{-4}$ and used cosine annealing learning rate scheduler, A batch size of 16 was chosen. PyTorch 1.10.1 was used with two NVIDIA V100 GPUs and CUDA 11.0. Data augmentation, including random crop, vertical flip, random rotation, and horizontal flip, is applied during the training.

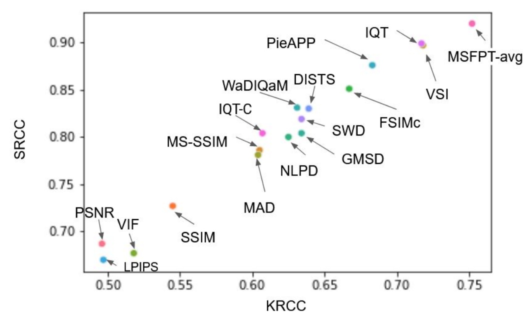
Table 2. Performence comparison over LIVE [41] and TID2013 [35] Datasets. [51]

Method	LIVE			TID2013		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	0.865	0.873	0.68	0.677	0.687	0.496
SSIM [45]	0.937	0.948	0.796	0.777	0.727	0.545
MS-SSIM [47]	0.94	0.951	0.805	0.83	0.786	0.605
VSI [53]	0.948	0.952	0.806	0.9	0.897	0.718
MAD [25]	0.968	0.967	0.842	0.827	0.781	0.604
VIF [40]	0.96	0.964	0.828	0.771	0.677	0.518
FSIMc [55]	0.961	0.965	0.836	0.877	0.851	0.667
NLPD [24]	0.932	0.937	0.778	0.839	0.8	0.625
GMSD [48]	0.957	0.96	0.827	0.855	0.804	0.634
WaDIQaM [6]		0.947	0.791	0.834	0.831	0.631
PieAPP [36]	0.908	0.919	0.75	0.859	0.876	0.683
LPIPS [56]	0.934	0.932	0.765	0.749	0.67	0.497
DISTS [12]	0.954	0.954	0.811	0.855	0.83	0.639
SWD [15]	-	-	-	-	0.819	0.634
IQT [9]	-	0.97	0.849	0.943	0.899	0.717
IQT-C [9]	-	0.917	0.737	-	0.804	0.607
MSFPT-1	0.962	0.976	0.874	**0.955**	**0.949**	**0.807**
MSFPT-2	0.958	0.964	0.846	0.872	0.857	0.673
MSFPT-3	0.944	0.955	0.824	0.853	0.828	0.635
MSFPT-0.5	0.963	0.976	**0.875**	0.831	0.796	0.598
MSFPT-avg	**0.972**	**0.977**	0.874	0.929	0.92	0.752

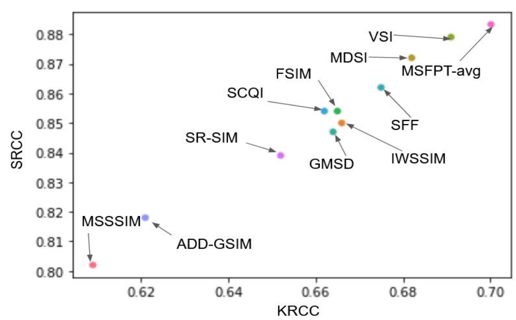
We compare MSFPT network with several state-of-the-art methods on all four datasets [17, 26, 35, 41] for IQA. The methods have deep learning-based methods such as PieAPP [37], LPIPS [56], SWD [16] and DISTS [12] and shallow methods like SSIM [45] and PSNR. For most cases our method shows more promising results than current deep



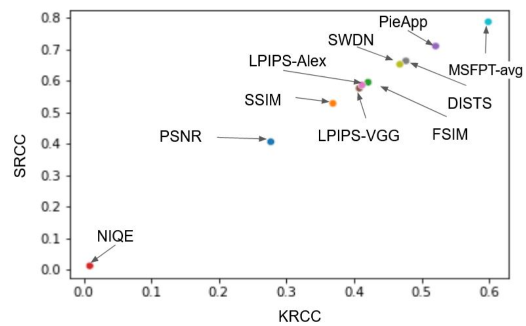
(a) LIVE Dataset



(b) TID 2013 Dataset



(c) KADID Dataset



(d) PIPAL Dataset

Figure 4. Quantitative comparison of IQA methods. (a) LIVE Dataset, (b)TID 2013 Dataset, (c)KADID Dataset, (d) PIPAL Dataset.

Table 3. Performence comparison over KADID Dataset. [26]

Method	KADID		
	PLCC	SRCC	KRCC
SSIM [45]	0.723	0.724	0.537
MS-SSIM [47]	0.801	0.802	0.609
IWSSIM [46]	0.846	0.850	0.666
MDSI [33]	0.873	0.872	0.682
VSI [53]	0.878	0.879	0.691
FSIM [55]	0.851	0.854	0.665
GMSD [48]	0.847	0.847	0.664
SFF [7]	0.862	0.862	0.675
SCQI [2]	0.853	0.854	0.662
ADD-GSIM [19]	0.817	0.818	0.621
SR-SIM [52]	0.834	0.839	0.652
MSFPT-1	0.822	0.846	0.653
MSFPT-2	0.796	0.799	0.613
MSFPT-3	0.667	0.674	0.495
MSFPT-0.5	0.857	0.857	0.672
MSFPT-avg	**0.888**	**0.883**	**0.700**

Table 4. Ablation study with respect to the different scales.

Model Name	Validation		
	Main Score	PLCC	SRCC
MSFPT-1	1.552	0.784	0.768
MSFPT-2	1.522	0.773	0.749
MSFPT-3	1.47	0.749	0.721
MSFPT-0.5	-	-	-
MSFPT-avg	1.598	0.810	0.788
Model Name	**Testing**		
	Main Score	**PLCC**	**SRCC**
MSFPT-avg	1.450	0.738	0.713
MSFPT-1	1.254	0.637	0.617
MSFPT + Bert + Scale1	1.383	0.699	0.684
MSFPT + Bert + Scale2	1.361	0.698	0.663
MSFPT + Bert + Scale3	1.182	0.621	0.561
MSFPT + Bert + Avg. of 1,2,3	1.44	0.73	0.71

learning-based methods. Our model out performs other deep-learning based models like IQT method [9] on LIVE [41] data set by 0.07 SRCC and 0.029 in KRCC. In case of TID2013 [35] by using weight sharing and multi-scale we outperform existing deep-learning models by 0.021 SRCC, and 0.034 KRCC. For KADID-10k [26], it outperforms various IQA methods like VSI by 0.01 PLCC, 0.004 SRCC and 0.009 KRCC.

3.3. Ablation study

The use of different information between various scales of input images is one of the vital characteristics in the proposed architecture. Four types of scales are available, i.e. 1, 2, 3 and 0.5, as mentioned in the Table 4. We conducted an ablation experiment to study the influence of input shape

Table 5. Performance comparison of the proposed algorithm in NTIRE IQA challenge, Testing phase.

Team Name	Main score	PLCC	SRCC
Anynomus1	1.651	0.826	0.822
Anynomus2	1.642	0.827	0.815
Anynomus3	1.64	0.823	0.817
Anynomus4	1.541	0.775	0.766
Anynomus5	1.538	0.772	0.765
Anynomus6	1.501	0.763	0.737
Pico Zen(ours)	1.45	0.738	0.713
Anynomus8	1.403	0.703	0.701

and transformer type, and the results of the performance evaluation are provided in Table 4.

In the proposed algorithm, we have used Attention is all you need [43] transformer, that gives significantly better performance over CNN based models. We have also tried to used Bert in the proposed algorithm and observed (from Table 4) that the Bert is giving slightly poorer performance. These results clearly validates that incorporating multi-scale features with the parallel transformers significantly improves the performances.

3.4. NTIRE 22 IQA Challenge Report

In both validation and testing phases, we use MSFPT model trained on PIPAL dataset on four different scales with batch size 16. Table 5 shows the competition's final ranking during the testing phase.

4. Conclusions

In this paper, we presented a full-reference image quality assessment algorithm based on parallel transformers and multi-scale CNN features. These features are trained for the purpose of quality prediction using transformers network with encoders and decoders. We conducted extensive experimental studies to show the superiority of using this combination of parallel transformers and multi-scale features as compared to other combination of networks. The proposed method outperforms current state-of-the-art image quality assessment methods in terms of performance.

Table 6. Performance comparison over Validation Dataset of NTIRE-2022 FR [18]

Model Name	Main Score	SRCC	PLCC
MSFPT-avg (our)	1.598	0.81	0.788
PSNR	0.503	0.234	0.269
NQM [10]	0.666	0.302	0.364
UQI [44]	0.966	0.461	0.505
SSIM [45]	0.696	0.319	0.377
MS-SSIM [47]	0.457	0.338	0.119
RFSIM [54]	0.539	0.254	0.285
GSM [29]	0.829	0.379	0.45
SRSIM [52]	1.155	0.529	0.626
FSIM [55]	1.005	0.452	0.553
VSI [53]	0.905	0.411	0.493
NIQE [32]	0.141	0.012	0.129
MA [30]	0.196	0.099	0.097
PI [4]	0.198	0.064	0.134
Brisque [31]	0.06	0.008	0.052
LPIPS-Alex [56]	1.175	0.569	0.606
LPIPS-VGG [56]	1.162	0.551	0.611
DISTS [12]	1.243	0.608	0.634

Table 7. Performance comparison over Testing Dataset of NTIRE-2022 FR [18]

Model Name	Main Score	SRCC	PLCC
MSFPT-avg (our)	1.45	0.738	0.713
PSNR	0.526	0.249	0.277
NQM [10]	0.76	0.364	0.395
UQI [44]	0.87	0.42	0.45
SSIM [45]	0.753	0.361	0.391
MS-SSIM [47]	0.532	0.369	0.163
RFSIM [54]	0.632	0.304	0.328
GSM [29]	0.874	0.409	0.465
SRSIM [52]	1.209	0.573	0.636
FSIM [55]	1.075	0.504	0.571
VSI [53]	0.975	0.458	0.517
NIQE [32]	0.166	0.034	0.132
MA [30]	0.287	0.14	0.147
PI [4]	0.249	0.104	0.145
Brisque [31]	0.14	0.071	0.069
LPIPS-Alex [56]	1.137	0.566	0.571
LPIPS-VGG [56]	1.228	0.595	0.633
DISTS [12]	1.342	0.655	0.687

References

[1] Edward Adelson, Charles Anderson, James Bergen, Peter Burt, and Joan Ogden. Pyramid methods in image processing. *RCA Eng.*, 29, 11 1983. 2

[2] Sung Ho Bae and Munchurl Kim. A novel image quality assessment with globally and locally consilient visual quality perception. *IEEE Transactions on Image Processing*, 25(5):2392–2406, May 2016. 6

[3] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking, 2021. 3

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. *CoRR*, abs/1711.06077, 2017. 7

[5] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *CoRR*, abs/1612.01697, 2016. 2

[6] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27:206–219, 01 2018. 5

[7] Hua-wen Chang, Hua Yang, Yong Gan, and Ming-Hui Wang. Sparse feature fidelity for perceptual image quality assessment. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 22, 06 2013. 6

[8] Pengfei Chen, Leida Li, Qingbo Wu, and Jinjian Wu. Spiq: A self-supervised pre-trained model for image quality assessment. *IEEE Signal Processing Letters*, 29:513–517, 2022. 2

[9] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 433–442, 2021. 2, 3, 5, 6

[10] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2000. 7

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, Apr 2021. 5, 7

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2, 3, 4

[14] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017. 2

[15] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric, 2020. 5

[16] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric, 2020. 5

[17] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. 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. 2, 5

[18] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Radu Timofte, et al. NTIRE 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 7

[19] Ke Gu, Shiqi Wang, Guangtao Zhai, Weisi Lin, Xiaokang Yang, and Wenjun Zhang. Analysis of distortion distribution for pooling in image quality prediction. *IEEE Transactions on Broadcasting*, 62:446–456, 2016. 6

[20] 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 (CVPR) Workshops*, pages 443–452, June 2021. 3

[21] Shuqiang Jiang, Weiqing Min, Linhu Liu, and Zhengdong Luo. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, 29:265–276, 2020. 2

[22] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 2

[23] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, 2017. 2

[24] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016:1–6, 02 2016. 5

[25] Eric Larson and Damon Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19:011006, 01 2010. 5

[26] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 06 2019. 2, 5, 6

[27] Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011. 1

[28] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012. 1

[29] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012. 7

[30] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *CoRR*, abs/1612.05890, 2016. 7

[31] 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, 7

[32] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 1, 7

[33] Hossein Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 08 2016. 6

[34] Fu-Zhao Ou, Yuan-Gen Wang, and Guopu Zhu. A novel blind image quality assessment method based on refined natural scene statistics. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1004–1008, 2019. 1

[35] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 2, 5, 6

[36] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *2018 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 1808–1817, 2018. 5

[37] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 2

[39] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012. 1

[40] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 5

[41] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 2, 5, 6

[42] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2, 3, 7

[44] Zhou Wang and A.C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. 7

[45] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 5, 6, 7

[46] Zhou Wang and Qiang Li. Li, q.: Information content weighting for perceptual image quality assessment. ieee image proc. 20(5), 1185-1198. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 20:1185–98, 11 2010. 6

[47] Zhou Wang, Eero P. Simoncelli, and Alan Conrad Bovik. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2:1398–1402 Vol.2, 2003. 5, 6, 7

[48] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *Image Processing, IEEE Transactions on*, 23, 08 2013. 1, 5, 6

[49] Dan Yang, Veli-Tapani Peltoketo, and Joni-Kristian Kämäräinen. Cnn-based cross-dataset no-reference image quality assessment. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3913–3921, 2019. 2

[50] Junyong You and Jari Korhonen. Transformer for image quality assessment. *CoRR*, abs/2101.01097, 2021. 2, 3, 4

[51] Chao Zeng and Sam Kwong. Learning transformer features for image quality assessment. *CoRR*, abs/2112.00485, 2021. 5

[52] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. *2012 19th IEEE International Conference on Image Processing*, pages 1473–1476, 2012. 6, 7

[53] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 5, 6, 7

[54] Lin Zhang, Lei Zhang, and Xuanqin Mou. Rfsim: A feature based image quality assessment metric using riesz transforms. In *2010 IEEE International Conference on Image Processing, ICIP 2010 - Proceedings*, pages 321–324, Dec. 2010. 2010 17th IEEE International Conference on Image Processing, ICIP 2010 ; Conference date: 26-09-2010 Through 29-09-2010. 7

[55] 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. 5, 6, 7

[56] 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. 1, 5, 7

[57] W. Zhou and Alan Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26:98–117, 01 2009. 1