NTIRE 2022 Challenge on Perceptual Image Quality Assessment

Jinjin Gu Haoming Cai Chao Dong Jimmy S. Ren Radu Timofte Yuan Gong
Shanshan Lao Shuwei Shi Jiahao Wang Sidi Yang Tianhe Wu Weihao Xia
Yujiu Yang Mingdeng Cao Cong Heng Lingzhi Fu Rongyu Zhang
Yusheng Zhang Hao Wang Hongjian Song Jing Wang Haotian Fan
Xiaoxia Hou Ming Sun Mading Li Kai Zhao Kun Yuan Zishang Kong
Mingda Wu Chuanchuan Zheng Marcos V. Conde Maxime Burchi Longtao Feng
Tao Zhang Yang Li Jingwen Xu Haiqiang Wang Yiting Liao Junlin Li
Kele Xu Tao Sun Yunsheng Xiong Abhisek Keshari Komal
Sadbhawana Thakur Vinit Jakhetiya Badri N Subudhi Hao-Hsiang Yang
Hua-En Chang Zhi-Kai Huang Wei-Ting Chen Sy-Yen Kuo Saikat Dutta
Sourya Dipta Das Nisarg A. Shah Anil Kumar Tiwari

Abstract

This paper reports on the NTIRE 2022 challenge on perceptual image quality assessment (IQA), held in conjunction with the New Trends in Image Restoration and Enhancement workshop (NTIRE) workshop at CVPR 2022. This challenge is held to address the emerging challenge of IQA by perceptual image processing algorithms. The output images of these algorithms have completely different characteristics from traditional distortions and are included in the PIPAL dataset used in this challenge. This challenge is divided into two tracks, a full-reference IQA track similar to the previous NTIRE IQA challenge and a new track that focuses on the no-reference IQA methods. The challenge has 192 and 179 registered participants for two tracks. In the final testing stage, 7 and 8 participating teams submitted their models and fact sheets. Almost all of them have achieved better results than existing IQA methods, and the winning method can demonstrate state-of-the-art performance.

1. Introduction

Assessing the perceptual quality of an image is a fundamental requirement in the fields of image acquisition, transmission, compression, reproduction, and processing. Image quality assessment (IQA) methods are tools that use computational models to measure the perceptual quality of images. As the "evaluation mechanism", IQA plays a critical role in guiding the development of image processing algorithms. However, distinguish perceptually better images is not an easy task [48, 26, 25], especially as newly-appeared image distortion types continue to challenge IQA methods, *e.g.*, Generative Adversarial Networks (GANs) based algorithms [24] and perceptual-oriented algorithms [32, 34, 61, 75]. IQA methods that are capable of automatically and accurately predicting subjective quality are in demand nowadays.

This NTIRE 2022 Perceptual Image Quality Assessment Challenge aims to push the developing state-of-the-art perceptual image quality assessment methods to deal with the novel GAN-based distortion types and gain new insights. We employ the PIPAL dataset [26] in this challenge, which is the only dataset including the results of perceptual-oriented algorithms. The PIPAL dataset contains 200 reference images, 29k distorted images and 1.88 million human judgements. The large size and diversity of distortion types of the PIPAL dataset allow us to benchmark these IQA methods.

This is the second perceptual IQA challenge held at the NTIRE workshop [28]. In the last challenge, several submitted entries significantly outperformed existing methods and achieved state-of-the-art performance in the full-reference IQA field. In this challenge, we included two tracks. The first track is similar to the NTIRE 2021 IQA challenge, focusing on full-reference methods. Considering the wide range of application scenarios and demands of no-reference methods, we set up a second track that focuses on no-reference IQA methods. We anticipate this new track to

*Jinjin Gu (jinjin.gu@sydney.edu.au), Haoming Cai, Chao Dong, Jimmy Ren and Radu Timofte are the NTIRE 2022 challenge organizers. The other authors participated in the challenge. Appendix.B and Appendix.C contain the authors' team names and affiliations. The NTIRE website: https://data.vision.ee.ethz.ch/cvl/ntire22/

push developing state-of-the-art no-reference IQA methods.

The challenge has 192 and 179 registered participants for two tracks, respectively. Among them, 7 and 8 participating teams submitted their models and fact sheets in the final testing stage, respectively. They introduce new technologies in network architectures, loss functions, ensemble methods, data augmentation methods, and *etc*. We present detailed challenge results in Sec 1.

This challenge is one of the NTIRE 2022 associated challenges: spectral recovery [2], spectral demosaicing [1], perceptual image quality assessment [27], inpainting [50], night photography rendering [20], efficient super-resolution [37], learning the super-resolution space [41], super-resolution and quality enhancement of compressed video [66], high dynamic range [47], stereo super-resolution [59], burst super-resolution [6].

2. Related Work

Full-Reference Image quality assessment (FR-IQA). FR-IQA methods evaluate the similarity between a distorted image and a given reference image and have been widely used to evaluate image/video processing algorithms. FR-IQA methods follow a long line of works, the most well-known of which is PSNR and SSIM [62]. SSIM introduces structural information in measuring image similarity and opens a precedent for evaluating image structure or feature similarity. After that, various FR-IQA methods have been proposed to bridge the gap between the results of IQA methods and human judgements [63, 69, 72, 52, 70]. Similar to other computer vision problems, advanced data-driven methods have also motivated the investigation of applications of IQA, such as LPIPS [74], PieAPP [49], WaDIQaM [9], SDW [25] and DISTS [17]. The 2021 NTIRE challenge has also brought some excellent FR-IQA methods, i.e., Cheon *et al*. [13] propose a transformer-based FR-IQA method IQT and win the first place at the challenge, Guo *et al*. [30] propose bilateral-branch multi-scale image quality estimation (IQMA) network, and Shi *et al*. [54] propose Region Adaptive Deformable Network (RADN).

No-Reference Image quality assessment (NR-IQA). In addition to the above FR-IQA methods, NR-IQA methods are proposed to assess image quality without a reference image. A typical NR-IQA is often based on natural image statistics. Natural images usually follow these natural image prior distributions, while distorted images often break such statistical regularities. Variation of methods have been used to extract natural image statistics [46, 43, 71, 51, 45, 73, 67]. In the era of deep learning, deep networks are anticipated to replace hand-crafted feature extraction and learn statistical priors on images, and many deep learning-based NR-IQA methods are proposed [33, 10, 38, 58, 76, 7, 65, 55, 78, 77]. More related to this work, Blau *et al*. [8] combine two NR-IQA methods, Ma [42] and NIQE [44] and propose the Perceptual Index (PI) method to measure the perceptual quality of super-resolution results without reference image. Although it can lead to the development of better perceptual-oriented algorithms compared with other FR-IQA methods that focus on evaluating distortion, its IQA performance is still unsatisfactory. In this challenge, we set a new track that focuses on NR-IQA methods and bring more advanced NR-IQA methods to this field.

Perceptual-oriented and GAN-based distortion. In the past years, benefiting from the invention of perceptual-oriented loss function [32, 61] and GANs [24]. many photo-realistic image generation and processing algorithms are proposed [34, 61, 60, 75, 12]. These perceptual-oriented algorithms greatly improve the perceptual effect of the output image. However, they also bring completely new characteristics to the output images. In general, these methods often fabricate seemingly realistic yet fake details and textures. They do not quite match the quality of detail loss, as they usually contain texture-like noise, or the quality of noise, the noise is similar to the ground truth in appearance but is not accurate. The quality evaluation of such images has been proved challenging for IQA methods [26]. Gu *et al*. [26] contribute an IQA dataset called Perceptual Image Processing ALgorithms dataset (PIPAL), including the results of Perceptual-oriented image processing algorithms. This data set is used to benchmark different IQA methods and is used as the training and testing dataset in this challenge.

3. The NTIRE Challenge on Perceptual IQA

We host the NTIRE 2022 Perceptual Image Quality Assessment Challenge to push developing state-of-the-art FR- and NR- IQA methods to deal with the novel GAN-based distortion types, compare different solutions, and gain new insights. Details about the challenge are as follows:

Tracks. We include two tracks: the FR-IQA track that focuses on evaluating full-reference IQA methods and a new NR-IQA track for no-reference IQA methods.

- Track 1: The task of this track is to obtain an FR-IQA method capable of producing high-quality perceptual similarity results between the given distorted images and the corresponding reference images with the best correlation to the reference ground truth MOS score.
- Track 2: The task of this track is to obtain an NR-IQA method capable of producing high-quality perceptual quality results with the best correlation to the reference ground truth MOS score. Only distorted images

Table 1. Quantitative results for the NTIRE 2022 Perceptual IQA challenge.

Rank	Team Name	Author/Method	PIPAL-NTIRE22-Test		
			Main Score	SRCC	PLCC
Track 1: Full-Reference IQA					
1	THU1919Group	shanshan	1.6511	0.8227	0.8284
2	Netease OPDAI	CongHeng.	1.6422	0.8152	0.8271
3	KS	JustTryTry	1.6404	0.8170	0.8235
4	JMU-CVLab	burchim	1.5406	0.7659	0.7747
5	Yahaha!	FLT	1.5375	0.7654	0.7722
6	debut_kele	debut	1.5006	0.7372	0.7634
7	Pico Zen	Komal	1.4504	0.7129	0.7375
8	Team Horizon	tensorcat	1.4032	0.7006	0.7027
	Baselines	IQT (NTIRE-21 Winner)	1.5884	0.7895	0.7989
		LPIPS-Alex	1.1369	0.5658	0.5711
		LPIPS-VGG	1.2278	0.5947	0.6331
		DISTS	1.3422	0.6548	0.6873
		SSIM	0.7530	0.3615	0.3915
		PSNR	0.5263	0.2493	0.2769
Track 2: No-Reference IQA					
1	THU_IIGROUP	THU_IIGROUP	1.4436	0.7040	0.7396
2	DTIQA	EvaLab.	1.4367	0.6996	0.7371
3	JMU-CVLab	nanashi	1.4219	0.6965	0.7254
4	KS	JustTryTry	1.4066	0.6808	0.7257
5	NetEase OPDAI	wanghao1003	1.3902	0.6705	0.7196
6	Withdrawn submission	anonymous	1.1828	0.5760	0.6068
7	NTU607QCO-IQA	mrchang87	1.1117	0.5269	0.5848
	Baselines	NIQE	0.1418	0.0300	0.1118
		MA	0.3978	0.1737	0.2242
		PI	0.2764	0.1234	0.1529
		Brisque	0.5722	0.2695	0.3027

are given in this track, and no reference images are available.

Dataset. Following NTIRE 2021 IQA challenge [28], we employ a subset of the PIPAL dataset as the training set and an extended version of the PIPAL dataset as the validation and the testing set. The PIPAL dataset includes traditional distortion types, image restoration results, compression results, and novel GAN-based image processing outputs. More than 1.88 million human judgements are collected to assign mean opinion scores (MOS) for PIPAL images using the Elo rating system [19]. The original PIPAL dataset includes 250 high-quality, diverse reference images, and each has 116 different distorted images. We use 200 of the 250 reference images and their distorted images as the training set (in total 200×116 distorted images). All training images and the MOS scores are publicly available.

The validation set and the testing set are selected from the extended version of the PIPAL dataset [28]. The validation set contains 25 reference images and 40 distorted images for each. The testing set contains 25 reference images and all the 66 distorted images for each reference image. The newly collected distortion types are all outputs of GAN-based image restoration algorithms or GAN-based compression algorithms. In total, 3300 additional images are collected. Note that for the participants, the training set and the validation/testing set contain completely different references and distorted images, ensuring the final results' objectivity. Methods trained with additional labelled IQA datasets (pre-training using non-IQA datasets such as ImageNet is allowed) will be disqualified from the final ranking for both tracks.

Evaluation protocol. Align with the challenge at NTIRE 2021 [28], our evaluation indicator, namely main score, consists of both Spearman rank-order correlation coefficient (SRCC) [53] and Person linear correlation coefficient

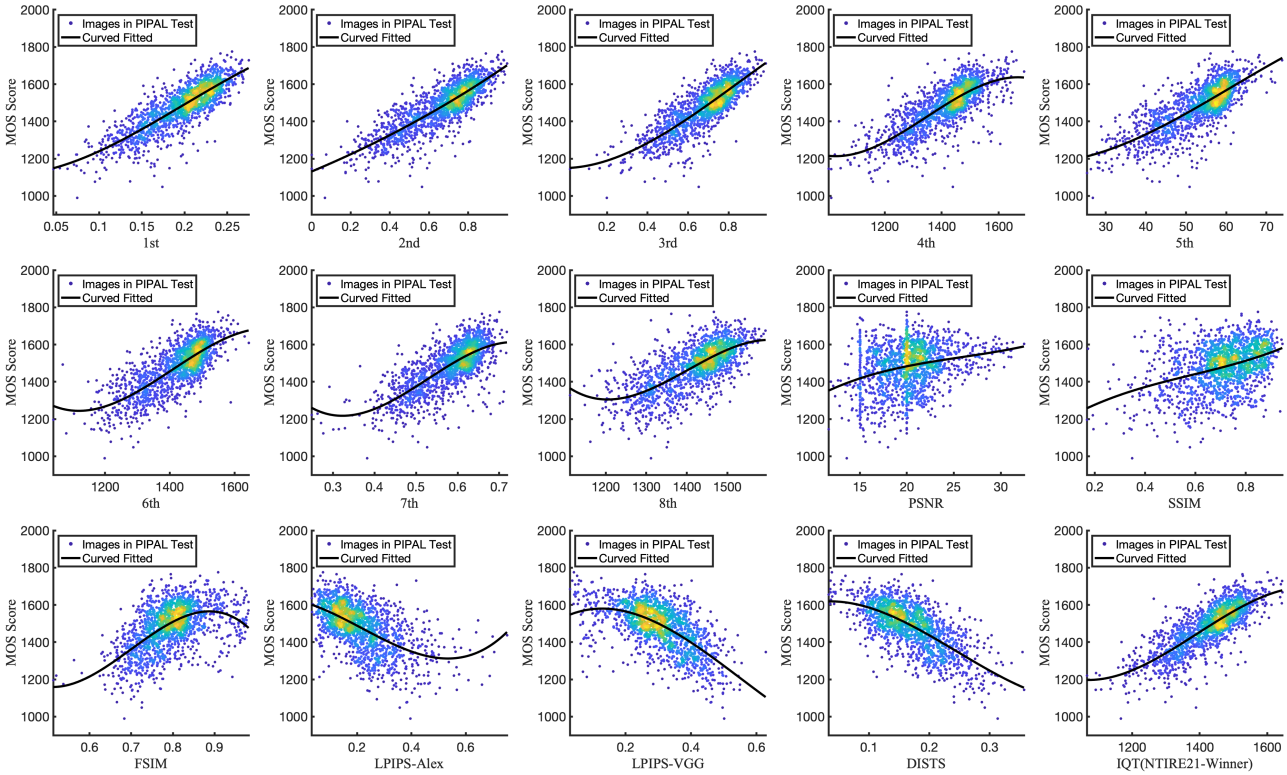

Figure 1. FR-IQA Track's Scatter plots of the objective scores vs. the MOS scores. The curves were obtained by a third-order polynomial nonlinear fitting.

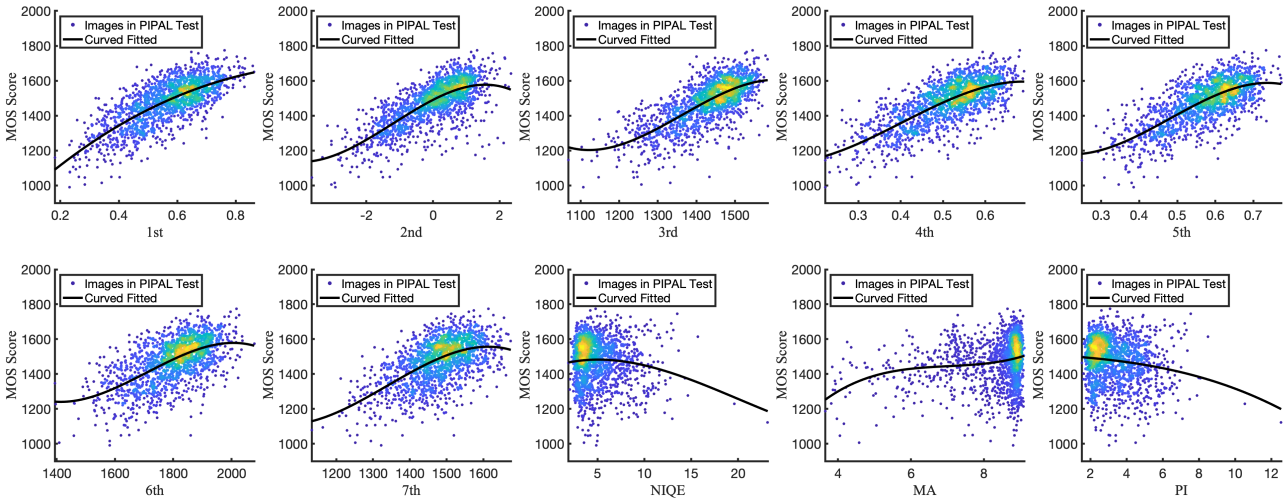

Figure 2. NR-IQA Track's Scatter plots of the objective scores vs. the MOS scores. The curves were obtained by a third-order polynomial nonlinear fitting.

(PLCC) [5]:

$$\text{Main Score} = |\text{SRCC}| + |\text{PLCC}|. \quad (1)$$

The SRCC evaluates the monotonicity of methods that whether the scores of high-quality images are higher (or lower) than low-quality images. The PLCC is often used to evaluate the accuracy of methods [53, 25]. Before calculating PLCC index, we perform the third-order polynomial nonlinear regression as suggested in the previous works [48, 26]. By combining SRCC and PLCC, our indicator can measure the performance of participating models in an all-round way.

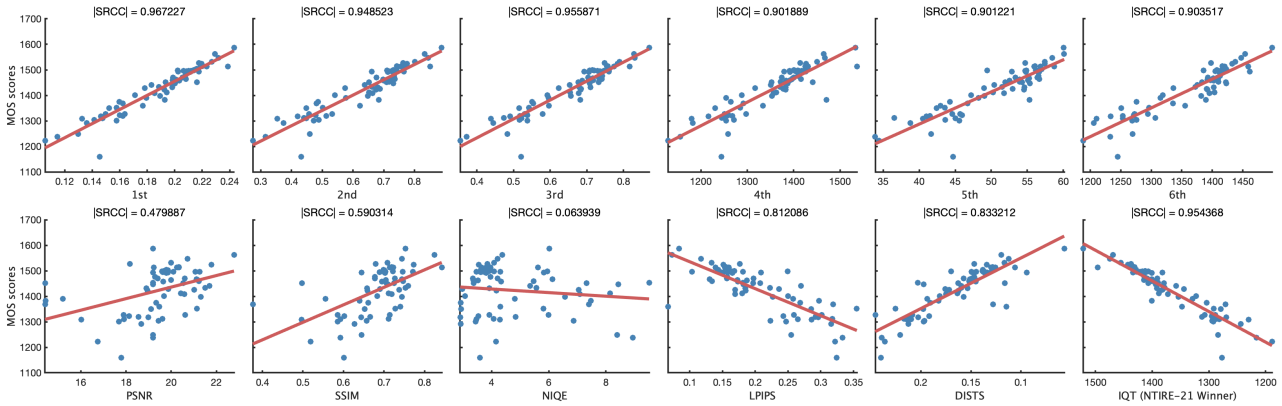

Figure 3. Analysis of FR-IQA methods in evaluating IR methods. Each point represents an algorithm. Higher correlations indicates better performance in evaluating perceptual image algorithm.

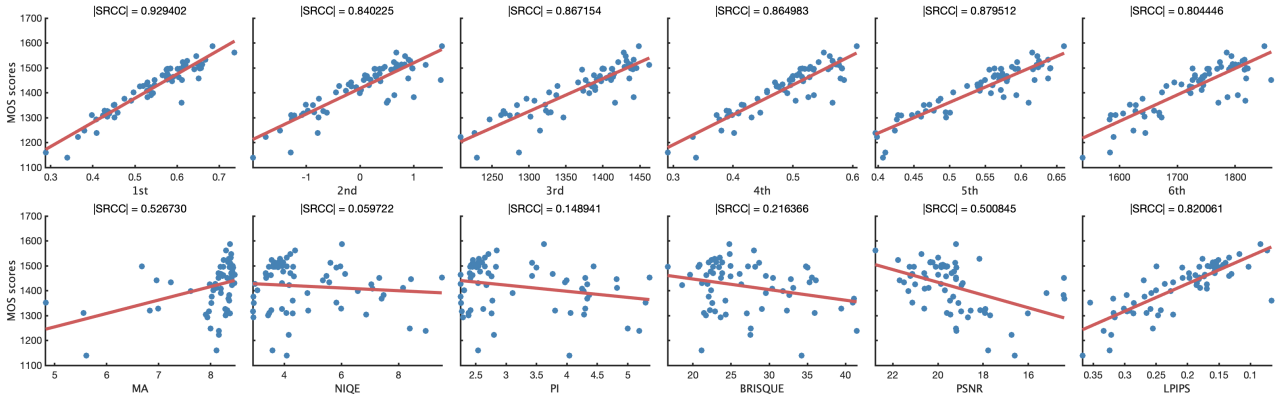

Figure 4. Analysis of NR-IQA methods in evaluating IR methods. Each point represents an algorithm. Higher correlations indicates better performance in evaluating perceptual image algorithm.

Challenge phases. The whole challenge consists of three phases: the developing phase, the validation phase, and the testing phase. In the developing phase, the participants can access to the reference and distorted images of the training set and also the MOS labels. This period is for the participants to familiarize themselves with the structure of the data and develop algorithms. In the validation phase, the participants can access the reference and distorted images of the training set and no labels are provided. The participants had the opportunity to test their solutions on the validation images and to receive immediate feedback by uploading their results to the server. A validation leaderboard is available. In the testing phase, the participants can access to the reference and distorted images of the training set. A final predicted perceptual similarity result is required before the challenge deadline. The participants also need to submit the executable file and a detailed description file of the proposed method. The final results were then made available to the participants.

4. Challenge Results

There are 8 and 7 teams participated in the testing phase of the challenge for the track 1 and track 2, respectively. Table 1 reports the main results and important information of these teams. The methods are briefly described in Section 5 and the team members are listed in Appendix B and Appendix C. We next analyze each track's result separately

4.1. Track 1: Full-Reference IQA Track

This is the second full-reference IQA challenge. In the last challenge, IQT [13] won the championship on this track using a transformer as the network backbone. This year, we use a more complex validation and testing dataset. According to the results in Table 1, we can see that this year's submitted methods have generally achieved comparable results. All valid entries achieved higher correlation performance than methods such as LPIPS [74], which are now widely used. Three teams surpassed last year's championship method. The champion team achieves an SRCC score of 0.823 and a PLCC score of 0.828, refreshing the state-of-the-art performance on PIPAL. Figure 1 shows the

scatter distributions of subjective MOS scores vs. the predicted scores by the top solutions and the other 7 IQA metrics on the PIPAL test set. The curves shown in Figure 1 were obtained by a third-order polynomial nonlinear fitting. One can observe that the objective scores predicted by the top solutions have higher correlations with the subjective evaluations than existing methods. In Figure 3, we show the scatter plots of subjective scores vs. the top solutions and some commonly-used IQA metrics for perceptual-oriented algorithms. Recall that an important goal of this challenge is to promote more promising IQA metrics for perceptual-oriented algorithms. As can be seen, the top solutions generally perform better in evaluating the images in the testing set. Among them, the correlation between the evaluation of the champion solution and the subjective score reaches 0.967, which surpasses the champion's performance of the last year.

4.2. Track 2: No-Reference IQA Track

It is the first time an NTIRE challenge focuses on no-reference IQA. NR-IQA is an indispensable part of algorithm evaluation, but widely used algorithms only show very limited performance on our test set. This year, we include this track to push the developing state-of-the-art NR-IQA methods to fill this gap. According to the results in Table 1, one can observe that all the valid entries surpass the current state-of-the-art performance, and some even achieve correlation performance comparable to FR methods. Figure 2 shows the scatter distributions of subjective MOS scores vs. the predicted scores by the top solutions and the other 3 NR-IQA metrics. The curves show compatible conclusions. In Figure 4, we show the scatter plots of subjective scores vs. the top solutions and some commonly-used IQA metrics for perceptual-oriented algorithms. It can be seen that the existing NR-IQA methods are not ideal in evaluating algorithms. The works produced in this challenge received high correlation scores, which means that these methods are closer to human judgment when used to evaluate images generated by perceptual-oriented algorithms. This also suggests that using these NR methods as metrics can lead to more visually friendly results. Among them, the correlation between the evaluation of the champion solution and the subjective score reaches 0.92, which greatly improves the practical value of NR-IQA as an algorithm metric.

5. Challenge Methods

We describe the submitted solution details in this section.

5.1. Track 1: Full-Reference IQA Track

5.1.1 THU1919Group

Team THU1919Group is the winner of the first track. They develop an Attention-based Hybrid Image Quality assessment network (AHIQ) to participate in the FR-IQA track. As is shown in Figure 5, their network takes pairs of reference images and distortion images as input and consists of three key components: the feature extraction module, the feature fusion module and the pixel pooling module. For the feature extraction module, the input pairs of reference images and distortion images first go through a vision transformer backbone ViT [18] and a convolution network (CNN) [31] for feature extraction. The convolution network is used to retain more spatial information, and the transformer network captures global semantic features. In the feature fusion module, the feature maps from later stages of ViT are used to obtain an offset map for deformable convolution. A simple 2-layer convolution network is used to project the feature after deformable convolution. In this way, features from the early stages of CNN can be better modified and utilized for further feature fusion. At last, they propose the pixel pooling module to assess the quality of distorted images. The pixel pooling module contains two branches. The first branch calculates scores for each pixel, and the second branch calculates the weight of each pixel score to the final evaluation score. By weighting all the pixel scores, the final score can be obtained.

Their network contains 140 million parameters. For optimization, they use the AdamW [40] optimizer with an initial learning rate(LR) of 10^{-4} and weight decay of10^{-5}. The minibatch size is 8. During testing, different backbone networks and fusion approaches are used for ensemble, such as models at different training epochs, directly concatenate the feature maps from CNN and ViT without deformable convolution, and use inception-resnetV2 as feature extraction and use ResNet152 as feature extraction.

5.1.2 Netease OPDAI

Team Netease OPDAI wins second place in the first track. They build their method based on the winner solution of the last year – the IQT network [13]. The difference is they concatenate the reference image and the distorted image instead of the difference operation. They also introduce central difference convolution and Siamese network structure to extract features of the distorted image and reference image, respectively. They use Swin [39] transformer as regression layer. Finally, they incorporate a residual network using the spatial gradient module.

For optimization, in addition to the conventional MSE loss, they also learn the distribution of quality scores by introducing the Kullback–Leibler scatter loss, and norm-in-

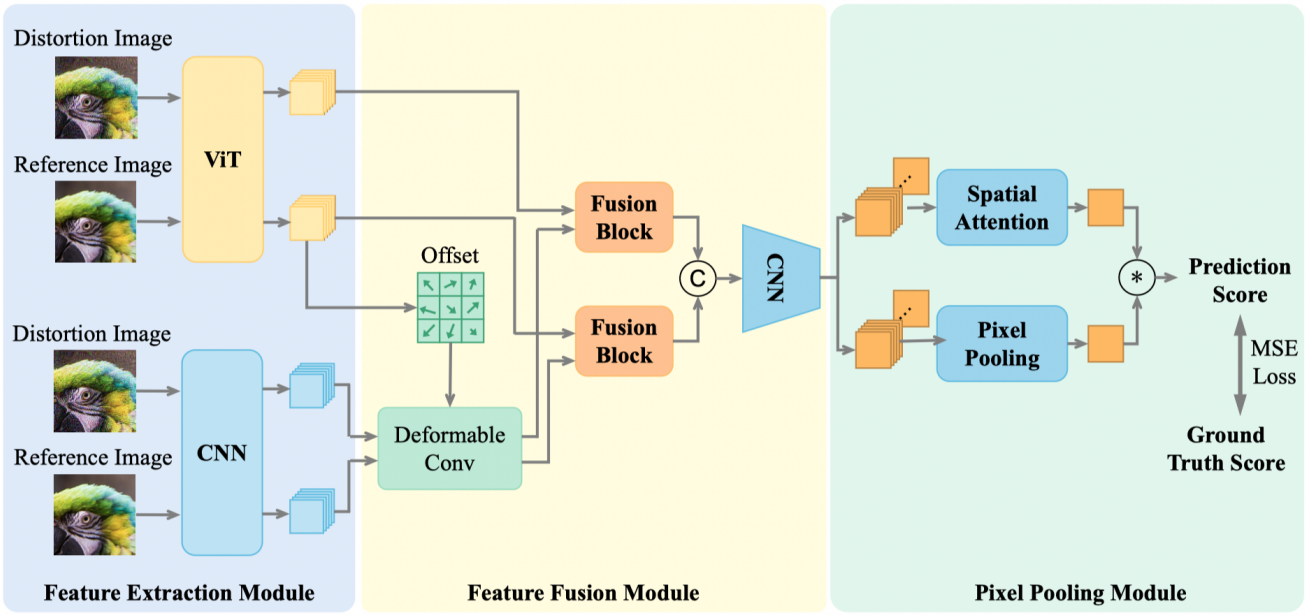

Figure 5. The overview of THUIIGROUP1919 team's Attention-based Hybrid Image Quality assessment network (AHIQ).

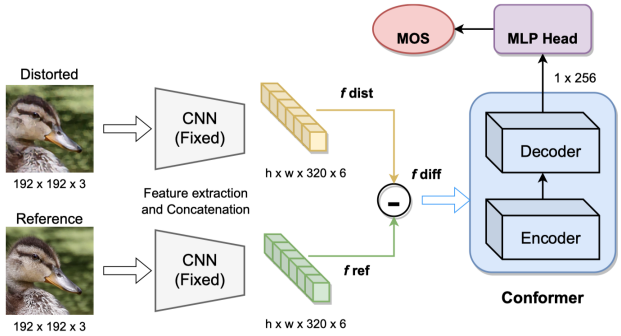

Figure 6. The network design of JMU-CVLab team's method.

norm loss [36]. They also introduce three data enhancement methods to help training. First, they dynamically optimize the frequency of difficult samples depending on the model fitting situation. Second, the model must adapt to random color space variations to improve generalization. And third, different positions of the distorted images are stitched together to increase the complexity of the dataset and to improve the robustness of the model. Their method has more than 276 million parameters.

5.1.3 KS

Team KS wins third place in the first track. They apply a bag of tricks for the IQA method with newly-appeared distortions. Firstly, they designed a new Multi-Scale Image Quality Network, called MSIQ-Net, to fully capture spatial distributions of distortion characteristics. MSIQ-Net takes a pair of distortion and reference images as input and generates multi-scale features using an FPN-like module. Among different scales, local texture distortion (*e.g.*, noise and blocking artifact) can be captured by low-level features, while global distributed distortion (*e.g.*, strange artifacts generated by GAN) can be learned from high-level features. A smooth module, which aggregates features adaptively, is also attached for the final representation.

They attach great importance to data processing during training. They discover that the labels of the training set have an unbalanced distribution. To prevent the model from being biased, they reconstruct the training data. First, for images with low/high MOS scores, they perform data augmentations (*e.g.*, horizontal flip, rotation) and increase the number of these images in the training set. Second, they follow the way in PIPAL and select high-frequency patches from SPAQ [21]. They then generate pseudo labels for these patches using a model trained on the given data. These patches, along with pseudo-labels, are added to the training process. Third, they apply a histogram equalization on labels and obtain an even distribution.

For the loss functions, in addition to the commonly used MSE loss, they also employ an extra PLCC-induced loss [36]. Assume we have N images in the training batch. Given the predicted quality scores $Y' = \{y'_1, \ldots, y'_N\}$ and the subjective quality scores $Y = \{y_1, \ldots, y_N\}$, the loss is defined as

$$\mathcal{L}_{plcc} = \left(1 - \frac{\sum_{i=1}^{N}(y'_i - a')(y_i - a)}{\sqrt{\sum_{i=1}^{N}(y'_i - a')^2 \sum_{i=1}^{N}(y_i - a)^2}}\right) * 0.5,$$

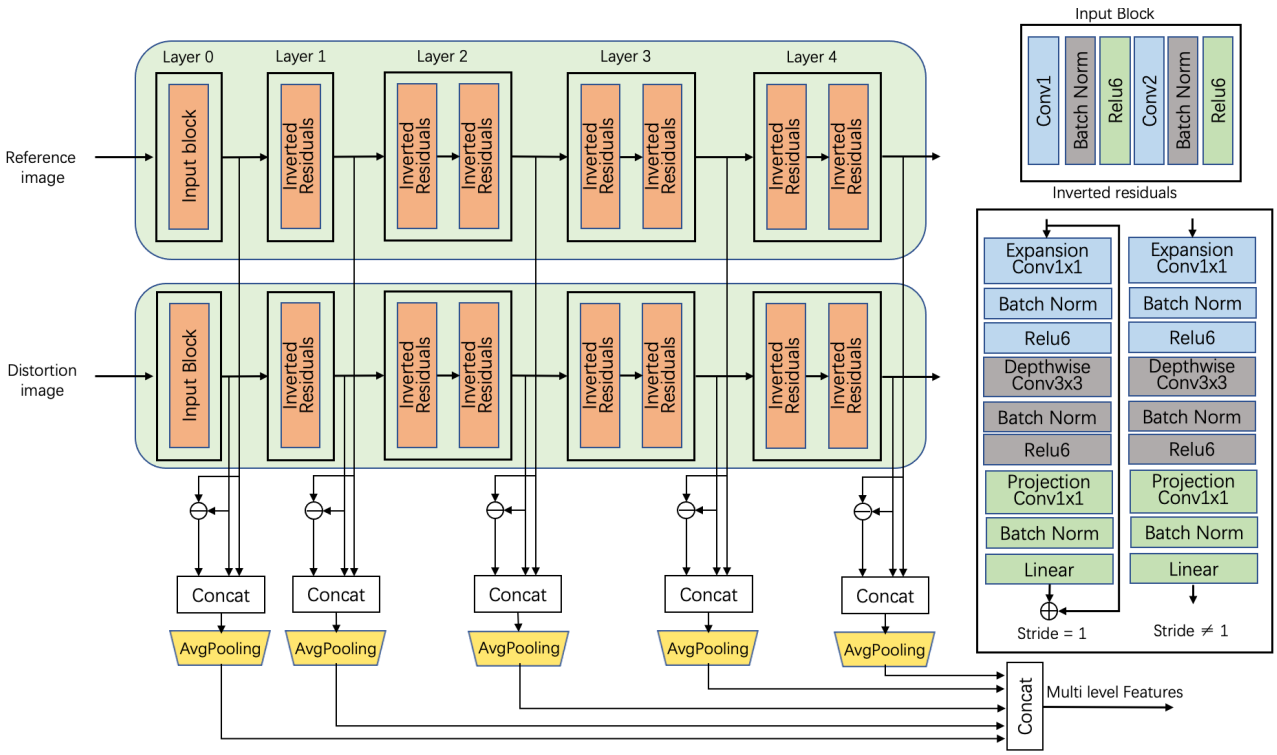

Figure 7. The framework design for feature extraction of Yahaha! team's solution.

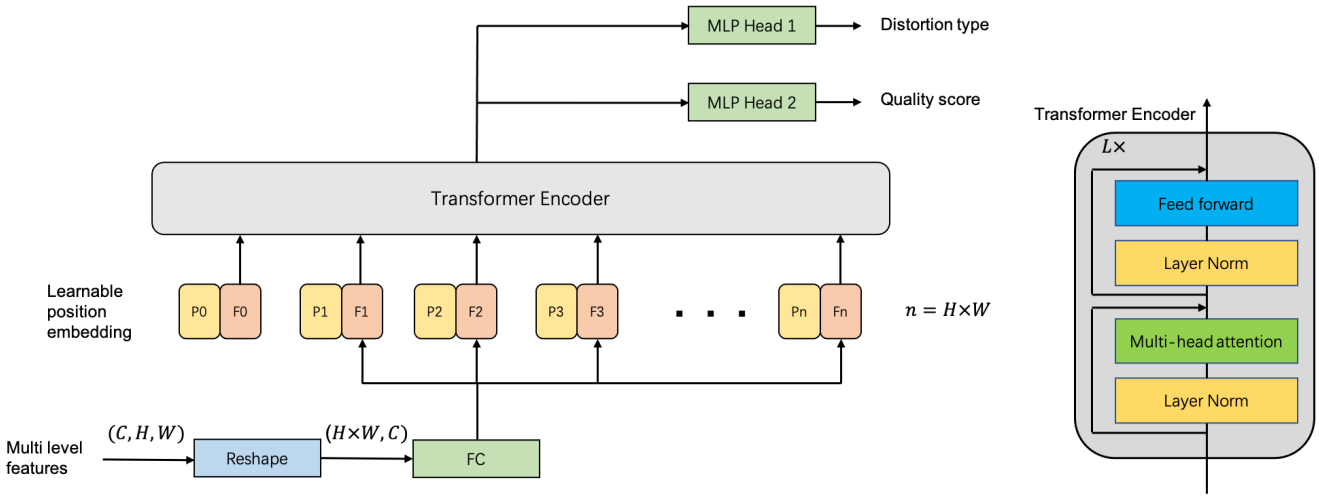

Figure 8. The framework of transformer of Yahaha! team's solution.

where a' and a are the mean values of Y' and Y, respectively.

The KS team also adopt the model ensemble strategy. Three main methods are used. First, by directly averaging the weights of multiple models trained with different hyper-parameters, the IQA model improves accuracy and robustness. Second, due to the model capacity and divergence, a single model may not make the perfect predictions for a given dataset, suffering from specific noise or bias. The combination can be implemented by averaging the output of each model. A weighted combination will also do the job, whose weights can come from linear regression or other schemes. Third, various augmentation types, which do not inflect the quality of images, are used for repeat predictions, including multi-crop, horizontal flip and random rotation. The average score of multiple views is used for the final prediction.

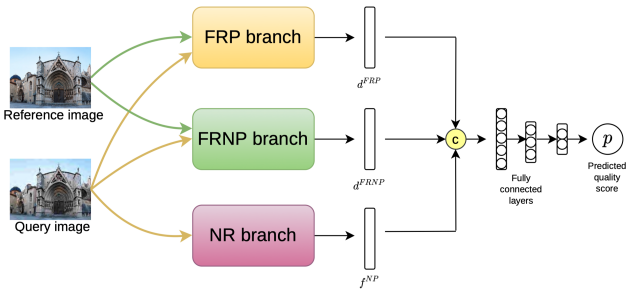

Figure 9. The framework design of the Horizon team's method.

5.1.4 JMU-CVLab

Team JMU-CVLab [15] proposes an IQA Conformer Network by improving the IQT [13] architecture. They use Inception-ResNet-v2 [57] network pre-trained on ImageNet [16] to extract reference and distorted images feature maps. The network weights are kept frozen, and a Conformer encoder-decoder is trained to regress MOS using the MSE loss. In their network, the `mixed5b`, `block35_2`, `block35_4`, `block35_6`, `block35_8` and `block35_10` feature maps are concatenated for the reference and distorted images generating f_{ref} and f_{dist}, respectively. In order to obtain the difference information between reference and distorted images, a difference feature map, $f_{diff} = f_{ref} - f_{dist}$ is also used. Concatenated feature maps are then projected using a point-wise convolution but not flattened to preserve spatial information. They used a single Conformer block [11, 29] for both encoder and decoder. The model hyper-parameters are: $L = 1$, $D = 128$, $H = 4$, $D_{feat} = 512$, and $D_{head} = 128$. The input image size of the backbone model is set to $192 \times 192 \times 3$, which generates feature maps of size 21×21. Their network design is shown in Figure 6. They also adopt the ensemble method to improve the performance of the final method. Their final submission is an ensemble of the proposed IQA Conformer Network and two pre-trained models: RADN [54], and ASNA [3].

5.1.5 Yahaha!

Team Yahaha! proposes a transformer-based full-reference image quality assessment framework leveraging multi-level features. The reference image and distortion image are fed to the backbone network separately. Their feature maps from different layers and difference maps of corresponding feature maps are downsampled and concatenated together. Then all these maps from different levels are fed to transformer layers for the joint training of distortion type prediction and perceptual score regression. A Siamese-network is used for extracting deep features from the reference image and distorted image, as shown in Figure 7. A pruned MobileNetv2 is used in this stage, and feature maps from 4 layers of MobileNetv2 are extracted for further processing. Then the extracted features are fed to a Transformer network to predict the final score, the architecture of which is shown in Figure 8. The Transformer encoder contains multi Transform layers, each consisting of a standard multi-head attention module and a feedforward module. Besides, layer normalization is adopted. The Transformer encoder is connected with two different fully connected layers to predict distortion type and opinion score, respectively.

For the optimization of the proposed method, they use a loss function consisting of four loss functions: L1 loss, cross-entropy loss, norm-in-norm loss [35] and relative-distance loss. The number of parameters for their proposed model is 68.853K. They also adapt ensemble strategy. The full training set is divided into five parts. These five parts are used as the validation set, while the other four parts are used for training. After obtaining five models, the submitted prediction scores for development and test are the average prediction results of these five models.

5.1.6 debut_kele

The debut_kele team's solution can be divided into two main parts: feature extraction and regression modelling. For the features, they extract different perceptual image quality metrics, which include SSIM, MS SSIM, CW-SSIM, GMSD, LPIPSvgg, DISTS, NLPD, FSIM, VSI, VIFs, VIF, MAD. Using gradient boosting trees, a regression model is built based on the pre-calculated IQA results. The model's 'max depth' parameter is set to 3, and the learning rate was set to 0.01. In addition, the feature subsample and the sample subsample values were set at 0.7 to prevent overfitting. The maximum iteration round is 10000, while the early-stopping round is set at 500. In their experiments, VIFs and VIF play a critical role than LPIPSvgg. Five-fold bagging-based ensemble strategy is used in their experiments.

5.1.7 Team Horizon

The Horizon team propose a Multi-branch Image Quality Assessment Network that consists of three parallel branches: (1) the full-reference pre-trained (FRP) branch, (2) the full-reference non-pretrained (FRNP) branch and (3) the no-reference (NR) branch. Both distorted and reference images are fed as input for the full-reference branches (FRP and FRNP), whereas in the no-reference branch, only the distorted image is provided as input. In each branch, they use a convolutional neural network-based encoder. In the FRP branch, they use an encoder from a classifier trained on ImageNet since these features are known to correlate well with perceptual quality. The weights of the FRP encoder are kept fixed throughout the training. They didn't use pre-trained encoders in FRNP and NR branches to enable the learning of discriminative features from the training data.

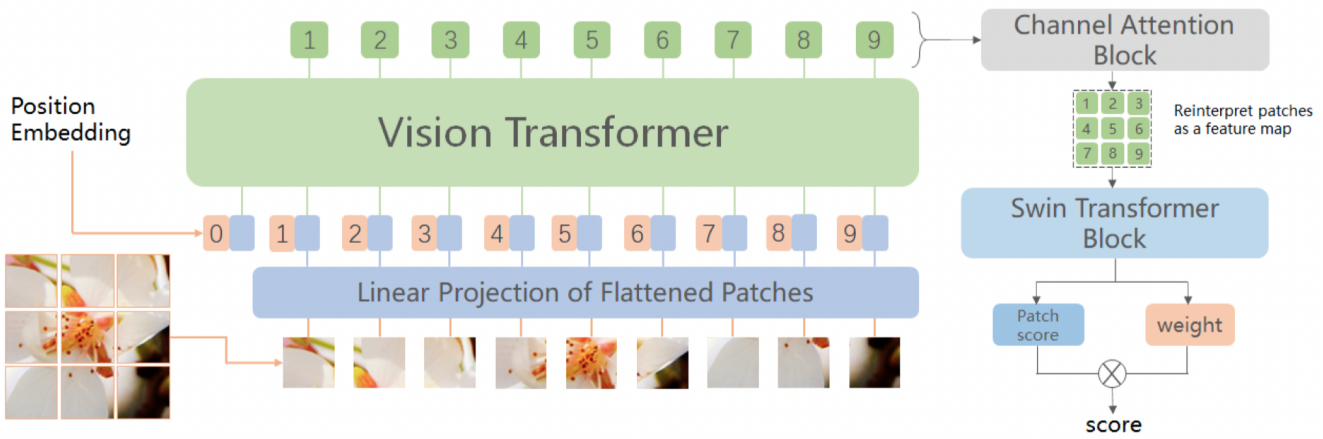

Figure 10. The model framework overview of the THU_IIGROUP team.

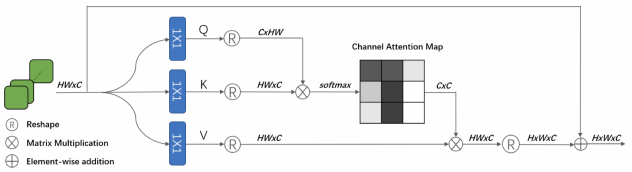

Figure 11. The Channel attention block used in THU_IIGROUP team's network.

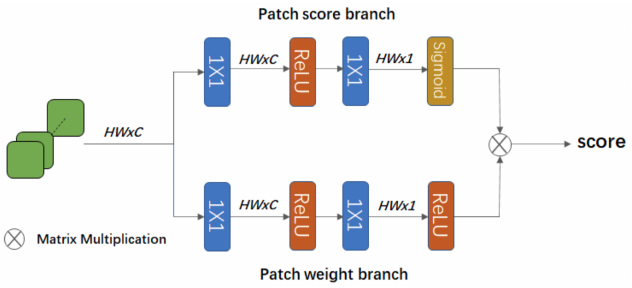

Figure 12. The Patch Weighted Branch used in THU_IIGROUP team's network.

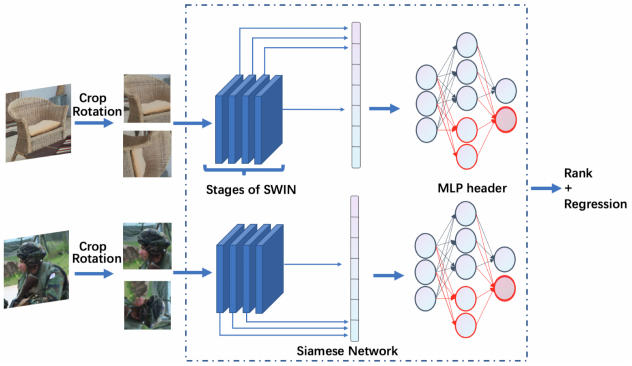

Figure 13. The Multi Stage fusing SWIN Transformer based on Siamese Network design proposed by team DTIQA.

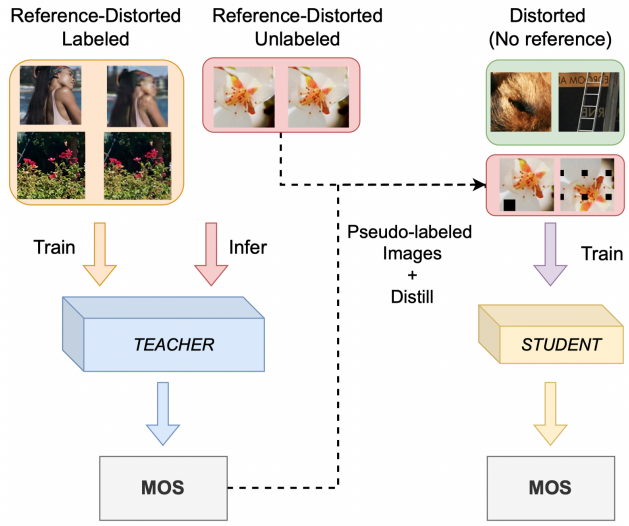

Figure 14. The blind noisy student setup proposed by team JMU-CVLab.

Full-reference branches extract features from both distorted and reference images and compute pixel-wise differences between distorted and reference features. But when the distortion is severe, computing pixel-wise difference, even in feature space, may not be optimal due to spatial misalignment. Hence, a No-reference branch focuses only on features related to the distortion present in the query image. The framework of the proposed method is shown in Figure 9.

5.2. Track 2: No-Reference IQA Track

5.2.1 THU_IIGROUP

Team THU_IIGROUP is the winner of the second track. In their method, they first cut of the image with a small size generates the global interaction between different regions

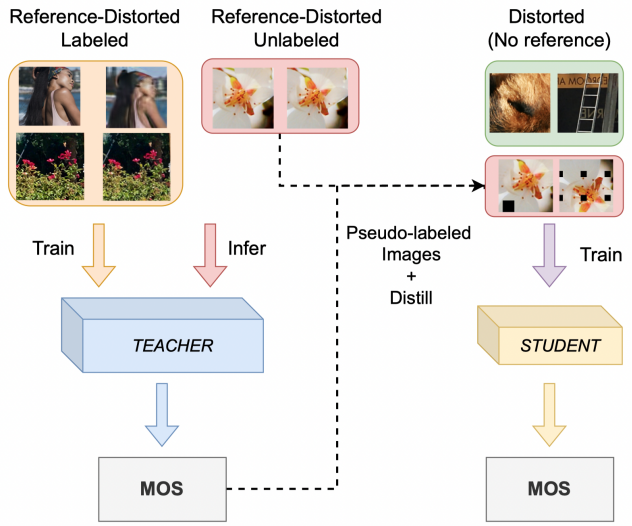

Figure 15. The blind noisy student setup proposed by team JMU-CVLab.

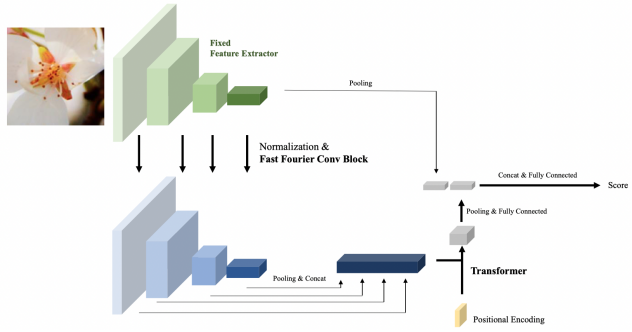

Figure 16. The method of an anonymous team.

of an image. It utilizes the pre-trained vision transformer [18] as the feature extractor to attain the feature of different patches from different positions. They select four layers of the vision transformer output as the main features. Second, to handle the difference among four layers, a channel attention block [68] is used to adjust the distribution of features from these four layers. Then, to reinforce the local connection with image patches, a swin transformer block [39] is applied to handle the perceptual quality of each patch. Finally, a regression head with two branches – patch score branch and weight branch [10] – is used to predict the quality score and the importance of each patch. The final score of the distorted image is generated by a multiplication. The overview of the method is shown in Figure 10, the channel attention block is shown in Figure 11, and the final patch weighted branch is shown in Figure 12. They also employ a bagging ensemble method. Three same models are trained for ensembles. (1) The model M1 without finetuning. (2) The model M2 finetune twice. (3) The model M3 finetune only once. The weight of M1, M2, and M3 are 0.25, 0.55 and 0.2.

5.2.2 DTIQA

Team DTIQA wins second place in track 2. Their main contributions can be divided into three parts. Firstly, they use a multi-stage Swin Transformer as the baseline and fine-tune the model on the PIPAL dataset. Secondly, several training tricks are used to improve the performance. These tricks include the loss function, data augmentation and test time augmentation. For the loss function, both ranking loss and regression loss are used. The regression loss is the Euclidean loss (MSE). The ranking loss explicitly exploits the relative ranking of image pairs available in the PIPAL dataset. The loss is formulated as:

$$\mathcal{L}_{rank} = \frac{2}{N}\sum_{i=0}^{N} e^{\hat{y}^{2i}-\hat{y}^{2i+1}}\mathbb{I}\{y^{2i} < y^{2i+1}\}.$$

For the training data augmentation, they use (1) random cropping image into patches, (2) random rotation, (3) weighted sampler to ensure that each batch sees a proportional number of mos scores and (4) colorspace augmentation. For the testing augmentation, they employ (1) random crop, then infer each image 80 times and get a harmonic mean score and (2) resize the image to 256 with different interpolation methods. Finally, they designed a model that combines Transformer and CNN for multiple resolutions. The framework of the proposed method is shown in Figure 13. As input to the Swin Transformer, they used the flattened output of several blocks of EfficientNet. Embedding and projection locations were also added to match the Transformer dimension and expanded with dummy tokens representing aggregated information.

5.2.3 JMU-CVLab

Team JMU-CVLab wins third place in track 2. In their method [15], a simple CNN backbone ϕ takes a distorted image x as input and aims to minimize the MOS y using the following loss function from [4], where $\phi(x) = y'$.

$$\mathcal{L} = \mathrm{MSE}(y, y') + (1 - \mathrm{Pearson}(\mathrm{y}, \mathrm{y}'))$$

They argue that the overfitting problem is the main problem in this track because there are only 200 reference images for 23200 distorted images. They use several augmentation methods: Horizontal and vertical flips. Rotations of 90/180/270 degrees. Take a random crop of size (224, 224). One of CutOut or GridMask as further regularization to ensure the model learns to assess the quality without looking at the entire image. The main trick in their method is called Noisy Student Training, a semi-supervised learning

approach that extends the idea of self-training and distillation. They distinguish a teacher model trained with full-reference pairs and a student model trained only with distorted images. The process is as follows: (1) train a teacher model using the PIPAL dataset with reference and distorted pairs, (2) infer on unlabelled samples and annotate the images – these MOS annotations are noisy and called pseudo-labels, (3) add the pseudo-labelled samples to the training set, and (4) train a student model that takes the distorted images as input and trained on the extended datasets.

The organizers note that the new data included in the solution of team JMU-CVLab contain the validation images from the NTIRE IQA challenge (2021 and 2022). These data do not contain the final test data, but the distortion type of these data is more similar to the distortion of the test data. We note that this method has the possibility of not generalising well to other distortion types.

5.2.4 NetEase OPDAI

Team NetEase OPDAI uses Swin Transformer [39] as the backbone. The Swin Transformer is pretrained on ImageNet [16] dataset. The network can extract more expressive features compared to Resnet [31]. The training uses MSE loss and KL-divergence loss. They also use the ensemble method to improve the final results. Two models are used: (1) Swin Transformer Large-224 as the backbone, 2-layer transformer layer to predict MOS score; data augmentation including flipping, cropping, and resizing, and (2) Swin Transformer Tiny-224 as the backbone, 1-layer transformer layer to predict MOS score; data augmentation including flipping, cropping and resizing. In the model (2), they remove the extremely hard distortion type. The results of these two models are then averaged as the final result.

5.2.5 Withdrawn submission

This team withdrew their submission and remains anonymous in this report. Their method is shown in Figure 16 summarized as follow. They build an NR-IQA model based on the model structure and loss of Transformers, Relative Ranking, and Self-Consistency (TReS) [23]. They use a pre-trained ResNeXt 101 model [64] on ImageNet [16] as a feature extractor and fixed it during the IQA training process. They use fast fourier convolution (FFC) [56, 14] so that the model can employ both global and local contexts of the image for quality assessment. FFC uses spectral transform as well as 2D convolutions to increase the receptive field size and to learn global context as well. The feature map of each level calculated by the feature extractor goes through normalization and FFC blocks. Then, these feature maps are combined via pooling and concatenation, passed through a transformer with positional encoding, and then passed through fully-connected layers to be a score value. For training, they use the L1 loss as the score difference loss when training the model. In addition, a self-consistency loss is used that allows similar feature vectors to be extracted for the rotated image and a relative ranking loss that considers sample ranking in a mini-batch which are used in TReS as well. They add an extra loss that narrows the difference between the predicted score difference and the ground truth score difference for every pair in the mini-batch. An Adam optimizer with a learning rate of 2×10^{-5} is used, and the learning rate is halved after every ten epochs.

5.2.6 NTU607QCO-IQA

The method of team NTU607QCO-IQA is also adapted from [23]. This model contains a backbone to extract multi-scale features and a linear layer neck to integrate features and output the final scores. Based on this architecture, they use the Res2Net [22] as the backbone. Furthermore, in the loss function part, they not only apply the L1 loss but also Pearson's correlation loss [3] and triplet loss. The Pearson's correlation loss is helpful for the model to increase the performance of PLCC. The triplet loss is used for surrogate ranking. They use Adamw optimizer with the learning rate of 0.0001 and the learning rate decrease of 0.1 every ten epochs. The total epoch is 50 and takes 11 hours. The batch size is 10. No data augmentation is used in the training phase.

Acknowledgments

We thank the NTIRE 2022 sponsors: Huawei, Reality Labs, Bending Spoons, MediaTek, OPPO, Oddity, Voyage81, ETH Zürich (Computer Vision Lab) and University of Würzburg (CAIDAS).

A. NTIRE 2022 Organizers

Title:
NTIRE 2022 Challenge on Perceptual Image Quality Assessment
Members:
Jinjin Gu[1,2] *(jinjin.gu@sydney.edu.au)*, Haoming Cai[3], Chao Dong[1,3], Jimmy S. Ren[4], Radu Timofte[5,6]
Affiliations:
[1] Shanghai AI Lab, Shanghai, China
[2] School of Electrical and Information Engineering, The University of Sydney
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[4] SenseTime Research
[5] University of Würzburg, Germany
[6] ETH Zürich, Switzerland

B. Track 1: Teams and Affiliations

THU1919Group

Title:
Attention Helps CNN See Better: Hybrid Image Quality Assessment Network
Members:
Yuan Gong[1] *(gong-y21@mails.tsinghua.edu.cn)*, *Shanshan Lao*[1] *(laoss21@mails.tsinghua.edu.cn)*, Shuwei Shi[1], Jiahao Wang[2], Sidi Yang[1], Tianhe Wu[1], Weihao Xia[3], Yujiu Yang[1]
Affiliations:
[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Department of Automation, Tsinghua University
[3] University College London

Netease OPDAI

Title:
An Algorithm for Reference Image Quality Assessment
Members:
Cong Heng[1] *(congheng@corp.netease.com)*, Lingzhi Fu[1], Rongyu Zhang[1], Yusheng Zhang[1], Hao Wang[1], Hongjian Song[1]
Affiliations:
[1] Netease

KS

Title:
Bag of Tricks for Practical Image Quality Assessment with Newly-appeared Distortions
Members:
Ming Sun[1] *(sunming03@kuaishou.com)*, Mading Li[1], Kai Zhao[1], Kun Yuan[1], Zishang Kong[1], Mingda Wu[1], Chuanchuan Zheng[1]
Affiliations:
[1] Kuaishou

JMU-CVLab

Title:
IQA Conformer Network
Members:
Maxime Burchi[1] *(marcos.conde-osorio@uni-wuerzburg.de)*, Marcos V. Conde[1]
Affiliations:
[1] University of Wurzburg, Computer Vision Lab

Yahaha!

Members:
Longtao Feng[1] *(fenglongtao@bytedance.com)*, Tao Zhang[1], Yang Li[1], Jingwen Xu[1], Haiqiang Wang[1], Yiting Liao[1], Junlin Li[1]
Affiliations:
[1] ByteDance

debut_kele

Title:
Gradient Boosting Trees -Based Perceptual Image Quality Assessment
Members:
Kele Xu[1] *(kelele.xu@gmail.com)*, Tao Sun[1], Yunsheng Xiong[1]
Affiliations:
[1] Key Laboratory for Parallel and Distributed Processing

Pico Zen

Title:
Multi Scale Image Quality Assessment with Transformers by weight sharing approach
Members:
Abhisek Keshari[1] *(2018ume0126@iitjammu.ac.in)*, Komal[1], Sadbhawana Thakur[1], Vinit Jakhetiya[1], Badri N Subudhi[1]
Affiliations:
[1] Indian Institute of Technology Jammu

Team Horizon

Title:
Multi-branch Image Quality Assessment Network
Members:
Saikat Dutta[1] *(saikat.dutta779@gmail.com)*, Sourya Dipta Das[2], Nisarg A. Shah[3], Anil Kumar Tiwari[3]
Affiliations:
[1] Indian Institute of Technology Madras
[2] Jadavpur University
[3] Indian Institute of Technology Jodhpur

C. Track 2: Teams and Affiliations

THU_IIGROUP

Title:
Multi-Dimension Attention Network for Image Quality Assessment
Members:
Sidi Yang[1] *(yangsd21@mails.tsinghua.edu.cn)*, *Tianhe Wu*[1]

(tianhe wu@foxmail.com), Shuwei Shi[1], Shanshan Lao[1], Yuan Gong[1], Mingdeng Cao[1], Jiahao Wang[1], Yujiu Yang[1]
Affiliations:
[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Department of Automation, Tsinghua University

DTIQA

Title:
Assessment Method Based on Transformer and Convolutional Neural Network
Members:
Jing Wang[1] *(wangjing.crystalw@bytedance.com)*, Haotian Fan[1], Xiaoxia Hou[1]
Affiliations:
[1] ByteDance

JMU-CVLab

Title:
Blind and Noisy IQA Students
Members:
Marcos V. Conde[1] *(marcos.conde-osorio@uni-wuerzburg.de)*, Maxime Burchi[1]
Affiliations:
[1] University of Würzburg, Computer Vision Lab

KS

Title:
Bag of Tricks for Practical Image Quality Assessment with Newly-appeared Distortions
Members:
Ming Sun[1] *(sunming03@kuaishou.com)*, Mading Li[1], Kai Zhao[1], Kun Yuan[1], Zishang Kong[1], Mingda Wu[1], Chuanchuan Zheng[1]
Affiliations:
[1] Kuaishou

NetEase OPDAI

Members:
Heng Cong[1] *(congheng@corp.netease.com)*, Lingzhi Fu[1], Rongyu Zhang[1], Yusheng Zhang[1], Hao Wang[1], Hongjian Song[1]
Affiliations:
[1] Netease

NTU607QCO-IQA

Title:
Multiple loss functions for no reference perceptual image quality assessment
Members:
Hao-Hsiang Yang[1] *(islike8399@gmail.com)*, Hua-En Chang[1], Zhi-Kai Huang[1], Wei-Ting Chen[1], Sy-Yen Kuo[1]
Affiliations:
[1] National Taiwan University

References

[1] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, et al. NTIRE 2022 spectral demosaicing challenge and dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[2] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, et al. NTIRE 2022 spectral recovery challenge and dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[3] Seyed Mehdi Ayyoubzadeh and Ali Royat. (asna) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 388–397, 2021. 9, 12

[4] Seyed Mehdi Ayyoubzadeh and Ali Royat. (asna) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 11

[5] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009. 4

[6] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 burst super-resolution challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[7] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018. 2

[8] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 2

[9] 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(1):206–219, 2017. 2

[10] 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(1):206–219, 2017. 2, 11

[11] Maxime Burchi and Valentin Vielzeuf. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. *arXiv preprint arXiv:2109.01163*, 2021. 9

[12] Haoming Cai, Jingwen He, Yu Qiao, and Chao Dong. Toward interactive modulation for photo-realistic image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 294–303, June 2021. 2

[13] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2, 5, 6, 9

[14] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 12

[15] Marcos V. Conde, Maxime Burchi, and Radu Timofte. Conformer and blind noisy students for improved image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 9, 11

[16] 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. Ieee, 2009. 9, 12

[17] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6, 11

[19] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978. 3

[20] Egor Ershov, Alex Savchik, Denis Shepelev, Nikola Banic, Michael S Brown, Radu Timofte, et al. NTIRE 2022 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[21] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 7

[22] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 12

[23] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 12

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[25] 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. *arXiv preprint arXiv:2011.15002*, 2020. 1, 2, 4

[26] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S Ren, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 4

[27] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy 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. 2

[28] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Yu Qiao, Shuhang Gu, and Radu Timofte. Ntire 2021 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 677–690, 2021. 1, 3

[29] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 9

[30] Haiyang Guo, Yi Bin, Yuqing Hou, Qing Zhang, and Hengliang Luo. Iqma network: Image quality multi-scale assessment network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[31] 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. 6, 12

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 2

[33] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 2

[34] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[35] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 789–797, 2020. 9

[36] Dingquan Li, Tingting Jiang, Ming Jiang, Vajira Lasantha Thambawita, and Haoliang Wang. Reproducibility companion paper: Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3615–3618, 2021. 7

[37] Yawei Li, Kai Zhang, Radu Timofte, Luc Van Gool, et al. NTIRE 2022 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[38] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–741, 2018. 2

[39] 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. 6, 11, 12

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[41] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 challenge on learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[42] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 2

[43] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2

[44] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 2

[45] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 2

[46] Zhaoqing Pan, Feng Yuan, Xu Wang, Long Xu, Shao Xiao, and Sam Kwong. No-reference image quality assessment via multi-branch convolutional neural networks. *IEEE Transactions on Artificial Intelligence*, 2022. 2

[47] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Ales Leonardis, Radu Timofte, et al. NTIRE 2022 challenge on high dynamic range imaging: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[48] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 1, 4

[49] 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*, pages 1808–1817, 2018. 2

[50] Andres Romero, Angela Castillo, Jose M Abril-Nova, Radu Timofte, et al. NTIRE 2022 image inpainting challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[51] 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. 2

[52] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005. 2

[53] 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. 3, 4

[54] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2, 9

[55] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 2

[56] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 12

[57] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 9

[58] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 2

[59] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2022 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[60] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 2

[61] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision*, pages 63–79. Springer, 2018. 1, 2

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[63] 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, 2003. 2

[64] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 12

[65] Bo Yan, Bahetiyaer Bare, and Weimin Tan. Naturalness-aware deep no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(10):2603–2615, 2019. 2

[66] Ren Yang, Radu Timofte, et al. NTIRE 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[67] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012. 2

[68] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 11

[69] 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. 2

[70] 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. 2

[71] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 2

[72] 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. 2

[73] Peng Zhang, Wengang Zhou, Lei Wu, and Houqiang Li. Som: Semantic obviousness metric for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2394–2402, 2015. 2

[74] 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. 2, 5

[75] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. 2019. 1, 2

[76] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 2

[77] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Learning to blindly assess image quality in the laboratory and wild. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 111–115. IEEE, 2020. 2

[78] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 2