NTIRE 2021 Challenge on Perceptual Image Quality Assessment

Jinjin Gu   Haoming Cai   Chao Dong   Jimmy S. Ren   Yu Qiao   Shuhang Gu
Radu Timofte   Manri Cheon   Sungjun Yoon   Byungyeon Kangg Kang   Junwoo Lee
Qing Zhang   Haiyang Guo   Yi Bin   Yuqing Hou   Hengliang Luo   Jingyu Guo
Zirui Wang   Hai Wang   Wenming Yang   Qingyan Bai   Shuwei Shi   Weihao Xia
Mingdeng Cao   Jiahao Wang   Yifan Chen   Yujiu Yang   Yang Li   Tao Zhang
Longtao Feng   Yiting Liao   Junlin Li   William Thong   Jose Costa Pereira
Ales Leonardis   Steven McDonagh   Kele Xu   Lehan Yang   Hengxing Cai
Pengfei Sun   Seyed Mehdi Ayyoubzadeh   Ali Royat   Sid Ahmed Fezza
Dounia Hammou   Wassim Hamidouche   Sewoong Ahn   Gwangjin Yoon
Koki Tsubota   Hiroaki Akutsu   Kiyoharu Aizawa

Abstract

This paper reports on the NTIRE 2021 challenge on perceptual image quality assessment (IQA), held in conjunction with the New Trends in Image Restoration and Enhancement workshop (NTIRE) workshop at CVPR 2021. As a new type of image processing technology, perceptual image processing algorithms based on Generative Adversarial Networks (GAN) have produced images with more realistic textures. These output images have completely different characteristics from traditional distortions, thus pose a new challenge for IQA methods to evaluate their visual quality. In comparison with previous IQA challenges, the training and testing datasets in this challenge include the outputs of perceptual image processing algorithms and the corresponding subjective scores. Thus they can be used to develop and evaluate IQA methods on GAN-based distortions. The challenge has 270 registered participants in total. In the final testing stage, 13 participating teams submitted their models and fact sheets. Almost all of them have achieved much better results than existing IQA methods, while the winning method can demonstrate state-of-the-art performance.
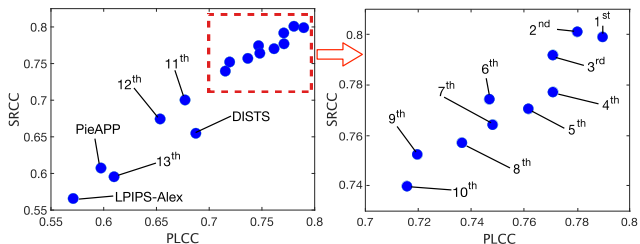
Figure 1. Quantitative comparison of IQA methods in the challenge. The right figure is the zoom-in view. SRCC represents Spearman rank order correlation coefficient and PLCC represents Pearson linear correlation coefficient. Higher coefficient matches perceptual score better. The top methods demonstrate the state-of-the-art performance.

1. Introduction

Image quality assessment (IQA) aims at using computational models to measure the perceptual quality of images, which are degraded during acquisition, compression, reproduction and post-processing operations. As the "evaluation mechanism", IQA plays a critical role in most image processing tasks, such as image super-resolution, denoising, compression and enhancement. Although it is easy for human beings to distinguish perceptually better images, it has been proved to be difficult for algorithms [38, 20]. Especially, on the basis of Generative Adversarial Networks (GANs) [18], perceptual image processing algorithms (or perceptual-oriented algorithms) [25, 30, 53, 62] have posed a great challenge for IQA methods, as they bring completely new characteristics to the output images [20]. It has been noticed that the contradiction between the quantitative evaluation results and the real perceptual quality is increasing [6, 7, 20]. This will also affect the development of image

processing algorithms, if the IQA methods cannot objectively compare their perceptual quality [7, 20]. Therefore, new IQA methods need to be proposed accordingly, to adapt new image processing algorithms.

The NTIRE 2021 challenge takes a step forward in benchmarking perceptual IQA – the task of predicting the perceptual quality of an image obtained by perceptual-oriented algorithms. We employ a new dataset called PIPAL [20] as our training set, which contains 200 reference images, 29k distorted images and 1.13M human judgements. Especially, this dataset includes the results of perceptual-oriented algorithms, which are missing in previous datasets. We also collect an extended dataset of PIPAL for validation and testing. This dataset contains 3,300 distorted images for 50 reference images, and all of them are the outputs of perceptual-oriented algorithms. We collect 753k human judgements to assign subjective scores for the extended images, ensuring the objectivity of the testing data.

The challenge has 270 registered participants in total. Among them, 13 participating teams have submitted their final solutions and fact sheets. They introduce new technologies in network architectures, loss functions, ensemble methods, data augmentation methods, and *etc*. The performance overview of these solutions is shown in Figure 1.

This challenge is one of the NTIRE 2021 associated challenges: nonhomogeneous dehazing [3], defocus deblurring using dual-pixel [1], depth guided image relighting [15], image deblurring [36], multi-modal aerial view imagery classification [31], learning the super-resolution space [32], quality enhancement of heavily compressed videos [56], video super-resolution [49], perceptual image quality assessment [21], burst super-resolution [5], high dynamic range [37].

2. Related Work

Image quality assessment (IQA). According to different usage scenarios, IQA methods can be divided in to full-reference methods (FR-IQA) and no-reference methods (NR-IQA). FR-IQA methods measure the perceptual similarity between two images, and have been widely used in the evaluation of image/video coding, restoration and communication quality. Beyond the most widely-used PSNR, FR-IQA methods follow a long line of works that can trace back to SSIM [54], which first introduces structural information in measuring image similarity. SSIM opened a precedent for the evaluation of image structure or features. After that, various FR-IQA methods have been proposed to bridge the gap between results of IQA methods and human judgements [55, 57, 59, 44, 58]. Similar to other computer vision problems, advanced data-driven methods have also motivated the investigation of applications of IQA. Zhang *et al*. [61] propose to use pre-trained deep networks to calculate the perceptual similarity and achieve good results. A

Figure 2. The difference between the traditional distortions (the first row) and the outputs of perceptual-oriented algorithms (the second row).

contemporaneous work [39] also propose to train deep IQA network using a pairwise-learning framework to predict the preference of one distorted image over the other. In addition to the above FR-IQA methods, NR-IQA methods are proposed to assess image quality without a reference image. Some popular NR-IQA methods include NIQE [35], [33], BRISQUE [34], and PI [7].

Perceptual-oriented and GAN-based distortion. In the past years, photo-realistic image generation has been evolving rapidly [30, 53, 52, 62], benefiting from the invention of perceptual-oriented loss function [25, 53] and GANs [18]. On the one hand, this kind of perceptual image restoration algorithm greatly improves the perceptual effect of the output image. On the other hand, it brings 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. An example of perceptual distortions is shown in Figure 2. The quality evaluation of such images has been proved challenging for IQA methods [20]. In order to evaluate and improve the performance of the IQA method against such perceptual distortions, Gu *et al*. [20] contribute a new IQA dataset called Perceptual Image Processing ALgorithms dataset (PIPAL), including the results of Perceptual-oriented image processing algorithms, which are missing in previous datasets. Recently, Gu *et al*. [19] propose to improve the IQA performance on these per-

Table 1. Quantitative results for the NTIRE 2021 Perceptual IQA challenge.

Rank	Team Name	Author/Method	Main Score	PIPAL		TID2013		LIVE	
				SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
1	LIPT	mrc	1.5885	0.7990	0.7896	0.8040	0.8440	0.9170	0.8970
2	MT-GTD	binyi	1.5811	0.8009	0.7803	0.7815	0.8265	0.9191	0.9132
3	The Amaurotia	gjy19	1.5625	0.7918	0.7707	0.7293	0.7931	0.9084	0.9074
4	THUIIGROUP1919	bqy2020	1.5480	0.7770	0.7709	0.7465	0.7960	0.9051	0.8777
5	Yahaha!	sherlocky	1.5317	0.7703	0.7615	—	—	—	—
6	Huawei Noah's Ark	wth	1.5212	0.7744	0.7468	—	—	—	—
7	debut_kele	debut_kele	1.5121	0.7641	0.7480	—	—	—	—
8	zhangtaotao	zhangtaotao	1.4936	0.7571	0.7366	—	—	—	—
9	MACS	alir	1.4717	0.7522	0.7194	0.7300	0.7800	0.9200	0.9400
10	orboai	orboai	1.4549	0.7397	0.7153	—	—	—	—
11	LION Team	sfezza	1.3774	0.7003	0.6771	—	—	—	—
12	SI analytics	Ahn	1.3280	0.6744	0.6535	0.5371	0.6264	0.9263	0.9211
13	tsubota	tsubota	1.2053	0.5955	0.6098	—	—	—	—
		LPIPS-VGG	1.2277	0.5947	0.6330	0.6695	0.7490	0.9433	0.9431
		LPIPS-Alex	1.1368	0.5658	0.5711	0.7444	0.7634	0.9211	0.9172
		PieAPP	1.2048	0.6074	0.5974	0.8478	0.8064	0.9182	0.9102
	Baselines	DISTS	1.3422	0.6548	0.6873	0.8184	0.8463	0.9468	0.9440
		SWD	1.2585	0.6243	0.6342	0.7895	0.8219	0.8832	0.8731
		FSIM	1.0748	0.5038	0.5709	0.8015	0.8560	0.9634	0.9491
		SSIM	0.7549	0.3614	0.3936	0.7414	0.7894	0.9479	0.9397
		PSNR	0.5263	0.2493	0.2769	0.6395	0.6541	0.8756	0.8686

ceptual distortions by explicitly considering the spatial misalignment using anti-aliasing pooling layers and spatially robust comparison operations in the IQA network.

3. The NTIRE Challenge on Perceptual IQA

We host the NTIRE 2021 Perceptual Image Quality Assessment Challenge and the objective are three-fold: (1) to push developing state-of-the-art perceptual image quality assessment algorithms to deal with the novel GAN-based distortion types; (2) to compare different solutions and gain new insights; and (3) to promote a novel large perceptual IQA dataset (PIPAL [20, 19]). Details about the challenge are as follows:

Task. The task of this challenge is to obtain an image quality assessment method capable to produce 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. Note that we do not restrict the participants to develop the full-reference IQA methods, and the blind IQA methods are also welcomed.

Dataset. 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 both traditional distortion types, image restoration results, compression results, and novel GAN-based image processing outputs. More than 1.13 million human judgements are collected to assign mean opinion scores (MOS) for PIPAL images using the Elo rating system [16]. The original PIPAL dataset includes 250 high-quality diverse reference images, 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.

We collect an extended version of the PIPAL dataset as the validation and testing set. We use the rest 50 reference images and collect 66 additional distorted images for each of them. 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. Thanks to the expandability of the Elo rating system used by the PIPAL dataset, we can assign MOS scores to new images with additional pairwise judgements without collecting from scratch. At last, 753k human judgements are involved in preparing the validation and testing set. The validation set contains 25 reference images and 40 distorted images for each of them. The testing set contains the rest 25 reference images and all the 66 distorted images
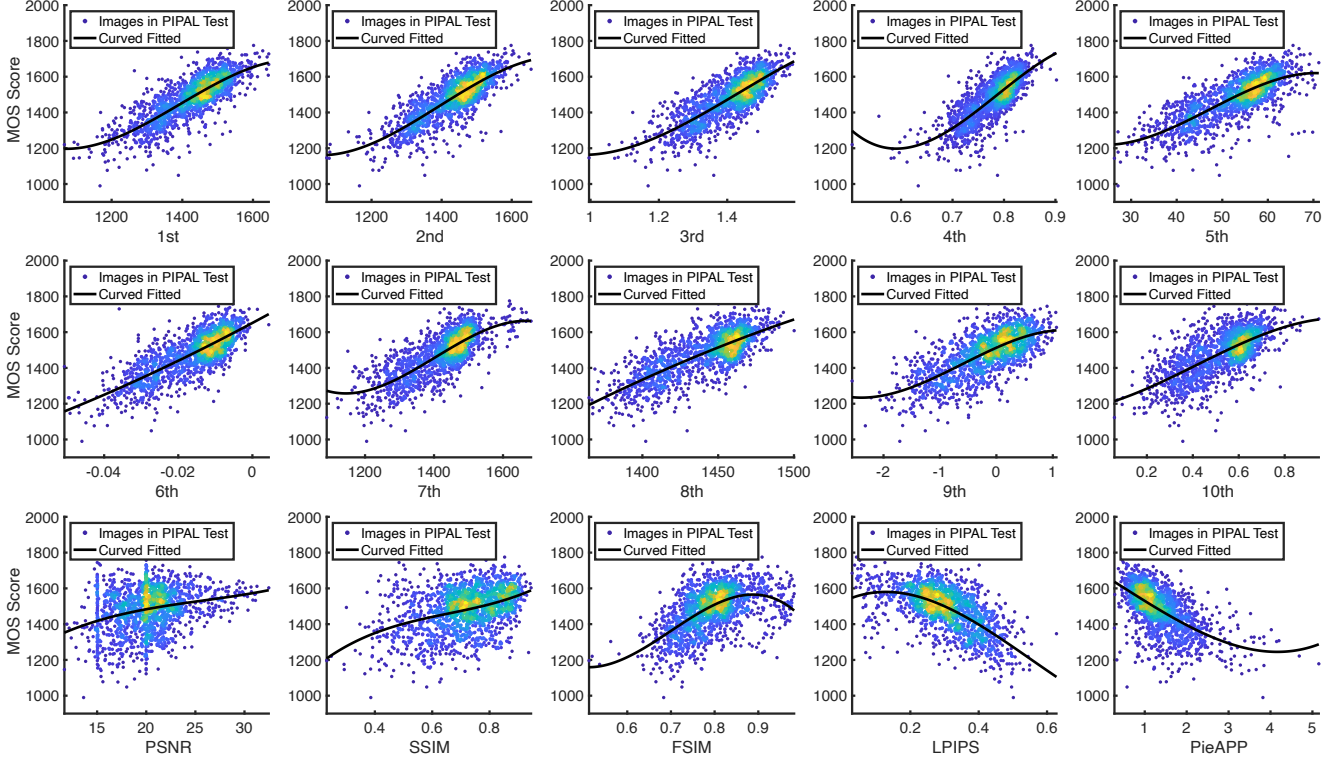
Figure 3. Scatter plots of the objective scores vs. the MOS scores.

for each reference image. Note that for the participants, the training set and the validation/testing set contain completely different reference and distorted images, which ensures the objectivity of the final results.

Evaluation protocol. Our evaluation indicator, namely main score, consists of both Spearman rank-order correlation coefficient (SRCC) [45] and Person linear correlation coefficient (PLCC) [4]:

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

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 [45, 19]. Before calculating PLCC index, we perform the third-order polynomial nonlinear regression as suggested in the previous works [38, 20]. By combining SRCC and PLCC, our indicator can measure the performance of participating models in an all-round way.

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 13 teams participated in the testing phase of the challenge. Table 1 reports the main results and important information of these teams. We also select some existing representative IQA methods as our baseline. Specifically, We choose PSNR, SSIM [54] and FSIM [59] as representative traditional IQA methods, PI [7] and NIQE [35] as representative blind IQA methods, and LPIPS [61], DISTS [12], PieAPP [39] and SWD [19] as representative deep-learning based methods. Table 1 reports the final test re-
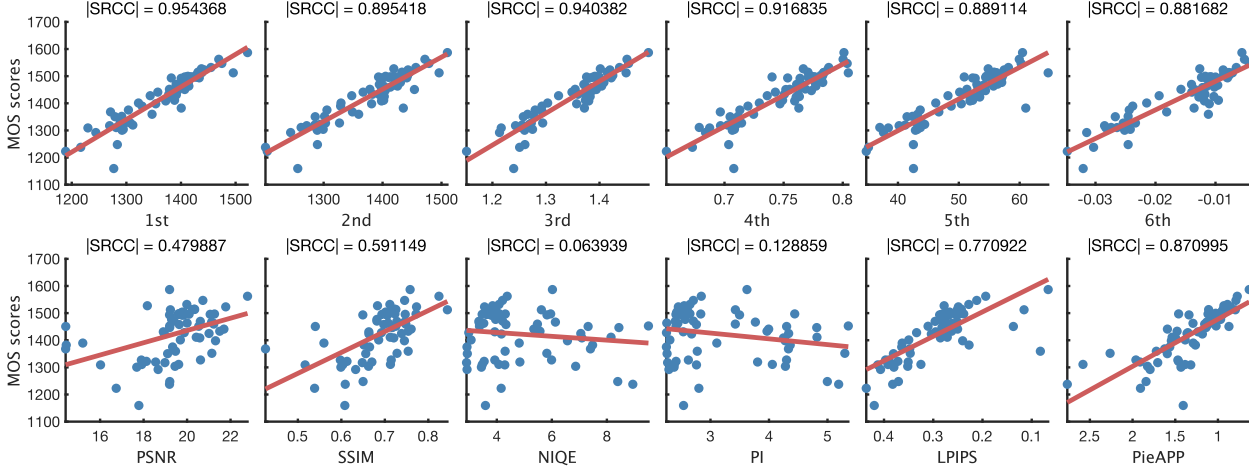
Figure 4. Analysis of IQA methods in evaluating IR methods. Each point represents an algorithm. Higher correlations indicates better performance in evaluating perceptual image algorithms.
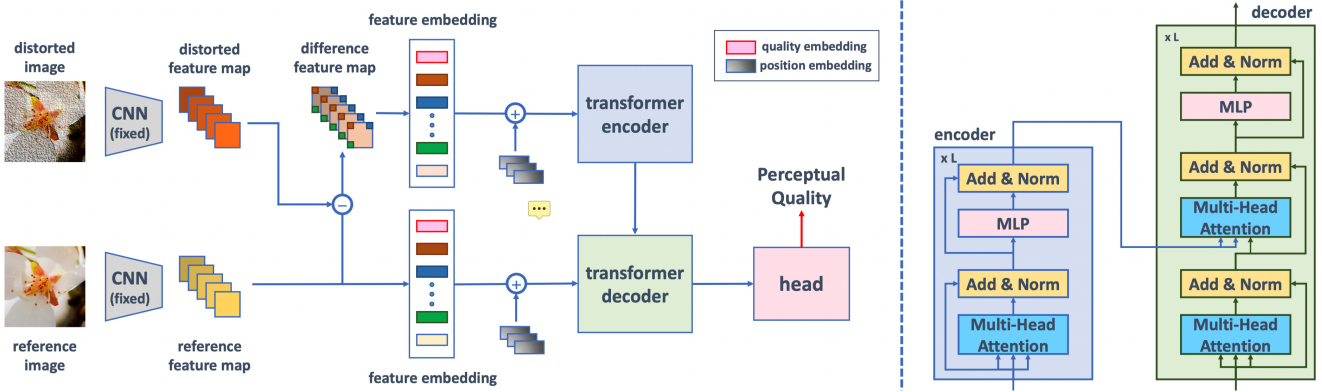

Figure 5. The overview of LIPT team's Image Quality Transformer (IQT) method.

sults and rankings of the challenge. The methods are briefly described in Section 5 and the team members are listed in Appendix A.

As shown in Table 1, 11 of 13 participating teams achieve an SRCC score higher than 0.75 on PIPAL, which significantly surpasses the highest performance of existing algorithms (0.65). The champion team achieves an SRCC score of 0.799 and a PLCC score of 0.790, refreshing the state-of-the-art performance on PIPAL. In order to evaluate their performance on traditional distortion types, we also report their results on TID2013 [38] and LIVE [46] dataset in Table 1. The top three teams in the challenge all achieve competitive results with existing methods on TID2013 and LIVE, showing their good generalization ability on traditional distortion types.

Figure 3 shows the scatter distributions of subjective MOS scores vs. the predicted scores by the top solutions and the other 5 IQA metrics on PIPAL test set. The curves shown in Figure 3 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. We then present the analysis of IQA methods as performance measures for perceptual image processing algorithms. Recall that an important goal of this challenge is to promote more promising IQA metrics for perceptual-oriented algorithms. In Figure 4, we show the scatter plots of subjective scores vs. the top solutions and some commonly-used IQA metrics for some perceptual-oriented algorithms. As can be seen, the top solutions perform well in evaluating the images in the testing set. Among them, the correlation between the evaluation of the champion solution (1st) and the subjective score reaches 0.95.

5. Challenge Methods
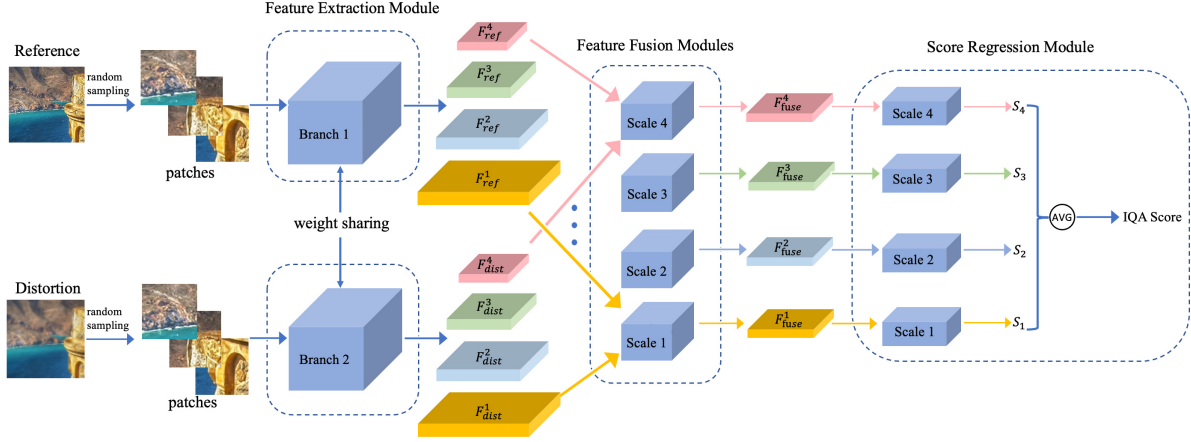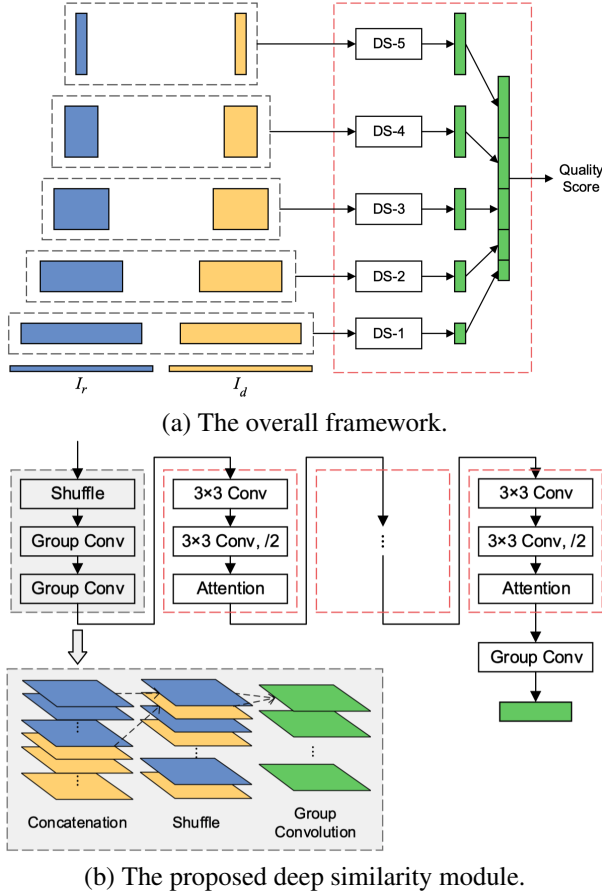
We describe the submitted solution details in this section.

Figure 6. The overview of MT-GTD team's bilateral-branch multi-scale image quality estimation (IQMA) network.

(a) The overall framework.

(b) The proposed deep similarity module.

Figure 7. The Amaurotia team: Learning to Learn a perceptual image path similarity metric.

5.1. LIPT

LIPT team is the winner of this challenge. They develop an image quality transformer (IQT), introduced in [10], that applies a transformer architecture to the perceptual IQA

task. Recently, the transformer-based models achieve impressive results in many vision tasks [14, 27]. However, this is the first time that the transformer technique [51] has been applied to the full-reference IQA task. The overview framework of their IQT method is illustrated in Figure 5. The IQT method consists of three parts, the feature extraction network, the transformer encoder and decoder, and the prediction head. Firstly, an Inception-ResNet-V2 network [50] pre-trained on ImageNet [42] is used to extract perceptual representations from both reference and distorted images. The extracted feature maps are then projected to vectors and a trainable extra quality embedding and position embedding are also added. Secondly, the transformer encoder takes the embedding of the feature map difference calculated between the reference feature map and the distorted feature map as input, and the output of the encoder is sent to the transformer decoder together with the embedded feature map of the reference image. The transformer encoder and decoder are based on the standard architecture of the transformer, which consists of multi-head attention modules, multi-layer perceptions and layer normalization. At last, the prediction head takes the output of the transformer decoder as the input and predicts the perceptual similarity score.

In the training phase of the IQT, M overlapping image patches of size 256×256 are cropped from both the reference image and the distorted image. The final quality score is obtained by averaging the quality scores of these patches. Horizontal flip and random rotation are applied as data augmentation during the training And the loss is calculated using a mean squared error between the predicted scores and the ground truth scores.

5.2. MT-GTD

MT-GTD team wins the second place in our challenge. They contribute a new bilateral-branch multi-scale image
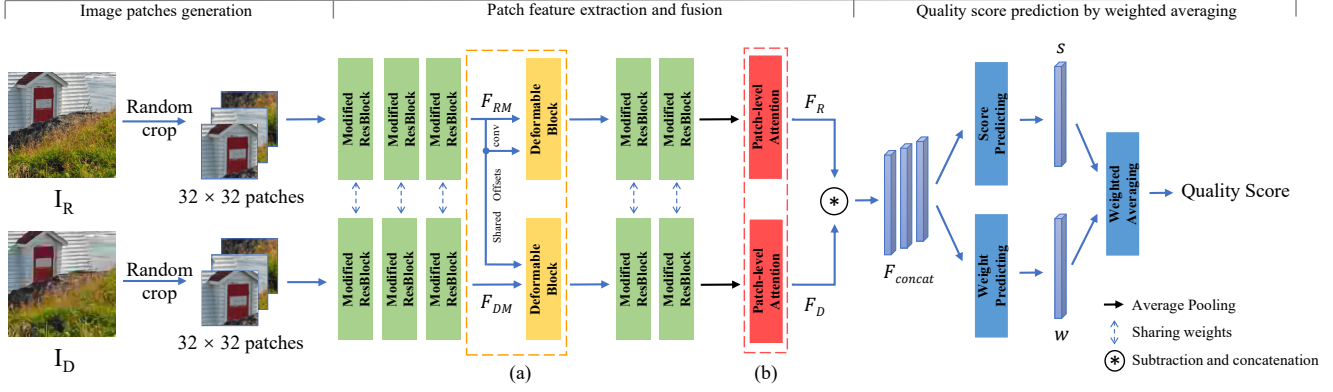
Figure 8. The overview of THUIIGROUP1919 team's Region Adaptive Deformable Network (RADN).

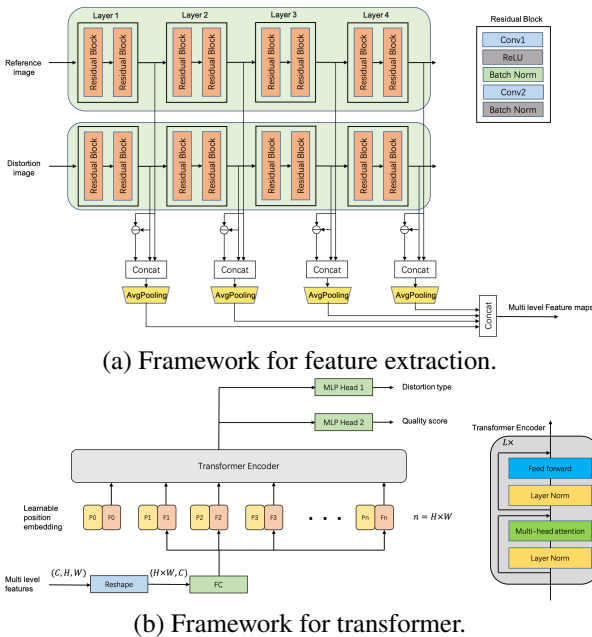(a) Framework for feature extraction.

(b) Framework for transformer.

Figure 9. The Yahaha team: A Transformer-based perceptual image quality assessment framework leveraging multi level features.

quality estimation (IQMA) network, which is detailed in [22]. At first, ResNet [24] pre-trained on ImageNet [29] is used as the feature extraction backbone. The IQMA network has two branches with Feature Pyramid Network (FPN)-like architecture to extract multi-scale features from patches of the reference image and corresponding patches of the distorted image separately. The feature maps of the same scale from both branches are then sent into several scale-specific feature fusion modules. Each module performs both a feature fusion operation and a pooling operation for corresponding features. Then several score regression modules are used to learn a quality score for each scale. Finally, image scores for different scales are fused as the quality score of the image. The overall framework is illustrated in Figure 6. It is worth noting that the parallel

integration method is used for ensemble models in the challenge. The MT-GTD team ensembles 9 models which are depicted in the testing description with an average score of all these models as the final score.

In the training phase of the IQMA network, a data augmentation operation is specifically designed to address the imbalance issue in the PIPAL training set. They observe that only a few images in the PIPAL training set have a subjective score of less than 1300. They re-sample these images using random horizontal flipping, vertical flipping and random rotations as data augmentation. The augmentation operation successfully increases the percentage of the images that have very small subjective scores. The training is driven using smooth L_1 loss as it is more robust to outliers.

5.3. The Amaurotia

The Amaurotia team extends the LPIPS metric [61] and proposed Learning to Learn a Perceptual Image Patch Similarity (L^2PIPS) method employed a new deep similarity (DS) module. Figure 7 (a) shows the overall framework of the proposed method. A ResNet-50 [24] network pre-trained on ImageNet [29] is used as the feature extraction backbone and multi-scale deep representations are extracted for comparison. Their main novelty lies in the DS module, which is designed to predict the similarities between feature pairs. In each DS module, a channel shuffle operation and group convolution are performed to compare the feature channels pair-wisely, as shown in Figure 7 (b). Moreover, frequency channel attention (FCA) [40] technique is also employed to perform attention mechanism. In the challenge, an additional cosine similarity is also used as an auxiliary output if each DS module.

The training phase of the proposed L^2PIPS has two stages. In the first 100 epochs, the ResNet-50 feature extraction network is frozen and the loss function is the L_1 loss. In the second phase, another 100 epochs training is performed. The weights of the feature extraction network are also trainable in this stage and the loss function is changed to mean
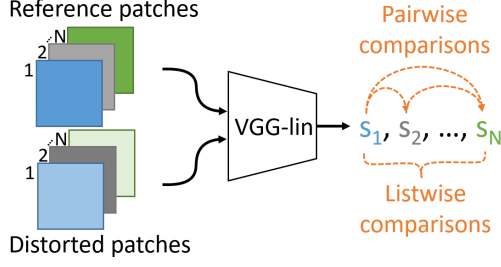
Figure 10. The Huawei Noah's Ark team: Image agnostic pairwise comparisons and differentiable correlation loss functions for full-reference image quality assessment.

square error loss.

5.4. THUIIGROUP1919

THUIIGROUP1919 proposes a new IQA network, namely Region Adaptive Deformable Network (RADN) [47], which is illustrated in Figure 8. Firstly, they modified the original residual block by removing batch normalization and adopting 3×3 convolution instead of 7×7. These modified residual blocks are trained from scratch to extract features for comparison. Note that no pre-trained model is used in the proposed RADN, which is different from most of the rest solutions. Considering that the human visual system's low sensitivity to the error and misalignment of the edges in perceptual distortions [19], reference-oriented deformable convolution modules are adopted to make better use of the reference information. The offset parameters are calculated using the reference image and the deformable convolutions use this offset to process the distorted image. Before being sent to quality score prediction, a novel patch-level attention module is designed to explore the correlations between local patches.

In the training phase of the RADN network, a contrastive pre-training strategy is proposed to make the model learn how to distinguish the image quality rather than directly guess the quality score. Similar to LPIPS [61] and SWD [19], a two-layer fully-connected network is employed to predict pairwise probabilities from the quality scores. The cross entropy loss function is used to calculate loss between the predicted probabilities and the ground truth probabilities obtained using Elo system in PIPAL dataset.

5.5. Yahaha

Yahaha team also adopts the transformer technique [51] to build their IQA method. A ResNet-18 [24] backbone network is first employed for feature extraction. As shown in Figure 9 (a), both reference image and distortion image are fed to the backbone network to obtain the feature maps of each layer. The feature maps and the difference of quality maps are concatenated together as $c_i = r_i \oplus d_i \oplus (r_i - d_i)$, where c_i is the concatenated feature from layer i, r_i and d_i

(a) Siamese-Difference architecture.

(b) Siamese-Difference with attention architecture.

(c) Siamese-Difference with ConvLSTM layer.
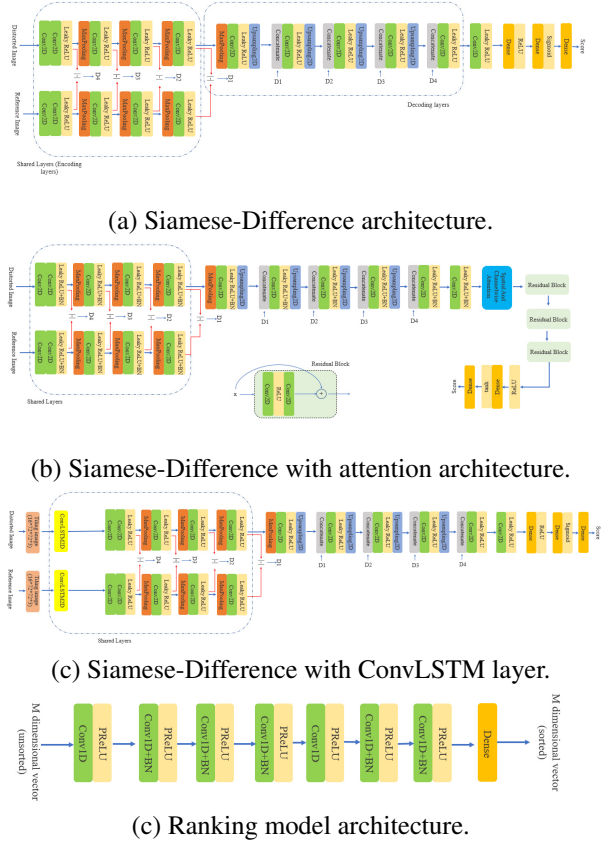

(c) Ranking model architecture.

Figure 11. The MACS team: Siamese-Difference Network for IQA.

are the reference and distorted image feature maps of layer i. Then, feature maps from different layers are downsampled to the same spatial resolution by average pooling before being sent to the transformer. Structure of transformer stage is shown in Figure 9 (b), features of each position are added with a learnable position embedding before fed to a standard transformer encoder. The transformer encoder is connected with two multi-layer fully-connected network heads for the prediction of distortion type and opinion score, respectively. The loss from these two parts are weighted summed during training.

In the training phase, the Yahaha team employs both L_1 loss and a relative-distance loss In relative-distance loss, the difference between prediction scores and the ground truth subjective scores are compared for each image pair. It can be formulated by:

$$Loss = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |(y_i - y_j) - (x_i - x_j)|,$$

where y_i and x_i represent the ith ground truth score and prediction score, respectively.
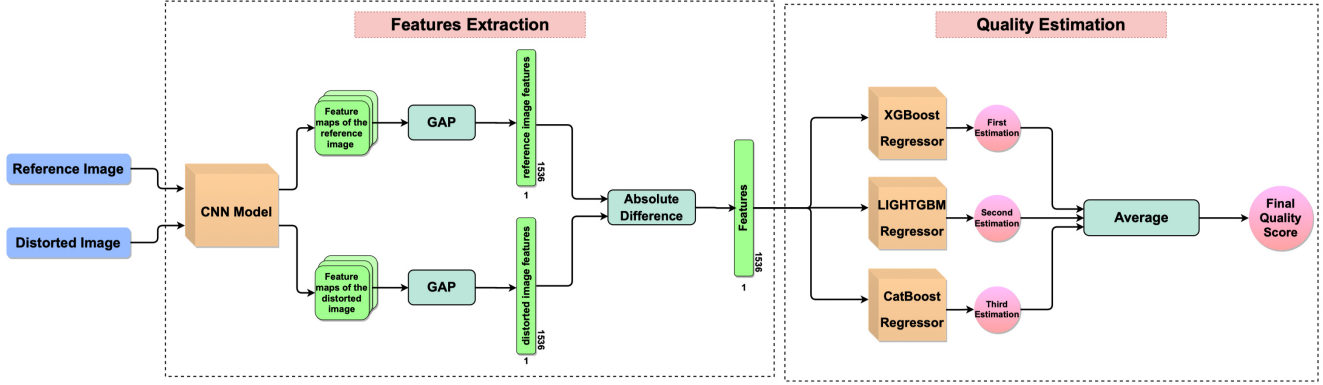
Figure 12. The LION team: Image Quality Estimation based on Ensemble of Gradient Boosting.
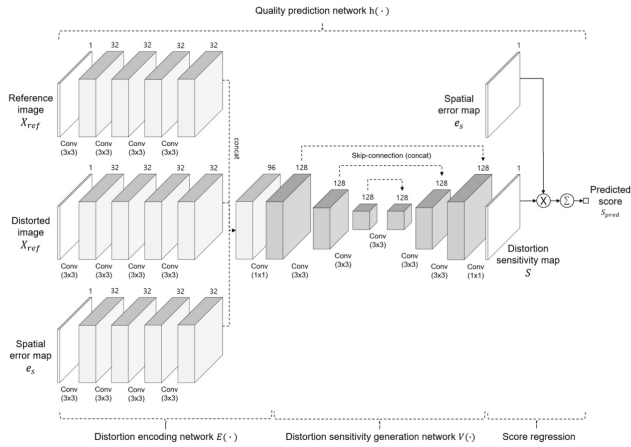

Figure 13. SI Analytics team: Deep Learning-based Distortion Sensitivity Prediction for Full-Reference Image Quality Assessment.

5.6. Huawei Noah's Ark

The innovative point of the method proposed by Huawei Noah's Ark team is the use of two special loss functions. The model is trained with a loss function that incorporates (a) pairwise comparisons, and (b) listwise comparisons. Having many and diverse comparisons with distorted images coming from different reference images provides a better signal during training, and better mimics how IQA is evaluated.

For the pairwise comparison loss, the Bradley-Terry (BT) sigmoid [8] is used to obtain a probability of which image in the pair has the best quality. The pair of distorted images can originate from different reference images, making the pairwise comparison agnostic to their reference image. Hence, the model is forced to compare pairs of images with different scene content, as is the case in IQA evaluation. For training, the MSE loss between the predicted probability from the BT model and its ground truth probability computed with the Elo logistic curve is used.

In addition to the pairwise comparison, novel differentiable regularizers derived from the PLCC and SRCC are also used as losses. The regularizers act as listwise comparisons where all samples in a mini-batch form an ordered list. Having listwise comparisons complement the pairwise comparisons, as every sample is now also compared to the rest of the mini-batch. Hence, the model is forced to learn subtle differences (*i.e.*, pairwise comparisons) as well as relative differences (*i.e.*, listwise comparisons). These two losses have the same weights in the training phase.

5.7. debut_kele

Debut_kele team employs ensemble technology to build their method. They mainly employs the complementary information provided by traditional IQA and deep learning-based IQA algorithms. The contributed solution can be divided into two main parts: feature extraction using different perceptual image quality assessment, which include SSIM, MS-SSIM, CW-SSIM, GMSD, LPIPS-VGG, DISTS, NLPD, FSIM, VSI, VIFs, VIF, and MAD. Then, a regression model is built using XGBoost [9] based on the pre-calculated IQA. The 'max depth' parameter of the model is set to 3 and learning rate was set at 0.01. In addition, feature subsample and the sample subsample values were set at 0.7 to prevent from overfitting. The maximum iteration round is set as 10000 while the early-stopping round is set 500. In the challenge, 5 fold bagging-based ensemble strategy is used.

5.8. MACS

MACS team proposes three different networks that all based on a Siamese-Difference architecture, which is detailed in [43]. The basic Siamese-Difference network are illustrated in Figure 11 (a). In the decoding part, the absolute values of the outputs are concatenated from the encoding part. Instead of using concatenation of the encoded images directly for the skip connections, the absolute difference of the encoded feature maps have been used. The

9

decoding part is not symmetric, which is different from the traditional Siamese networks The authors argue that this architecture has more capability in representing useful features to estimate the quality score rather than the traditional Siamese network architecture. In the second network, channel-wise attention and spatial attention are used to improve the Siamese-Difference architecture, as shown in Figure 11 (b). To mitigate the issue of high computational cost of the proposed network on big images, the third network first tiled each input into patches with size 72×72 and stack them as an image bag. These patches are sent to a ConvLSTM Layer, as shown in Figure 11 (c). In the challenge, each network is trained separately, and the averaged score of them are used as the final score.

In the training phase, multiple loss functions are combined to achieve better results. Firstly, MSE loss is used as the most common loss function for a regression task. A new differenitable pearson loss is also employed to increase the PLCC performance. At last, a novel rank loss is proposed in order to improve the SRCC rank correlation performance. Inspired by SoDeep [17], a ranker network is trained to learn how to sort M inputs. The network architecture is shown in Figure 11 (d). To train this network, M dimensional random vectors are generated and the network is trained to learn how to sort the inputs by minimizing mean absolute error between the sorted input and the output of the network. After training this network, this network can be used as a differentiable sorting function. Suppose $R(S)$ and $R(S')$ are the estimated ranked vectors of S and S', respectively. The squared difference between these two vectors is used in order to increase SRCC measure as $L_{Rank} = \frac{1}{M}\|R(S) - R(S')\|^2$.

5.9. LION

LION team also employs boosting strategy to build their solution, which is introduced in [23]. Firstly, the VGG-16 network [48] pre-trained on the ImageNet [29] is used as the feature extraction module. First, an analysis is performed on features extracted from different layers to determine if a particular layer of the model provides more relevant feature maps for image quality assessment task. For each layer of the VGG-16 model, a distance between reference and distorted images is computed using extracted feature maps. The correlation between the calculated distances and the subjective scores is computed using SRCC. Three intermediate layers (block4 conv2, block4 conv3 and block5 conv1) are highlighted by analyzing these SRCC values and are considered to be the best candidates for the evaluation of the perceptually image quality. When performing image quality estimation, the reference and distorted images are fed to the VGG-16 network to derive the feature vector from the three selected intermediate convolution layers. Then, global average pooling is performed to reduce the feature maps into

a manageable size. Finally, to provide the feature vector to the regression part of the framework, the absolute difference between the extracted features of the reference and distorted images is calculated. The feature vector is then fed into the three Gradient Boosting regression models, which are XGBoost [9], LightGBM [26] and CatBoost [13], to be regressed to predict three image quality scores. The three models are ensemble models using decision trees. The idea behind using three Gradient Boosting regression models, is to combine the predictions of these models to form an ensemble that outperforms a single trained model.

5.10. SI Analytics

SI Analytics team proposes a method motivated by the visual sensitivity map, which is detailed in [2]. The visual sensitivity map refers to a weighting map describing the degree of visual importance of each pixel to the human visual system (HVS). As shown in Figure 13, their method firstly predicts visual sensitivity map allocating local weights to the pixels according to their local spatial characteristics based on the reference image, the distorted image, and the spatial error map. The spatial error map is defined by $\frac{log(1/((X_{ref}-X_{dis})^2+\varepsilon/255^2))}{log(255^2/\varepsilon)}$, where X_{ref} and X_{dis} are gray-scale reference image and distorted image normalized to $[0,1]$, and $\varepsilon = 1$. With the spatial error map and predicted visual sensitivity map, the Hadamard product and global average pooling are applied in series to obtain the final subjective score. Different from DeepQA [28], they use UNet [41] structure to conserve the spatial information of input images instead of using down-sampling operations for predicting distortion sensitivity maps. Moreover, there are more convolution layers in the encoding network to enlarge the receptive fields.

In the training phase, they normalize the ground-truth score into $[0,1]$ and oversample those images outside the range of $[0.4, 0.8]$ to relieve score imbalance problems. The loss function adopts the combination of mean-square error and L2 loss between predicted quality score and subjective scores.

5.11. tsubota

Tsubota team builds their method by modifying PieAPP [39] on feature extraction network and pooling layers. Specifically, they replace the VGG [48] network in PieAPP with AlexNet [29] pre-trained on ImageNet [11] image and adopt Blurpool [60] to avoid aliasing, not commonly used Maxpool.

In the training phase, they firstly train their model on the PieAPP dataset to obtain a pre-trained model. Then, they finetune the pre-trained model on the PIPAL dataset. For those trainings, the input image was randomly cropped into 36 patches with a resolution of 64×64 pixels. In the testing phase, they use a 64×64 sliding window with the

stride size set to 6 pixels. The rest of settings are same with PieAPP.

Acknowledgements

A. Teams and Affiliations

NTIRE 2021 Team

Title:
NTIRE 2021 Challenge on Perceptual Image Quality Assessment
Members:
Jinjin Gu[1] *(jinjin.gu@sydney.edu.au)*, Haoming Cai[2], Chao Dong[2], Jimmy S. Ren[3], Yu Qiao[2], Shuhang Gu[4], Radu Timofte[5]
Affiliations:
[1] School of Electrical and Information Engineering, The University of Sydney
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] SenseTime Research
[4] OPPO Research
[5] Computer Vision Lab, ETH Zurich, Switzerland

LIPT

Title:
Perceptual Image Quality Assessment with Transformers
Members:
Manri Cheon[1] *(manri.cheon@lge.com)*, Sungjun Yoon[1], Byungyeon Kangg Kang[1], Junwoo Lee[1]
Affiliations:
[1] LG Electronics

MT-GTD

Title:
IQMA Network: Image Quality Multi-Scale Assessment Network
Members:
Qing Zhang[1] *(zhangqing31@meituan.com)*, Haiyang Guo[1], Yi Bin[1], Yuqing Hou[1], Hengliang Luo[1]
Affiliations:
[1] Meituan Group

The Amaurotia

Title:
L^{2}PIPS: Learn to Learn a Perceptual Image Patch Similarity Metric

Members:
Jingyu Guo[1] *(gjy19@mails.tsinghua.edu.cn)*, Zirui Wang[1], Hai Wang[1], Wenming Yang[2]
Affiliations:
[1] Department of Electronic Engineering, Tsinghua University
[2] Shenzhen International Graduate School, Department of Electronic Engineering, Tsinghua University

THUIIGROUP1919

Title:
WResNet: A Light Residual Network for FR-IQA
Members:
Qingyan Bai[1] *(baiqingyan1998@163.com)*, *Shuwei Shi*[1] *(ssw20@mails.tsinghua.edu.cn)*, Weihao Xia[1], Mingdeng Cao[2], Jiahao Wang[2], Yifan Chen[1], Yujiu Yang[1]
Affiliations:
[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Department of Automation, Tsinghua University

Yahaha

Title:
A Transformer-based perceptual image quality assessment framework leveraging multi level features
Members:
Yang Li[1] *(liyang.00@pku.edu.cn)*, Tao Zhang[2], Longtao Feng[2], Yiting Liao[2], Junlin Li[2]
Affiliations:
[1] Peking University
[2] Bytedance Inc.

Huawei Noah's Ark

Members:
William Thong[1] *(william.thong@huawei.com)*, Jose Costa Pereira[1], Ales Leonardis[1], Steven McDonagh[1]
Affiliations:
[1] Huawei Noah's Ark Lab

debut_kele

Members:
Kele Xu[1] *(kelele.xu@gmail.com)*, Lehan Yang[2], Hengxing Cai[3], Pengfei Sun[45]
Affiliations:
[1] Key Laboratory for Parallel and Distributed Processing
[2] The University of Sydney
[3] 4Paradigm Inc.
[4] University of Zurich
[5] ETH Zurich

MACS

Title:

IQA Using SIDIS: Siamese Differential Network with Surrogate Loss function

Members:

Seyed Mehdi Ayyoubzadeh[1] (zhangqing31@meituan.com), Ali Royat[2]

Affiliations:

[1] ECE Department, McMaster University

[2] EE Department, Sharif University of Technology

LION

Title:

Image Quality Estimation based on Ensemble of Gradient Boosting

Members:

Sid Ahmed Fezza[1] (sfezza@inttic.dz), Dounia Hammou[1], Wassim Hamidouche[2]

Affiliations:

[1] National Institute of Telecommunications and ICT, Oran, Algeria

[2] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France

SI Analytics

Title:

Deep Learning-based Distortion Sensitivity Generation for Full-Reference Image Quality

Members:

Sewoong Ahn[1] (sfezza@inttic.dz), Gwangjin Yoon[1]

Affiliations:

[1] SI Analytics

tsubota

Title:

PieAPP Using ImageNet Pre-trained AlexNet

Members:

Koki Tsubota[1] (tsubota@hal.t.u-tokyo.ac.jp), Hiroaki Akutsu[2], Kiyoharu Aizawa[1]

Affiliations:

[1] The University of Tokyo

[2] Hitachi, Ltd.

References

[1] Abdullah Abuolaim, Radu Timofte, Michael S Brown, et al. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[2] Sewoong Ahn, Yeji Choi, and Kwangjin Yoon. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 10

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

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

[5] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[6] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[7] 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. 1, 2, 4

[8] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 9

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 9, 10

[10] 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. 6

[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. Ieee, 2009. 10

[12] 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. 4

[13] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018. 10

[14] 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

[15] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021 depth guided image relighting challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

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

[17] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10792–10801, 2019. 10

[18] 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

[19] 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. 2, 3, 4, 8

[20] 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, 3, 4

[21] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[22] 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. 7

[23] Dounia HAMMOU, Sid Ahmed FEZZA, and Wassim Hamidouche. Egb: Image quality assessment based on ensemble of gradient boosting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 10

[24] 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. 7, 8

[25] 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

[26] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017. 10

[27] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 6

[28] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684, 2017. 10

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural net-

works. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 7, 10

[30] 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

[31] Jerrick Liu, Oliver Nina, Radu Timofte, et al. NTIRE 2021 multi-modal aerial view object classification challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[32] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 learning the super-resolution space challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[33] 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

[34] 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. 2

[35] 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, 4

[36] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[37] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[38] 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, 5

[39] 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, 4, 10

[40] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. *arXiv preprint arXiv:2012.11879*, 2020. 7

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 10

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[43] Ali Royat Seyed Mehdi Ayyoubzadeh. (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. 9

[44] 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

[45] 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. 4

[46] 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. 5

[47] 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. 8

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 10

[49] Sanghyun Son, Suyoung Lee, Seungjun Nah, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[50] 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. 6

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017. 6, 8

[52] 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

[53] 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

[54] 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, 4

[55] 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

[56] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[57] 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

[58] 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

[59] 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, 4

[60] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pages 7324–7334. PMLR, 2019. 10

[61] 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, 4, 7, 8

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