Deep Portrait Quality Assessment. A NTIRE 2024 Challenge Survey

Nicolas Chahine * Marcos V. Conde *† Daniela Carfora * Gabriel Pacianotto *
Benoit Pochon * Sira Ferradans * Radu Timofte * Zhichao Duan Xinrui Xu
Yipo Huang Quan Yuan Xiangfei Sheng Zhichao Yang Leida Li Haotian Fan
Fangyuan Kong Yifang Xu Wei Sun Weixia Zhang Yanwei Jiang Haoning Wu
Zicheng Zhang Jun Jia Yingjie Zhou Zhongpeng Ji Xiongkuo Min Weisi Lin
Guangtao Zhai Xiaoqi Wang Junqi Liu Zixi Guo Yun Zhang Zewen Chen
Wen Wang Juan Wang Bing Li

Figure 1. Sample portraits from the *NTIRE 2024 Portrait Quality Assessment Challenge* testing set.

Abstract

This paper reviews the NTIRE 2024 Portrait Quality Assessment Challenge, highlighting the proposed solutions and results. This challenge aims to obtain an efficient deep neural network capable of estimating the perceptual quality of real portrait photos. The methods must generalize to diverse scenes and diverse lighting conditions (indoor, outdoor, low-light), movement, blur, and other challenging conditions. In the challenge, 140 participants registered, and 35 submitted results during the challenge period. The performance of the top 5 submissions is reviewed and provided here as a gauge for the current state-of-the-art in Portrait Quality Assessment.

1. Introduction

Portrait Quality Assessment (PQA) is becoming increasingly important in a variety of fields, from social media engagement to professional photography. The subjective nature of aesthetic appreciation, combined with the technical complexities of image capture and processing, makes PQA a challenging task. While Redi et al. [25] have explored the attributes that contribute to the perceived beauty of portraits, the utility-focused approach of Face Image Quality Assessment (FIQA) [27] underscores the diversity of criteria required for different quality assessment contexts.

The widespread use of smartphones has democratized portrait photography, yet achieving professional-quality images remains a challenge due to hardware limitations and the intricacies of advanced image processing techniques. Traditional objective quality assessment methods often fall short, as they typically do not account for the non-linear processing involved in modern photography, such as multi-image fusion and AI enhancements [32]. This gap has led to the rise of Blind Image Quality Assessment (BIQA) approaches, which evaluate image quality without the need for reference images. However, these methods frequently overlook the scene-specific semantics that significantly influence perceived quality, leading to a "one-size-fits-all" approach that is rarely effective across varied conditions. The challenges of domain shift and generalization — where the

∗ Marcos V. Conde, Nicolas Chahine, Daniela Carfora, Gabriel Pacianotto, Sira Ferradans, Benoit Pochon, Radu Timofte are the challenge organizers, while the other authors participated in the challenge.

Marcos V. Conde († corresponding author) and Radu Timofte are with University of Würzburg, CAIDAS & IFI, Computer Vision Lab.

Nicolas Chahine, Daniela Carfora, Gabriel Pacianotto, Sira Ferradans, Benoit Pochon are with DXOMARK.

NTIRE 2024 webpage: https://cvlai.net/ntire/2024
Code and Dataset: https://github.com/DXOMARK-Research/PIQ2023

quality assessment model fails to adapt to different conditions — remain as significant obstacles [41].

This paper introduces several novel frameworks aimed at addressing the shortcomings of current PQA methods, particularly in handling domain shifts and ensuring generalizability to unseen portrait conditions. Through an organized challenge, we seek to explore and validate these frameworks, setting new standards for PQA that can adapt to the diverse and dynamic nature of portrait photography. Our challenge relies on the PIQ23 public dataset [3] and a private portrait dataset designed specifically to explore the aforementioned challenges.

Related Computer Vision Challenges Our challenge is one of the NTIRE 2024 Workshop associated challenges on: dense and non-homogeneous dehazing [1], night photography rendering [2], blind compressed image enhancement [38], shadow removal [33], efficient super resolution [26], image super resolution (×4) [7], light field image super-resolution [36], stereo image super-resolution [35], HR depth from images of specular and transparent surfaces [40], bracketing image restoration and enhancement [43], portrait quality assessment [4], quality assessment for AI-generated content [21], restore any image model (RAIM) in the wild [19], RAW image super-resolution [9], short-form UGC video quality assessment [18], low light enhancement [22].

2. Portrait IQA Challenge

In this challenge, we introduce the PIQ benchmark [4], based on the PIQ23 portrait dataset [3] published in 2023, composed of diverse skin tone photographs in challenging scenarios for smartphone cameras. The dataset is divided into 50 "scenes" defined by their illumination condition, target distance, framing, posture, background, etc. Every scene has around 100 images collected from multiple smartphones and covering various subjects. Each scene is separately annotated according to three image quality attributes (detail/noise, exposure/contrast, and overall) using pairwise comparisons, which yields precise and consistent quality insights when applied to image groups with similar content. Around 600k comparisons in total (for the 3 features) were collected from 30 experts in controlled visualization conditions (calibrated screens, fixed eye-to-screen distance, controlled background illumination, etc. [3]). These annotations were converted to JODs (Just Objectionable Difference), quality units where 1 unit apart means that 75% of the observers can see the quality difference between two images, using psychometric scaling algorithms. By design, each scene has an independent quality scale where the scores of the scenes are not inter-comparable. This introduces a challenge when training machine learning models.

Test Dataset and Evaluation In this challenge, we proposed to focus on the *overall* attribute and a *"generalization split"* (we will refer to this as the challenge testing set, hidden/private test), that is, to evaluate the capacity of the models to generalize outside the training scenes when evaluating the overall quality of the portrait. We cannot expect the ML model to correctly estimate the JOD quality value of the image since it is dependent on the scene, but it should be able to correctly rank a set of images according to their quality.

We split the **evaluation procedure** into two phases. For the preliminary testing phase, we propose to use the public PIQ23 (Fig. 2) test set with no scene overlap with the training set (the images of scene 1 in the training set cannot be in the testing set). The final testing phase is based on a private test set composed of 96 single-person scenes of 7 images each, taken with 6 high-quality smartphone images and 1 DSLR capture edited by a photographer used as the quality reference (Figs. 1 and 3).

The participants do not have access to the challenge generalization testset. The results are obtained by executing their submitted models to ensure reproducibility, and basic runtime requirements on commercial GPUs.

2.1. Baseline Models

We have chosen to compare the proposed models with multiple baseline methods from the HyperIQA family (HyperIQA [29], SEM-HyperIQA [3] and FHIQA [5]) which are specifically designed to tackle the domain shift and scene semantics understanding, and that have proven performance on the PIQ23 dataset.

HyperIQA uses the HyperNetwork architecture to incorporate semantic information into image quality predictions. Building upon HyperIQA, SEM-HyperIQA introduces a multitasking approach that allows for scene-specific rescaling of quality scores. It employs a multi-layer perceptron (MLP) to predict the scene category of an image, which is then used to adjust the quality scores of individual patches through a scene-specific multiplier and offset. However, it assumes that each image belongs to a known scene category, which limits its ability to generalize to new scenes. FHIQA (Fig. 4) extends the concept of quality score rescaling by utilizing the entire scene prediction vector, rather than relying on a single scene category. This vector represents the similarity of the input image to all known scene categories, allowing FHIQA to rescale the pre-quality score based on a weighted combination of these similarities. The key innovation of FHIQA lies in its potential to generalize to new scene categories not included in the training set, by leveraging the information encoded in the classification weights.

https://cvlai.net/ntire/2024/

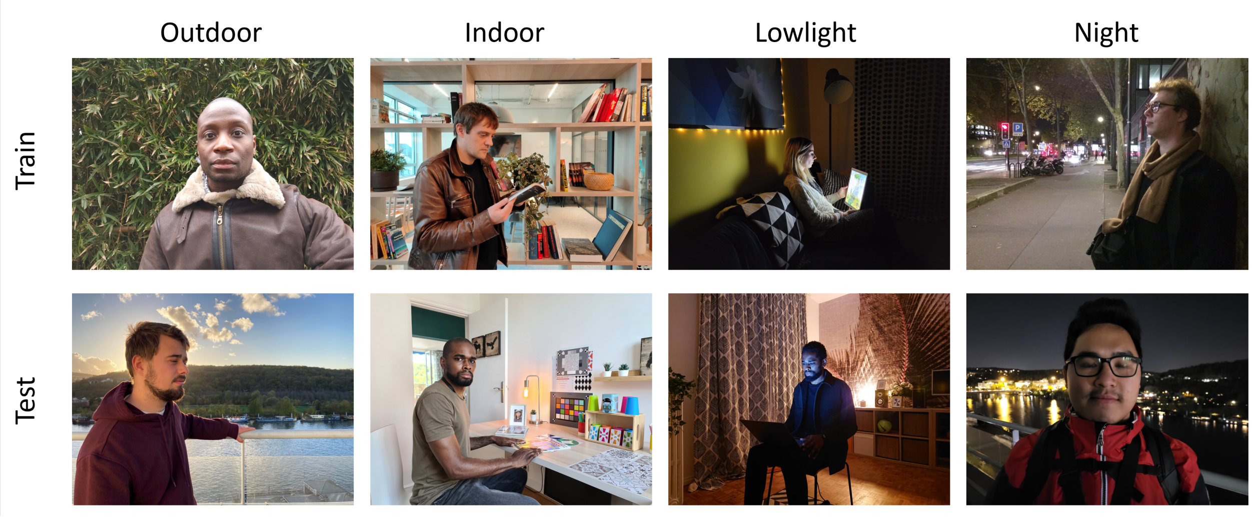

Figure 2. Examples from the **train/test split of PIQ23** [3]. The test set incorporates various framing settings, backgrounds, subject characteristics, and weather conditions that are significantly distinct from the training set.

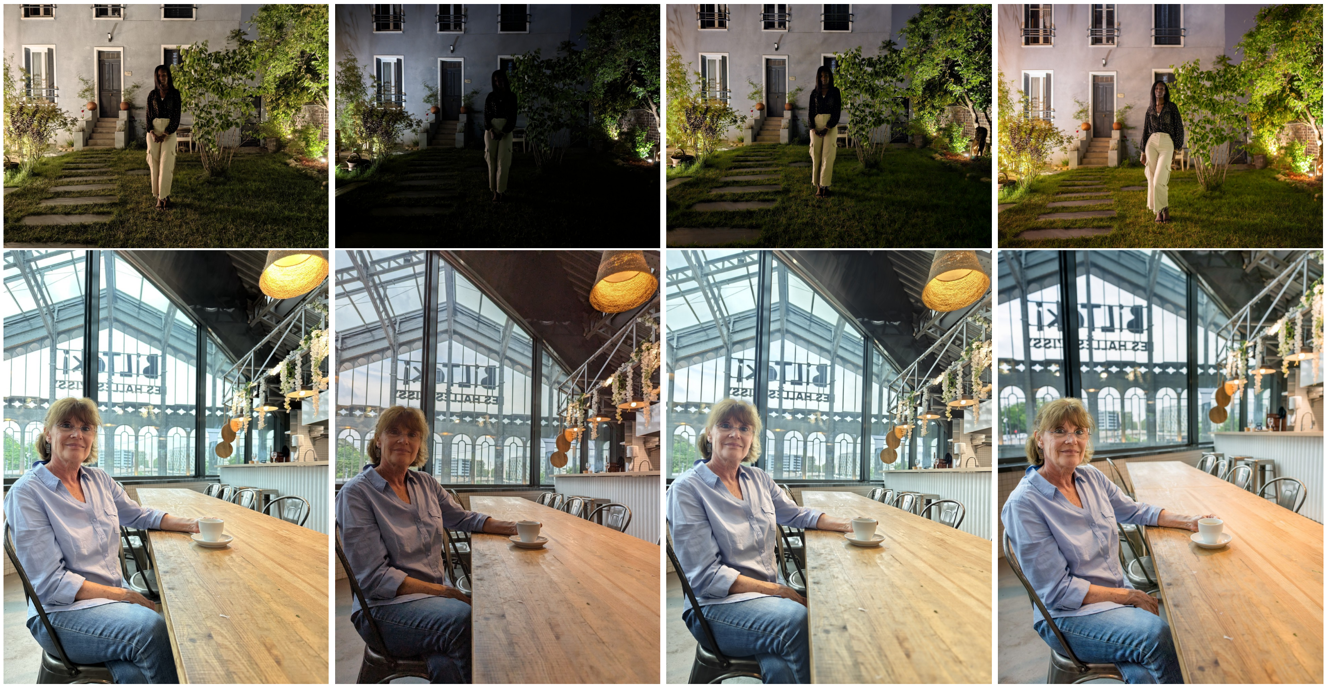

Figure 3. Sample images from two scenes of the **challenge generalization test set**. The three first image columns were taken with different smartphone devices, while the last column of images was taken with a DSLR camera and edited by a professional photographer.

3. Challenge Results and Methods

In the following sections we describe the best challenge solutions. Note that the method descriptions were provided by each team as their contribution to this survey.

In Tab. 1 we provide the benchmark using standard correlation metrics. We can observe that all the methods struggle to generalize in the challenge testset. The reason is that the new test images were captured using high-quality smartphones, extending the PIQ23 [3] dataset. The models struggle with this quality domain gap, which indicates that the model performance highly depends on the device used for capturing the data.

In Tab. 2 we provide the final ranking and additional information of the methods.

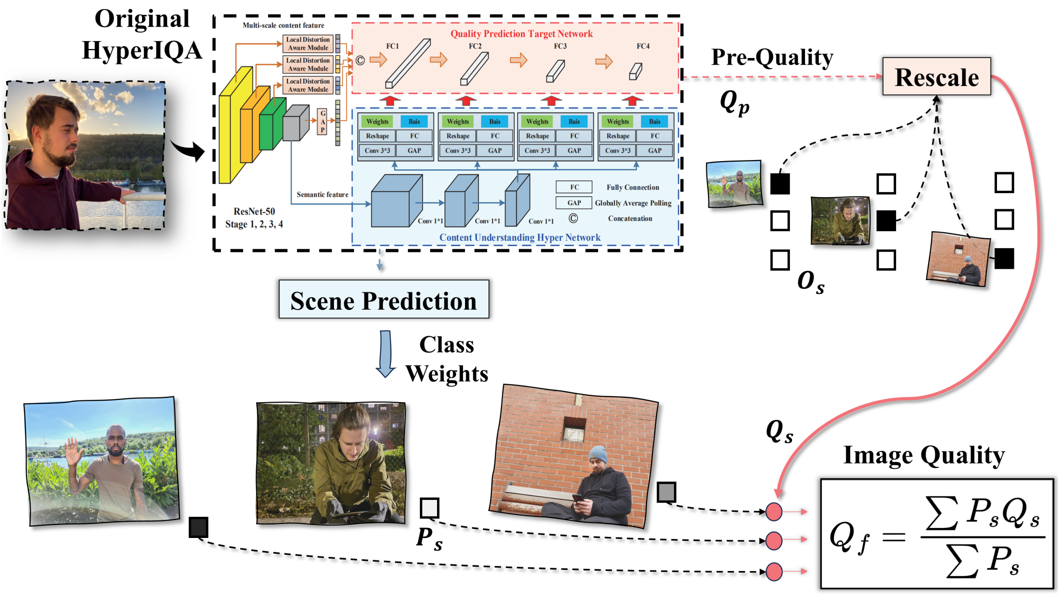
Original HyperIQA
Multi-scale content feature
Local Distortion Aware Module
Local Distortion Aware Module
Local Distortion Aware Module
GAP
ResNet-50
Stage 1, 2, 3, 4
Semantic feature
Quality Prediction Target Network
FC1
FC2
FC3
FC4
Weights
Bais
Reshape
FC
Conv 3*3
GAP
Conv 1*1
Conv 1*1
Conv 1*1
FC Fully Connection
GAP Globally Average Polling
© Concatenation
Content Understanding Hyper Network
Pre-Quality
Q_p
Rescale
O_s
Scene Prediction
Class Weights
P_s
Q_s
Image Quality
$Q_f = \frac{\sum P_s Q_s}{\sum P_s}$

Figure 4. Diagram of FULL-HyperIQA (FHIQA). The figure illustrates how FHIQA processes input images, extracts semantic information, and adapts the quality prediction based on scene-specific evaluations.

Team	Method	PIQ23 Test [3]			Challenge Test		
		SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
Xidian-IPPL	RQ-Net (Sec. 3.1)	0.820	0.839	0.621	**0.554**	0.597	0.381
BDVQAGroup	BDVQA (Sec. 3.2)	0.849	**0.866**	0.667	0.393	0.575	0.333
SJTU MMLab	PQE (Sec. 3.3)	**0.864**	0.857	**0.690**	0.411	0.544	0.333
I²Group	MoNet (Sec. 3.4)	0.760	0.791	0.580	0.357	0.433	0.286
SECE-SYSU	SAR (Sec. 3.5)	0.828	0.855	0.651	0.304	0.453	0.238
Baseline 1	HyperIQA [29]	0.740	0.736	0.550	0.429	0.560	0.333
Baseline 2	SEM-HyperIQA [3]	0.749	0.752	0.558	0.518	0.605	0.333
Baseline 3	FULL-HyperIQA [5]	0.778	0.784	0.586	0.536	**0.633**	**0.429**

Table 1. **Challenge Benchmark.** SRCC: Spearman Rank Correlation Coefficient, PLCC: Pearson Linear Correlation Coefficient, KRCC: Kendal Rank Correlation Coefficient. The correlation metrics are calculated per scene, and the final result corresponds to the median of scene-wise metrics. We highlight the **best** and second best.

Team	Method	PIQ23 Test [3]	Challenge Test	# Params. (M)	Extra Data	Train Res.
Xidian-IPPL	RQ-Net (Sec. 3.1)	0.751	**0.517**	**57.6**	Yes	224
BDVQAGroup	BDVQA (Sec. 3.2)	0.779	0.433	794	No	384
SJTU MMLab	PQE (Sec. 3.3)	**0.811**	0.429	174	Yes	384
I²Group	MoNet (Sec. 3.4)	0.710	0.368	408	No	384
SECE-SYSU	SAR (Sec. 3.5)	0.777	0.315	149	No	224
Baseline 1	HyperIQA [29]	0.676	0.456	128	No	1300px
Baseline 2	SEM-HyperIQA [3]	0.690	0.501	145	No	1300px
Baseline 3	FULL-HyperIQA [5]	0.711	0.515	145	No	1300px

Table 2. The final metric for each testing set consists of the median of the scene-wise average of the SRCC, PLCC, and KRCC correlations. We also provide the training resolution in pixels (px), number of parameters (in Millions), and if the team used additional data for training.

3.1. RQ-Net: Towards Robust Cross-scene Relative Quality Assessment

Team Xidian IPPL

Zhichao Duan, Xinrui Xu, Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Leida Li

Xidian University

Contact: `zach@stu.xidian.edu.cn`

The team presents a method for Robust Cross-scene **R**elative **Q**uality Assessment.

RQ-Net is a method to predict the relative quality of images. It consists of two branches: global quality perception and local quality perception. As shown in Fig. 5. A downsampled version of image used as input to the global branch, and multiple patches cropped from HD image are used as input to the local branch. This simple design is inspired by some previous work [17, 37]. Both branches use ViT-B/16 [10] with shared weights as the backbone and are initialized with CLIP [24] pre-trained weights. After ViT encoding, the class (global) features and grid (local) features are fused by a Global-Local Feature-aware Block and the relative quality scores of the images are predicted. We propose the following two main contributions to achieve Robust Cross-scene "Relative Quality" Assessment:

(1) **Scale-shift invariant loss.** Attempting to train with data from different scenes/domain for cross-scene generalization is difficult and inappropriate due to the different quality label scales of images between scenes and the presence of domain shift. We propose to predict quality in a "relative quality space" with scale-shift invariant loss to handle this ambiguity. In a mini-batch prediction, let S be the number of scene categories in the batch and K be the sample size of each category. Then we define the scale-shift invariant loss as:

$$\mathcal{L} = \frac{1}{SK}\sum_{i=1}^{S}\sum_{j=1}^{K}\|\hat{q}_{ij} - \hat{q}^{*}_{ij}\| \tag{1}$$

where $\hat{q}_{ij}$ and $\hat{q}^{*}_{ij}$ are the scores of the prediction and ground truth after mapping them into relative quality space. For K samples in each scene, we use a simple and robust way to map predictions and ground truth to a zero-shift and unit-scaled quality space:

$$t(q) = median(q), \quad s(q) = mean(\|q - t(q)\|)$$
$$\hat{q} = \frac{q - t(q)}{s(q)}, \quad \hat{q}^{*} = \frac{q^{*} - t(q*)}{s(q^{*})} \tag{2}$$

We uses the parameters S, K and a custom `torch.utils.data.Sampler` to balance the scene richness and sample richness of the training process.

(2) **Pre-training with mixed multi-source data.** Since we train the model in relative quality space, multiple datasets can be easily blended for joint tuning. We propose two pre-training strategies and use bagging ensemble method, RQ$_{general}$ and RQ$_{portrait}$ are trained for ensembles. Specifically, we mix four datasets SPAQ [12], KonIQ-10k [14], LIVE In the Wild [13] and RBID [8] and consider them to be from four different scenes (domains). RQ$_{general}$ is pre-trained in the relative quality space and fine-tuned on the PIQ23. In addition, we construct the PIQ23-Face dataset by masking the regions outside the face in the PIQ23 image. RQ$_{portrait}$ is pretrained on PIQ23-Face with the same strategy, but using three separate Global-Local Feature-aware Blocks to perceive detail, exposure and overall quality. Finally fine-tuned on PIQ23. The two pre-training approaches greatly promote the model's cross-scene evaluation capability and the robustness of portrait evaluation. The model is illustrated in Fig. 5.

We use ViT-B/16 [10] as the backbone and are initialized with CLIP [24] **pre-trained** weights.

We pre-train using two types of datasets and then fine-tune on the PIQ23-Overall provided by the challenge. The data used for pre-training is:
(1) External public datasets: SPAQ [12], KonIQ-10k [14], LIVE In the Wild [13] and RBID [8].
(2) PIQ23-Face: Obtained by pre-processing the PIQ23 dataset. Regions other than the face in the original image are masked to 0, and the detail, exposure and overall quality scores in the original dataset are used for supervised multi-task learning.

Training: The original HD image is first resized to 244×244, then randomly crop and flip for augmentation. The cropped image of size 224×224 is used as input for the global branch. For the local branch, the HD image is also applied with random flip and then divided into 7×7=49 squares. The 32×32 sized mini-patches cropped from each square are re-spliced into 224×224 sized inputs.

Inference: The inputs for testing and training are the same. But for testing augmentation, four-corner, top, bottom, left, right, and center crops are used instead of randomly flipping and cropping the images.

Implementation details The team implemented RQ-Net by PyTorch and train it on two NVIDIA 4090 GPUs. The original HD image is first resized to 244×244, then randomly crop and flip for augmentation. The cropped image of size 224×224 is used as input for the global branch. For the local branch, the HD image is also applied with random flip and then divided into 7×7=49 squares. The 32×32 sized mini-patches cropped from each square are re-spliced into 224×224 sized inputs. The inputs for testing and training are the same. But for testing augmentation, four-corner, top, bottom, left, right, and center crops are used instead of randomly flipping and cropping the images.

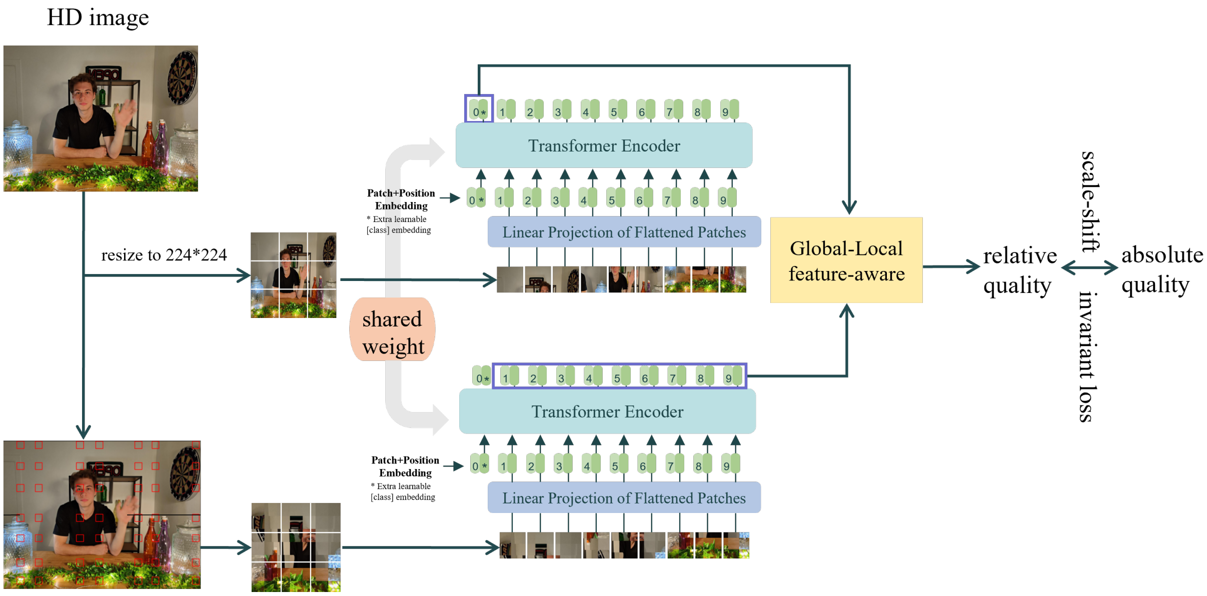

Figure 5. Diagram of the **RQ-Net** proposed by Team Xidian IPPL.

We use the Adam optimizer with weight decay of 1×10^{-5} to train RQ-Net, with mini-batch size of 128 (S=4, K=32). We use the cosine decay learning rate strategy, with a maximum learning rate of 1×10^{-5}.

Four models with the same structure were eventually trained for ensemble. RQ*general-M1*, RQ*portrait-M1*, RQ*general-M2* and RQ*portrait-M2*, where general/portrait denotes the two pre-training strategies introduced previously, and M1/M2 denotes different training set division strategies.

3.2. Ranking based vision transformer network for image quality assessment.

Team BDVQA Group

Haotian Fan, Fangyuan Kong, Yifang Xu

ByteDance Inc

The team proposed a method based on MSTRIQ Wang et al. [34], a Swin-Transformer based method. We raise several training and inference tricks to increase the performance of this method. We combined rank loss and mse loss to increase the model same-scene ranking ability.

The merged ranking loss is given by:

$$loss_{merged\_loss} = \frac{2}{N}\sum_{i=0:2:N} \begin{cases} e^{\hat{y}^i-\hat{y}^{i+1}} + (y^i-\hat{y})^2, & \text{if } y^i < y^{i+1} \\ (y^i-\hat{y})^2, & \text{others} \end{cases} \quad (3)$$

We used several data augmentation method to increase the training dataset and enhance robustness of our model: Random Crop image into patches, Random Rotation.

We also use *Test time augmentation (TTA)* can perform random modifications to the testing images. The following TTA methods are implemented to increase our model performance: (i) FiveCrop and TenCrop, (ii) Random crop then inference each image 18 times averaged.

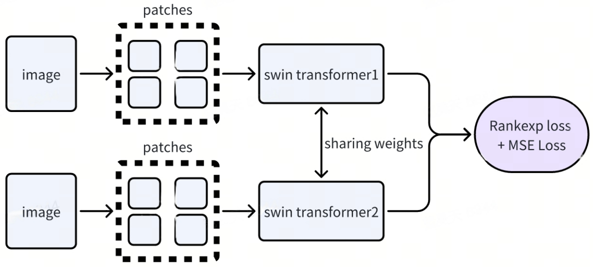

Figure 6. Siamese Swin Transformer [23] approach proposed by Team BDVQA.

We use Swin transformer [23] pre-trained on ImageNet. No additional datasets were used. The data pre-processing consists on random resized crops. The overall training method is illustrated in Fig. 6.

Implementation details The model is implemented in Pytorch. The estimated training time is 2h using 8 A100 GPUs (40G). The models are trained using AdamW optimizer and learning rate $2e^{-5}$.

The input images are augmented using random resized crop to 384 x 384. During inference, we use test time augmentation (TTA) of random crops 18 times, and average the results to produce the final output.

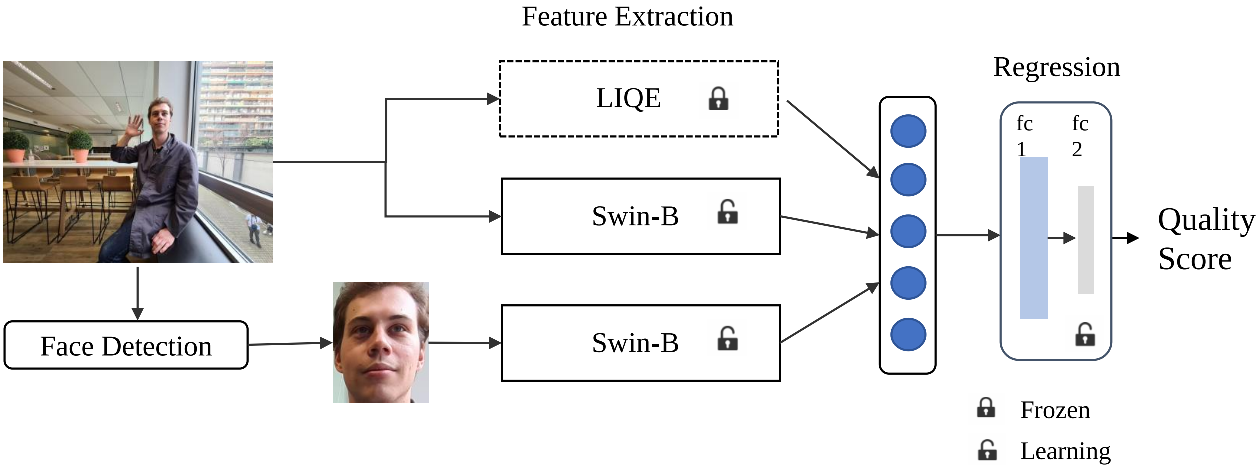

Figure 7. PQE method proposed by Team SJTU MMLab.

3.3. PQE: A Portrait Quality Evaluator by Analyzing the Characteristics of Facial and Full Images

Team SJTU MMLab

Wei Sun [1], *Weixia Zhang* [1], *Yanwei Jiang* [1], *Haoning Wu* [2], *Zicheng Zhang* [1], *Jun Jia* [1], *Yingjie Zhou* [1], *Zhongpeng Ji* [3], *Xiongkuo Min* [1], *Weisi Lin* [2], *Guangtao Zhai* [1]

[1] *Shanghai Jiao Tong University*
[2] *Nanyang Technological University*
[3] *Huawei*

Contact: sunguwei@sjtu.edu.cn

The team introduces a two-branch portrait quality assessment model motivated by the influence of both facial and background components on portrait quality. Thus, employing a single neural network on the portrait image is insufficient to model the quality relationship between the facial and the background components.

To address this problem, we propose a two-branch neural network (each branch consisting of a Swin Transformer-B [23]) for portrait quality assessment, where two branches are used to model the quality characteristics of the full and the facial components respectively.

Moreover, the shooting scene (including luminance, environment, etc.) also impact the perception of portrait quality. Therefore, we perform LIQE [42], a CLIP based scene classification and quality evaluation model, to extract scene and quality features for the full image. Subsequently, we concatenate these features and utilize a two-layer MLP to derive the quality scores. We employ the learning-to-rank training method [42] and use the fidelity loss [31] as the loss function to optimize the model.

We use LIQE, a **pre-trained** model trained on LIVE [28], CSIQ [15], KADID-10k [20], BID [8], CLIVE [13], and KonIQ-10k [14] to extract scene and quality features. The branch for the entire image is pre-trained on the LSVQ [39] dataset and the branch for the facial image is pre-trained on the GFIQA [30] dataset. For images in the PIQ23 dataset, we use yolo-face package to extract the face images from the full portrait image.

Training: We use the fidelity loss to train our model. Specifically, for an image pair ($\boldsymbol{x}$, $\boldsymbol{y}$) from the same scene in the PIQ23 dataset, we compute a binary label according to their ground-truth JODs:

$$p(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} 0 & \text{if} \quad \mathrm{q(x)} \geq \mathrm{q(y))} \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

We estimate the probability of $\boldsymbol{x}$ perceived better than $\boldsymbol{y}$ as

$$\hat{p}(\boldsymbol{x},\boldsymbol{y}) = \Phi(\frac{\hat{q}(\boldsymbol{x}) - \hat{q}(\boldsymbol{y})}{\sqrt{2}}), \tag{5}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function, and the variance is fixed to one. We adopt the fidelity loss to optimize the model:

$$\begin{aligned} \ell(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) =& 1 - \sqrt{p(\boldsymbol{x},\boldsymbol{y})\hat{p}(\boldsymbol{x},\boldsymbol{y})} \\ & - \sqrt{(1-p(\boldsymbol{x},\boldsymbol{y}))(1-\hat{p}(\boldsymbol{x},\boldsymbol{y}))}. \end{aligned} \tag{6}$$

During training, both the resolutions of full and facial images are resized to 384×384. We train the model on 2 NVIDIA RTX 3090 GPUs with a batch size of 6. The training epoches are set as 10. Learning rate is 1×10^{-5}.

We use the Swin Transformer-B as the backbone, which is a hybrid network structure. We use yolo-face as the face detector, which is a CNN network structure. LIQE is the transformer based network structure. The model is illustrated Fig. 7.

Implementation details

- Optimizer: Adam
- Learning rate: 1×10^{-5}
- GPUS: 2 NVIDIA RTX 3090
- Datasets: We use the LSVQ dataset to pre-train the branch for the full image to obtain a robust quality-aware feature representation and use the GFIQA dataset to pre-train the branch for the facial image to obtain facial-related quality feature representation. We train the whole model on the PIQ23 dataset.
- Training Time: 2 hours
- Training Strategy: Pair-wise training
- Augmentations: Randomly crop

3.4. A Mean-Opinion Network For Portrait Quality Assessment: MoNet

Team I*^{2}*Group

Zewen Chen [1,2], *Wen Wang* [3], *Juan Wang* [1], *Bing Li* [1]

[1] *State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA*
[2] *School of Artificial Intelligence, University of Chinese Academy of Sciences*
[3] *Beijing Jiaotong University*

Contact: chenzewen2022@ia.ac.cn

We take the dataset annotation process, where different annotators will annotate different opinion scores for the same image and the average of theses scores is applied as the label, namely mean opinion score (MOS). Thus, a novel network architecture called mean-opinion network (MoNet) is proposed [6]. Mimicking the human rating process, we develop a multi-view attention learning (MAL) module for the MoNet to implicitly learn diverse opinion features by capturing complementary contexts from various perspectives. The opinion features collected from different MALs are integrated into a comprehensive quality score, effectively relieving the impacts of hyper-parameter configurations on the performance, facilitating more robust quality score assessment. To be more alignment with this challenge, we additionally take a full connection (FC) layer to get the scenes classification.

Global Method Description We present a novel network called mean-opinion network (MoNet), which collects various opinions by capturing diverse attention contexts to make a comprehensive decision on the image quality score. Fig. 8 shows the network architecture of the MoNet, which mainly consists of three parts: i) a pre-trained ViT is employed for multi-level feature perception, ii) multi-view attention learning (MAL) modules are proposed for opinion collection, and iii) an image quality score regression module is designed for quality estimation.

A) Multi-level Semantic Perception. Given an image $I \in \mathbb{R}^{H \times W \times 3}$, we firstly crop it into C patches with the size of $S \times S$, where H and W denote the height and width of the image and $C = \frac{H \times W}{S^2}$. Then the patches are flattened and fed into a linear projection with the dimension of D, producing the embedding feature $\mathbf{E} \in \mathbb{R}^{C \times D}$. Subsequently, the features $\mathbf{E}$ sequentially traverses 12 transformer blocks, resulting in a set of multi-level features. Finally, the outputs of N transformer blocks are selected and used as basic features, denoted as f_j $(1 \leq j \leq N)$.

B) Multi-view Attention Learning Module. The critical part of the MoNet is the multi-view attention learning (MAL) module. The motivation behind it is that individuals often have diverse subjective perceptions and regions of interest when viewing the same image. To this end, we employ multiple MALs to learn attentions from different viewpoints. Each MAL is initialized with different weights and updated independently to encourage diversity and avoid redundant output features. The number of MALs can be flexibly set as a hyper-parameter. We show in our results its effect on the performance of our model.

As shown in Fig. 8, the MAL starts from N self-attentions (SAs), each of which is responsible to process a basic feature $\mathbf{f}_j$ $(1 \leq j \leq N)$. The outputs of all the SAs are concatenated, forming a multi-level aggregated feature $\mathbf{F} \in \mathbb{R}^{C \times D \times N}$. Then $\mathbf{F}$ passes through two branches, *i.e.*, a pixel-wise SA branch and a channel-wise SA branch, which apply a SA across spatial and channel dimensions, respectively, to capture complementary non-local contexts and generate multi-view attention maps. In particular, for the channel-wise SA, the feature $\mathbf{F}$ is first reshaped and permuted to convert the size from $C \times D \times N$ to $D \times (C \times N)$. After the SA, the output feature is permuted and reshaped back to the original size $C \times D \times N$. Subsequently, the outputs of the two branches are added and average pooled, generating an opinion feature. The design of the two branches has two key advantages. First, implementing the SA in different dimensions promotes diverse attention learning, yielding complementary information. Second, contextualized long-range relationships are aggregated, benefiting global quality perception.

C) Image Quality Score Regression. Assuming that M opinion features are generated from M MALs employed in the MoNet. To derive a global quality score from the collected opinion features, we utilize an additional MAL. The MAL integrates diverse contextual perspectives, resulting in a comprehensive opinion feature that captures essential information. This feature is then processed through a transformer block, three convolutional layers with kernel sizes of 5×5, 3×3, and 3×3 to reduce the number of channels, followed by two fully connected layers that transform the feature size from 128 to 64 and from 64 to 1. Finally, we obtain a predicted quality score from the MoNet. For the

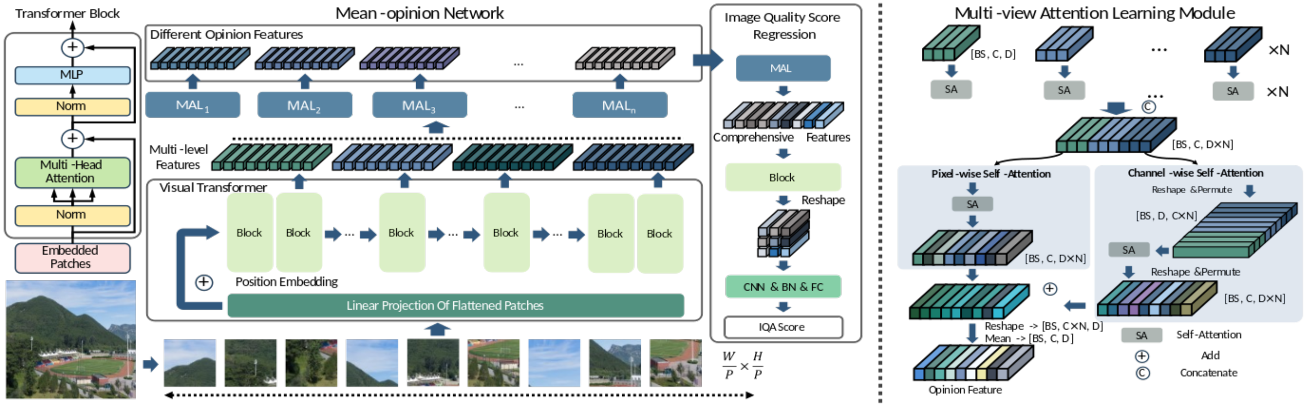

Figure 8. Network architecture of the proposed MoNet (left) and multi-view attention learning module (right).

scene classification, we additionally take a FC layer simiar to the image quality score regression.

Implementation details The pre-trained ViT model `vit_base_patch16_384` is used as the backbone of the MoNet. We use $N = 4$ transformer blocks to extract basic features, namely the 3rd, 6th, 9th and 12th blocks. The default number of the MAL is set to $M = 5$. We use the Adam optimizer with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-5}. The learning rate is adjusted using the Cosine Annealing for every 50 epochs. We train our model for 100 epochs with a batch size of 11 on one RTX3090. We take the mean square error (MSE) loss to reduce the discrepancy between the predicted scores and ground truths (GT). And we take the cross-entropy loss for scene classification.

3.5. Scene Adaptive Global Context and Local Facial Perception Network for Portrait Image Quality Assessment

Team SECE-SYSU

Xiaoqi Wang, Junqi Liu, Zixi Guo, and Yun Zhang

School of Electronics and Communication Engineering,
Sun Yat-sen University, China

Contact: `zhangyun2@mail.sysu.edu.cn`

The facial region is pivotal for portrait image quality evaluation, yet it typically occupies only a small portion of the entire image, which poses a challenge for deep neural networks that tend to capture global semantics and context. Furthermore, scene-dependent variations in portrait quality scores introduce additional complexities [5]. To address these issues, this solution proposes a scene-adaptive global context and local facial perception network. The proposed method first leverages a face detector [16] to precisely localize the facial region within the global image. Then, vision Transformer is employed to model quality-centric embeddings of both local facial region and global image. To address scene-specific quality biases, we formulate a scene recognition task and leverage scene category to adaptively select scene-specific global and local facial regressors. Finally, a global local gating network dynamically adjusts the weighting of quality predictions from the two branches, resulting in the final quality score.

Global Method Description We propose a novel solution that scene-adaptively evaluates the global image and detected local facial image through a face detector, ultimately fusing the local and global assessments to obtain a final quality score through a global local gating network. The proposed model, illustrated in Figure 9, is structured as follows: **Face Detector** employs a lightweight Dual Shot Face Detector [16] for robust facial localization within portrait images. The initial confidence threshold is empirically set to 0.8, achieving a 99% face detection rate on PIQ23 database [3]. In the absence of a detected face at this threshold, the Detector iteratively decreases the confidence threshold (0.7, 0.6, 0.5, 0.4, 0.2) until at least one face is successfully identified. The longer edge of the detected face bounding box determines the cropping dimension, ensuring a minimum resolution of 512x512 pixels. **Feature Extraction Module** leverages the first 10 Transformer layers from the ViT-Base [11], initialized with pre-trained weights from the CLIP visual encoder [24]. The initial 6 blocks are frozen, and the remaining parameters are fine-tuned on PIQ23 database. The outputs of the first 6 blocks are fed into a convolutional layer and linear layer for scene classification. The global and local embeddings are derived by mapping the concatenated features from the last three ViT blocks to the ViT's embedding space via a convolutional projection. **Global Local Gating Network** is constructed to weight local and global quality predictions, comprising an input layer, two fully-connected hidden layers of

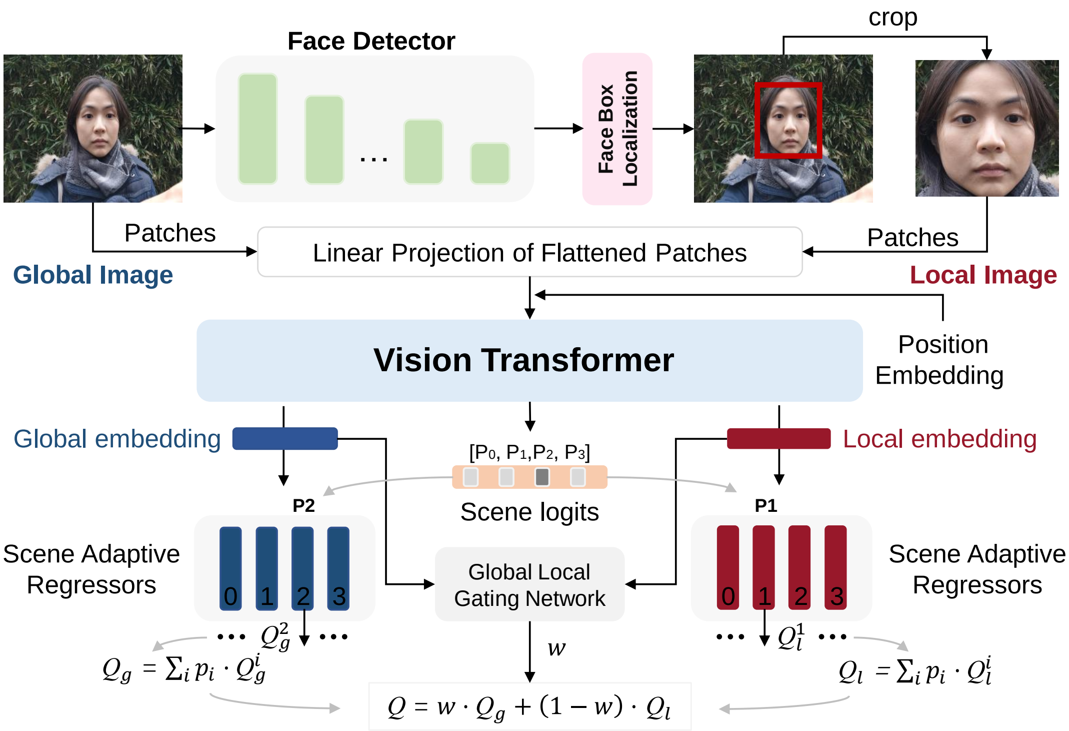
Face Detector
crop
Face Box Localization
...
Patches
Patches
Linear Projection of Flattened Patches
Global Image
Local Image
Position Embedding
Vision Transformer
Global embedding
Local embedding
[P0, P1,P2, P3]
Scene logits
P2
P1
Scene Adaptive Regressors
0 1 2 3
Global Local Gating Network
0 1 2 3
Scene Adaptive Regressors
Q_g^2
Q_l^1
w
$Q_g = \sum_i p_i \cdot Q_g^i$
$Q_l = \sum_i p_i \cdot Q_l^i$
$Q = w \cdot Q_g + (1-w) \cdot Q_l$

Figure 9. Overview the method proposed by Team SECE-SYSU. The model leverages a face detector for facial localization. A Vision Transformer then extracts quality-aware embeddings from both the local facial region and the global image. Scene classification guides the selection of scene-specific regressors for global and local quality prediction. A gating network dynamically fuses these predictions for the final quality score.

sizes 128 and 64 with ReLU activations, and a single output neuron with a sigmoid activation. **Scene Adaptive Regressors**, implemented as linear layers, are selected based on the ground truth scene categories (training phase). The model is *trained* under three scenarios: global-only, local-only, and joint global-local, with respective probabilities of 0.3, 0.3, and 0.4. During the *testing* phase, global and local image are jointly processed, and their scores are weighted by the predicted scene probabilities and individual scores. The gated fusion network then weights the local and global scores to yield the final prediction.

Implementation details The proposed model was constructed using the PyTorch framework and trained on an NVIDIA GeForce RTX 3090 GPU (24G). The experiment employed an 80-epoch training regimen with a batch size of 16. The AdamW optimizer with betas of 0.9 and 0.999 was utilized for training, initialized with a learning rate of 1e-5 and an L2 weight decay of 1e-5. A cosine annealing learning rate scheduler was adopted, with a warm-up phase reaching a maximum learning rate of 1e-4 and a minimum learning rate of 0 over 30 cycles. The objective function was the Huber loss, with a hyperparameter of 0.2. Data from the PIQ23 dataset was split, with 90% of samples from each scene used for training and the remaining 10% used for testing. During preprocessing, input images were subdivided into 224 ×224 patches. For training, a single patch was randomly sampled per image and underwent random flipping for data augmentation. The training phase took approximately 8 hours. In the testing phase, 30 patches were densely sampled from each image, and the final prediction was obtained by averaging the predicted results. The proposed model achieves an SROCC of 0.8335 and a PLCC of 0.8422 on PIQ23 following the scene-based data partitioning of Chahine et al [3]. Our model achieves an inference time of 755ms per image on an Intel i5-12400F CPU with 16GB RAM and an NVIDIA GeForce RTX 2080 8GB GPU (model complexity details in Table 3).

Module	Face detector	Other modules
MACs	3.36G	26.11G
FLOPs	3.36G	35.43G
Params	14.22M	135.02M
Inference time (per image): 755 ms		

Table 3. Efficiency study of Team SECE-SYSU.

Acknowledgements

This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2024 sponsors: Meta Reality Labs, OPPO, KuaiShou, Huawei and University of Würzburg (Computer Vision Lab).

The organizers thank the DXOMARK engineers and photographs for their time investment and fruitful discussions about image quality.

References

[1] Cosmin Ancuti, Codruta O Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[2] Nikola Banić, Egor Ershov, Artyom Panshin, Oleg Karasev, Sergey Korchagin, Shepelev Lev, Alexandr Startsev, Daniil Vladimirov, Ekaterina Zaychenkova, Dmitrii R Iarchuk, Maria Efimova, Radu Timofte, Arseniy Terekhin, et al. NTIRE 2024 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[3] Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Théo Cayla, Sira Ferradans, and Jean Ponce. An image quality assessment dataset for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9968–9978, 2023. 2, 3, 4, 9, 10

[4] Nicolas Chahine, Marcos V. Conde, Gabriel Pacianotto, Daniela Carfora, Benoit Pochon, Sira Ferradans, Radu Timofte, et al. Deep portrait quality assessment. A NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 2

[5] Nicolas Chahine, Sira Ferradans, Javier Vazquez-Corral, and Jean Ponce. Generalized portrait quality assessment. *arXiv preprint arXiv:2402.09178*, 2024. 2, 4, 9

[6] Zewen Chen, Juan Wang, Bing Li, Chunfeng Yuan, Weiming Hu, Junxian Liu, Peng Li, Yan Wang, Youqun Zhang, and Congxuan Zhang. Gmc-iqa: Exploiting global-correlation and mean-opinion consistency for no-reference image quality assessment. *arXiv preprint arXiv:2401.10511*, 2024. 8

[7] Zheng Chen, Zongwei WU, Eduard Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. NTIRE 2024 challenge on image super-resolution (×4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[8] Alexandre Ciancio, Eduardo AB da Silva, Amir Said, Ramin Samadani, Pere Obrador, et al. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing*, 20(1):64–75, 2010. 5, 7

[9] Marcos V. Conde, Florin-Alexandru Vasluianu, Radu Timofte, et al. Deep raw image super-resolution. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 5

[11] 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. 9

[12] 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. 5

[13] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 5, 7

[14] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 5, 7

[15] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010. 7

[16] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 9

[17] Leida Li, Tianshu Song, Jinjian Wu, Weisheng Dong, Jiansheng Qian, and Guangming Shi. Blind image quality index for authentic distortions with local and global deep feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8512–8523, 2022. 5

[18] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[19] Jie Liang, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Radu Timofte, Lei Zhang, et al. NTIRE 2024 restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[20] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 7

[21] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[22] Xiaoning Liu, Zongwei WU, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[23] 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, 7

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 9

[25] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. The beauty of capturing faces: Rating the quality of digital portraits. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2015. 1

[26] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth NTIRE 2024 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[27] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–49, 2022. 1

[28] 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. 7

[29] 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, 4

[30] Shaolin Su, Hanhe Lin, Vlad Hosu, Oliver Wiedemann, Jinqiu Sun, Yu Zhu, Hantao Liu, Yanning Zhang, and Dietmar Saupe. Going the extra mile in face image quality assessment: A novel database and model. *IEEE Transactions on Multimedia*, 2023. 7

[31] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 383–390, 2007. 7

[32] Oliver van Zwanenberg, Sophie Triantaphillidou, Robin Jenkin, and Alexandra Psarrou. Edge detection techniques for quantifying spatial imaging system performance and image quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[33] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei WU, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[34] Jing Wang, Haotian Fan, Xiaoxia Hou, Yitian Xu, Tao Li, Xuechao Lu, and Lean Fu. Mstriq: No reference image quality assessment based on swin transformer with multi-stage fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1269–1278, 2022. 6

[35] Longguang Wang, Yulan Guo, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, Radu Timofte, et al. NTIRE 2024 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[36] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2024 challenge on light field image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[37] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pages 538–554. Springer, 2022. 5

[38] Ren Yang, Radu Timofte, et al. NTIRE 2024 challenge on blind enhancement of compressed image: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[39] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14019–14029, 2021. 7

[40] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2024 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[41] Emin Zerman, Giuseppe Valenzise, and Frederic Dufaux. An extensive performance evaluation of full-reference hdr image quality metrics. *Quality and User Experience*, 2(1):1–16, 2017. 2

[42] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 7

[43] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. NTIRE 2024 challenge on bracketing image restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2