CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/185127/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Wang, Huasheng, Liu, Jiang, Tan, Hongchen, Lou, Jianxun, Liu, Xiaochang, Zhou, Wei, Chen, Ying, Whitaker, Roger , Colombo, Walter and Liu, Hantao 2026. KSIQA: A knowledge-sharing model for no-reference image quality assessment. IEEE Transactions on Neural Networks and Learning Systems 10.1109/tnnls.2026.3656757

Publishers page: https://doi.org/10.1109/tnnls.2026.3656757

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

KSIQA: A Knowledge-Sharing Model for No-Reference Image Quality Assessment

Huasheng Wang, Jiang Liu, *Graduate Student Member, IEEE*, Hongchen Tan, Jianxun Lou, *Member, IEEE*, Xiaochang Liu, Wei Zhou, Ying Chen, *Senior Member, IEEE*, Roger Whitaker, Walter Colombo, *Member, IEEE*, and Hantao Liu, *Member, IEEE*

***Abstract*—No-reference image quality assessment (NR-IQA) aims to quantitatively measure human perception of visual quality without comparing a distorted image to a reference. Despite recent advances, existing NR-IQR approaches often demonstrate insufficient ability to capture perceptual cues in the absence of a reference, limiting their generalisability across diverse and complex real-world image degradations. These limitations hinder their ability to match the reliability of full-reference IQA (FR-IQA) counterparts. A key challenge, therefore, is to enable NR-IQA models to emulate the reference-aware reasoning exhibited by humans and FR-IQA methods. To address this challenge, we propose a novel NR-IQA model based on a knowledge-sharing (KS) strategy to simulate this capability and predict image quality more effectively. Specifically, we designate an FR-IQA model as the teacher and an NR-IQA model as the student. Unlike conventional knowledge distillation (KD), our proposed architecture enables the NR-IQA student and FR-IQA teacher to share a decoder rather than being independent models. Furthermore, the student model contains a Mental Imagery Generation (MIG) module to learn mental imagery as the reference. To fully exploit local and global information, we adopt a vision transformer (ViT) branch and a convolutional neural network branch for feature extraction (FE). Finally, a quality-aware regressor (QAR) combined with deep ordinal regression is constructed to infer the quality score. Experiments show that our proposed NR-IQA model, KSIQA, has class-leading performance against current no-reference (NR) techniques across widespread benchmark datasets.**

***Index Terms*—Image quality, objective metric, perception, subjective experiment.**

I. Introduction

IMAGE quality assessment (IQA) has emerged as a pivotal technique in a wide range of real-world applications [1], [2], [3], [4], [5], [6], [7], including monitoring the perceptual quality of compressed content on social media, evaluating the performance of image restoration algorithms (e.g., denoising and super-resolution), and optimizing image processing pipelines in modern camera systems. In contrast to the laborious and expensive subjective methods that are used to obtain IQA ratings from human viewers, objective IQA models that adopt computational algorithms to emulate the visual quality perception of the human visual system (HVS) offer a rapid and flexible technological solution. IQA models can be generally classified into full-reference IQA (FR-IQA) [8], [9], [10], reduced-reference IQA (RR-IQA) [11], and no-reference IQA (NR-IQA) [12], [13], [14], [15], depending on the degree of usage of the pristine/reference image in the model. Since a reference is often rarely available in real-world scenarios, NR-IQA has recently gained significant attention due to its high practical relevance.

Deep learning techniques have significantly advanced the NR-IQA, and state-of-the-art (SOTA) NR-IQA models benefit from adopting various convolutional neural network (CNN) architectures and vision transformer (ViT), as well as the utilization of novel loss functions. For example, MANIQA [13] presents the Transposed Attention Block and the Scale Swin Transformer (ST) Block as innovations to enhance global and local feature interactions. Meanwhile, the multi-dimensional strategy used in MANIQA efficiently facilitates interactions amongst diverse image regions. To capture the relative ordering between different scores on the perceptual quality scale, a novel loss function, i.e., deep ordinal loss (DO-loss), is proposed in [12] to boost the prediction performance of deep learning-based NR-IQA.

Despite the advancements made by existing NR-IQA methods, there is considerable scope for performance improvements. Innovations are driven by the observations of viewers' behavior in assessing image quality. Based on the assumption that humans use visual priors of high-quality (HQ) images to judge the perceived quality of the low-quality (LQ) images in the NR-IQA context, the model in [16] incorporates explicit HQ prior distributions—using the so-called nonaligned reference (NAR) images with similar scenes to LQ inputs—to measure the distributional disparities between HQ and LQ images. The requirement of using similar HQ scenes to simulate the priors limits the practicality of this method in real-world scenarios. To address this problem, the method proposed in [17] employs content-variant HQ images as reference

Received 20 February 2025; revised 25 July 2025 and 27 October 2025; accepted 12 January 2026. *(Corresponding author: Jiang Liu.)*

Huasheng Wang and Ying Chen are with Alibaba Group, Hangzhou 310023, China (e-mail: lufei.whs@taobao.com).

Jiang Liu, Wei Zhou, Roger Whitaker, Walter Colombo, and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. (e-mail: liuj137@cardiff.ac.uk; zhouw26@cardiff.ac.uk; WhitakerRM@cardiff.ac.uk; ColomboG@cardiff.ac.uk; liuh35@cardiff.ac.uk).

Hongchen Tan is with the College of Future Technology, Dalian University of Technology, Dalian 116024, China (e-mail: tanhongchenphd@bjut.edu.cn).

Jianxun Lou is with the School of Computer Science, Northeast Electric Power University, Jilin 132013, China (e-mail: jianxunlou@neepu.edu.cn).

Xiaochang Liu is with the School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China (e-mail: liuxch68@mail2.sysu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2026.3656757

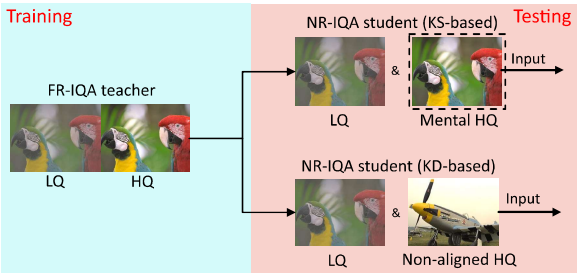


Fig. 1. Comparison of the proposed KS strategy (top branch) versus conventional KD strategy (bottom branch) for transferring HQ and LQ distribution difference information from FR-IQA teacher to NR-IQA student. Note that the KD student model requires both nonaligned HQ and LQ images as input, while the KS student model (proposed) only requires an LQ image as input—a mental HQ imagery is generated from the LQ to simulate the mental reference.

priors and explores the transfer of the HQ–LQ distribution difference information of FR-IQA to the HQ–LQ disparity information of NR-IQA using knowledge distillation (KD). KD approaches typically adopt an offline training paradigm, where an FR-IQA teacher model (comprising its own encoder and decoder) is first trained and then frozen. Subsequently, a separate NR-IQA student model (also with its own encoder and decoder) is trained to replicate intermediate outputs of the fixed teacher. However, this unidirectional and decoupled knowledge transfer poses several intrinsic limitations: 1) the frozen teacher cannot adapt or coevolve with the student during training; 2) the architectural separation between the models may lead to misalignment in their respective feature spaces; and 3) such methods often require both the distorted image and external nonaligned or content-variant HQ image during inference, limiting their practicality as truly single-input NR-IQA solutions. To tackle these challenges, in this article, we propose to enable tightly integrated, decoder-level knowledge sharing (KS) between the FR-IQA teacher and the NR-IQA student. This unified architecture provides an effective form of implicit inductive bias alignment. More specifically, it allows the student model to learn latent quality representations without explicitly mimicking the teacher's outputs, which helps mitigate the feature space misalignment issue often observed in KD. This approach simulates reference-aware reasoning in a structurally coherent way, thus improving both learning efficiency and quality prediction accuracy.

This approach gives a more efficient and reliable deep learning-based approach for NR-IQA. In particular, the novel teacher–student architecture only requires information from reference images during the training phase and abstains from using any reference information/priors during the inference phase. As previously noted [6], when humans evaluate the quality of a single distorted image, they tend to compare it to a mental model of a reference and render the assessment by comparisons. To simulate this process in a coherent and efficient structure, as shown in Fig. 1, we employ an imagery counterpart generation module to construct a mental imagery, representing the pristine reference for the distorted input of the NR-IQA student model. Meanwhile, we devise a novel KS strategy where the FR-IQA teacher and NR-IQA student networks are designed to share the same regressor. Critically, these networks are then jointly trained to enhance the student's feature expression capability.

To the best of our knowledge, this article pioneers the introduction of the novel KS strategy, which involves the sharing of the decoder from the teacher model and simultaneous joint training with the student model. The contributions of this article are as follows.

1) A novel deep learning network architecture for NR-IQA, named knowledge-sharing IQA (KSIQA), is proposed. In contrast to the conventional KD strategy, KSIQA provides an efficient knowledge transfer approach where the FR-IQA teacher model and NR-IQA student model share a decoder and are jointly trained for enhanced collaborative learning.
2) A streamlined simulation of visual priors for NR-IQA is proposed. An imagery counterpart generation model is built into the network architecture to stimulate the mental imagery of the reference image.
3) The proposed KSIQA consistently outperforms SOTA no-reference (NR) IQA models across benchmark IQA datasets. Notably, our architecture demonstrates a strong generalization ability.

II. Related Work

A. Image Quality Assessment

There have been substantial developments in IQA. Traditional methods aim to simulate the sensitivity of the HVS to different image signals by incorporating various explicit vision models, such as the modeling of perceived structural information in natural scenes [8] and natural scene statistics (NSS) [18], [19], [20]. In recent years, deep learning-based IQA models have shown significant advancements, providing more accurate and reliable prediction of perceived image quality [16], [17], [21], [22], [23], [24], [25], [26]. In contrast to traditional IQA methods that rely on combining handcrafted image features, deep learning-based IQA methods have the advantage of directly extracting discriminatory features from images. SOTA deep learning techniques have been successfully applied in IQA [12], [13], [27], [28], [29], [30], [31], [32], [33], [34]. For example, AHIQ [21] introduces a dual-branch feature extraction (FE) backbone, where one branch utilizes ViT to capture global semantic features, while the other branch utilizes convolutional neural network (CNN) to capture local image features. Meta-IQA [29] employs metalearning to train a network to distinguish different distortion types, incorporating prior knowledge of IQA. Hyper-IQA [34] separates features into low- and high-level components, enhancing the former through transformations of the latter. TReS [32] combines CNNs and transformer features for both local and nonlocal representations of distorted images. It also employs self-consistency and ranking loss to improve model robustness. DOR-IQA [12] integrates the deep ordinal loss (DO-loss) function into the IQA model to enhance the accuracy of prediction. MANIQA [13] introduces a multidimensional attention network, effectively utilizing multidimensional interactions in both channel and spatial dimensions. It should be noted that

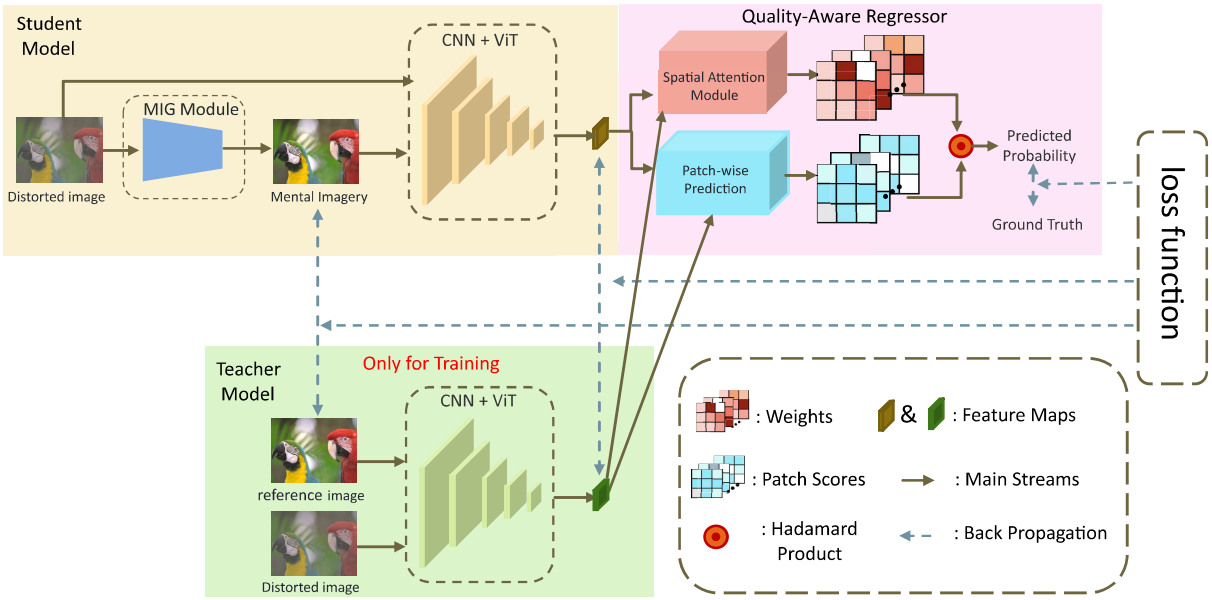


Fig. 2. Schematic overview of our proposed KSIQA framework. It consists of an FR-IQA teacher and an NR-IQA student that share a decoder and are jointly trained. After training, the NR-IQA student is utilized for inference. The architecture comprises three pivotal elements: the MIG module to learn to generate mental imagery as the reference; the FE module that combines CNN and ViT, and the QAR to infer a quality score.

NR-IQA is a more challenging problem than FR-IQA, albeit with the application of advanced deep learning technologies. Recently, an NR-IQA framework [15] has been proposed that exploits vision–language correspondence within a multitask learning paradigm. Similarly, the method in [35] trains multimodal models to align with text-defined quality levels. However, both approaches still face difficulties in accurately modeling the nuanced and continuous characteristics of human perceptual judgments.

B. Knowledge Transfer

Knowledge transfer between models (i.e., often referred to as teacher and student models) has been extensively studied and is relevant to our work. This concept has been further advanced in [36] as KD, which aims to make a student model's distribution match that of a teacher model. Cross-modal distillation has been widely applied in deep learning to transfer knowledge from one neural network to another network to solve various visual tasks, e.g., action recognition [37] and depth estimation [38]. However, these traditional KD methods rely on static, one-way knowledge transfer, which is often constrained by the capacity gap between the teacher and student models [39]. Recent research has explored architectural innovations to enhance knowledge transfer beyond simply matching output distributions. For example, Chen et al. [40] proposed a strategy to directly use the classifier from a pretrained teacher model for the student model and only train the student's encoder to align its feature representations. This method achieves competitive performance, demonstrating the value of sharing decision-making components. Similarly, Ben-Baruch et al. [41] investigated classifier sharing for enhanced representation distillation, showing improved accuracy across various tasks and confirming the effectiveness of sharing decision layers to facilitate more efficient knowledge transfer.

In the area of IQA, the CVRKD-IQA model [17] transfers the knowledge of discrepancy information between HQ–LQ images learned from an FR-IQA teacher to encourage an NR-IQA student to reproduce the same representations. These KD-based IQA models typically employ distinct and independent models for teacher and student, following an offline process where the teacher is first pretrained and then frozen, which results in a static knowledge flow that struggles to bridge the inherent capacity gap between the models. In the proposed KS approach, we design a novel network architecture using a shared decoder for both the student and teacher models. Moreover, the teacher and student networks are jointly trained, facilitating collaborative learning and knowledge exchange throughout the training process.

III. Method

We present our KSIQA framework. As illustrated in Fig. 2, it employs paired reference and distorted images as input for the teacher model, and the distorted image alone as the input for the student model. The architecture comprises three pivotal elements: the mental imagery generation (MIG) module, the FE module, and the quality-aware regressor (QAR).

A. Mental Imagery Generation Module

To address the challenge posed by the wide range of distortion types present in IQA applications, we introduce an MIG module to generate robust imagery as a proxy reference for the input to the NR-IQA model, with supervision provided by the reference input of the FR-IQA model. Conceptually, the MIG module seeks to simulate the potential reference imagery for the distorted input, mimicking a projection from the mind.

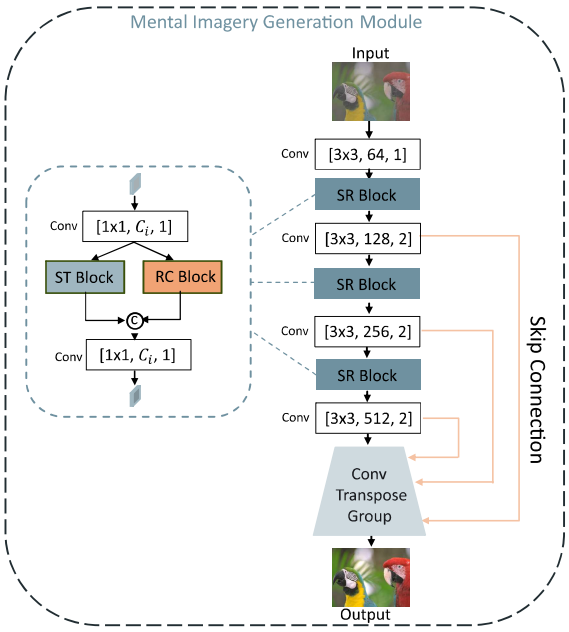


Fig. 3. Schematic of the imagery counterpart generation (MIG) module. The module generates a "mental imagery" as a proxy reference from a given distorted input, employing a UNet-like backbone built with SR blocks that combine ST and RC blocks.

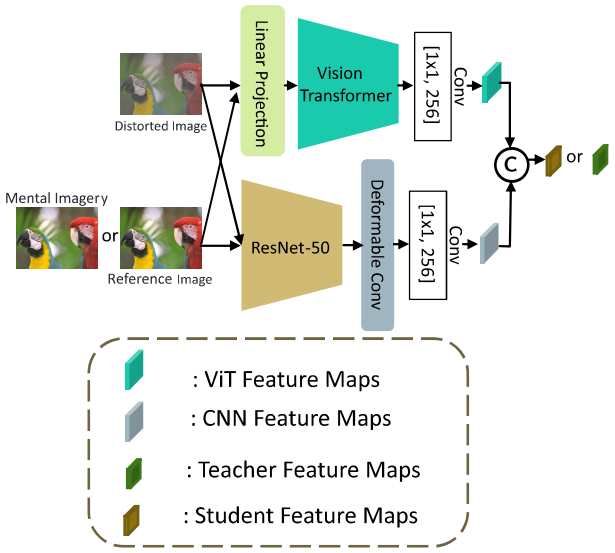


Fig. 4. Schematic of the FE module. The module consists of two parallel branches: a ViT branch for capturing global semantic information, and a CNN (ResNet-50) branch for extracting local textural features.

The MIG module generates quality-relevant content from the original undistorted version of the affected image, rather than reproducing the original. The design of this module is inspired by the ResNet [42] and ST [43], as illustrated in Fig. 3.

More specifically, we denote the distorted image as $I_d \in \mathbb{R}^{224\times224\times3}$ and feed it into the MIG module. By using a 3×3 convolutional layer, we transform this image into a feature map with 64 channels, denoted as $F_d \in \mathbb{R}^{224\times224\times64}$. This feature map then undergoes enhancement via an ST block [43] and a residual convolutional (RC) block [44] that form the primary building block, i.e., the spatial reduction (SR) block. Within each SR block, the input undergoes a 1×1 convolutional operation and is followed by two separate operations, i.e., an ST block and a residual 3×3 convolutional (RC) block. The outputs of these blocks are concatenated and passed through another 1×1 convolutional layer to produce the residual of the input. It should be noted that the SR block maintains the dimensional size of feature maps passing through this block. In addition, "SConv" and "TConv" refer to 2×2 strided convolution with a stride of 2 and 2×2 transposed convolution with a stride of 2, respectively.

The primary architecture of the MIG module mirrors that of a UNet backbone, which comprises four scales. Each scale incorporates a residual connection between an Sconv layer for downscaling and a TConv layer for upscaling. The number of channels in each layer ranges from 64 to 512 across the four scales. Concerning the decoder of the MIG module, similar to the encoder, an SR block is employed prior to each TConv layer within the Convolutional Transpose (TConv) group to augment the feature representation capability.

Finally, a 3×3 convolutional layer is applied to convert these feature maps into a mental imagery of reference, denoted as $I_{\text{mi}} \in R^{224\times224\times3}$. Following this and by using feature supervision from the FR-IQA teacher network, the backward propagation process can further constrain the generation of I_{mi} to make it more closely relevant for the IQA task by gathering information from the reference image $I_r \in \mathbb{R}^{224\times224\times3}$. This effectively resolves the challenge posed by multiple distortion types existed in diverse IQA tasks.

B. FE Module

Fig. 4 illustrates the proposed dual-branch FE module, comprising a ViT branch [45] and a CNN branch [42]. The transformer-based feature extractor focuses on capturing broad semantic information, using self-attention mechanisms. This enables the network to proficiently model distant features and encode image patches into representative forms. Given the importance of local details in human perception of visual quality, the CNN branch contributes to modeling localized information.

To implement the FE module, the reference image for the FR-IQA teacher (or the simulated mental imagery for the NR-IQA student) and the distorted image are each fed into both ViT and CNN branches. Linear projection is applied to ensure the input requirements of ViT are met. For the transformer branch, the output sequences are reshaped into feature maps $f_{\text{t}} \in \mathbb{R}^{28\times28\times3840}$. For the CNN branch, we derive shallow feature maps using ResNet, resulting in $f_c \in \mathbb{R}^{56\times56\times768}$.

To align and fuse f_{t} and f_c, we employ a 1×1 convolution to convert the dimension of channels of f_{t} to 256. In addition, the feature maps output from the CNN branch undergo a deformable convolution [46], enhancing their adaptability for the subsequent feature fusion. We also use a 1×1 convolution to reduce the dimension of f_c to 256. Ultimately, the resulting f_{t} and f_c feature maps are concatenated. Now, let I_d and I_r pass through the FE module of the teacher model, generating

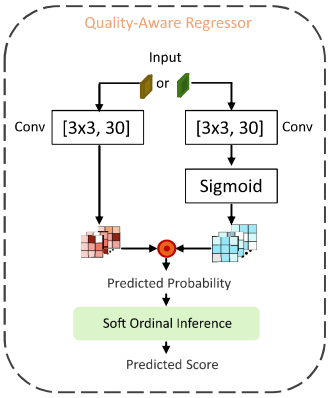


Fig. 5. Schematic of the QAR module. The module employs a dual-branch attention mechanism to produce a probability distribution of the quality scores, which is then converted into a final score using deep ordinal regression.

feature maps F^{t}_{dt}, F^{t}_{dc}, F^{t}_{rt}, and $F^{\mathrm{t}}_{\mathrm{RC}}$. These feature maps are combined as follows:

$$F^{\mathrm{t}}_{d} = \mathrm{Concat}\left[F^{\mathrm{t}}_{dc}, F^{\mathrm{t}}_{dt}\right]; F^{\mathrm{t}}_{r} = \mathrm{Concat}\left[F^{\mathrm{t}}_{\mathrm{RC}}, F^{\mathrm{t}}_{rt}\right]. \tag{1}$$

The final output of the teacher model is as follows:

$$F^{\mathrm{t}} = \mathrm{Concat}\left[F^{\mathrm{t}}_{d}, F^{\mathrm{t}}_{r}, F^{\mathrm{t}}_{d} - F^{\mathrm{t}}_{r}\right]. \tag{2}$$

Similarly, I_d and I_{mi} pass through the FE module of the student model, resulting in feature maps F^{s}_{dt}, F^{s}_{dc}, F^{s}_{mit}, and F^{s}_{mic}. These feature maps are combined as follows:

$$F^{s}_{d} = \mathrm{Concat}\left[F^{s}_{dc}, F^{s}_{dt}\right]; F^{s}_{\mathrm{mi}} = \mathrm{Concat}\left[F^{s}_{\mathrm{mic}}, F^{s}_{\mathrm{mit}}\right]. \tag{3}$$

The final output of the student model is as follows:

$$F^{s} = \mathrm{Concat}\left[F^{s}_{d}, F^{s}_{\mathrm{mi}}, F^{s}_{d} - F^{s}_{\mathrm{mi}}\right]. \tag{4}$$

Notably, the hybrid FE module combines elements of both CNNs and transformers to leverage their respective strengths, making sure both local textural information and global semantic information are comprehensively modeled.

C. QAR Module

Acknowledging that each element within a deep feature map corresponds to a unique patch of the input image, the spatial dimension's content is indispensable for overall image quality perception. However, the application of conventional pooling strategies, such as max-pooling and average-pooling, incurs information loss and dismisses the intricate interconnections amongst image patches. To overcome this problem and inspired by the approach taken in [21], we construct a QAR module, consisting of a patchwise prediction branch and a spatial attention branch, as depicted in Fig. 5. The input $F^{s} \in \mathbb{R}^{28\times28\times256}$ or $F^{\mathrm{t}} \in \mathbb{R}^{28\times28\times256}$ is fed into the module, where the prediction branch computes the probability score for each element in the feature map, while the spatial attention branch calculates a corresponding attention map after a sigmoid operation. The final score is obtained through a weighted summation of the individual probability scores. In this article, the interval for the probability distribution is set to 30 as per the findings of [12]. This operation can be expressed as follows:

$$s_p = \frac{s \odot w}{\sum w} \tag{5}$$

where s that has the dimensions of $28 \times 28 \times 30$ represents the probability map of the scores, w that has the dimensions of $28 \times 28 \times 30$ corresponds to the associated attention map, and $\odot$ denotes the elementwise Hadamard product. Finally, $s_p \in \mathbb{R}^{1\times1\times30}$ signifies the predicted score's probability. It should be noted that the Hadamard product preserves the dimensionality of the channels, hence the resulting s_p maintains a 30-D confidence distribution of score probabilities.

Rather than utilizing the mean-squared error (MSE) loss between the predicted score and the ground truth for training, we incorporate deep ordinal regression [12]. This approach enables the model to factor in the relative ordering between different ratings on the perceptual quality scale, using a soft ordinal inference to transform the predicted probabilities to a continuous variable for image quality.

D. Loss Function

Unlike the conventional KD task in [17], our KSIQA model employs a shared QAR between the teacher and student models. This approach offers two advantages. First, it reduces the overall number of model parameters, improving model efficacy and generalization. Second, the need for a multitude of loss functions is circumvented by joining teacher and student models to collaboratively train the QAR. We incorporate a total of three loss functions for our KSIQA model: 1) the MSE loss between I_r and I_{mi}; 2) the MSE loss to enforce the teacher model's feature maps F_{t} to guide the generation of the student model's feature maps F_s; and 3) the DO-loss to constrain the final outputs s_{pt} and s_{ps}. The final loss formulation is as follows:

$$\begin{aligned} L_{\mathrm{total}} = {} & L_2\left(I_r, I_{\mathrm{mi}}\right) + L_2\left(F^{\mathrm{t}}_{dt}, F^{s}_{dt}\right) + L_2\left(F^{\mathrm{t}}_{rt}, F^{s}_{\mathrm{mit}}\right) \\ & + L_2\left(F^{\mathrm{t}}_{dc}, F^{s}_{dc}\right) + L_2\left(F^{\mathrm{t}}_{\mathrm{RC}}, F^{s}_{\mathrm{mic}}\right) \\ & + 0.5 \times L_{\mathrm{DO}}\left(s_{ps}, s_{gt}\right) + 0.5 \times L_{\mathrm{DO}}\left(s_{pt}, s_{gt}\right) \end{aligned} \tag{6}$$

where s_{gt} denotes the discretized ground truth scores as detailed in [12]. The coefficient of L_{DO} is introduced to balance its contribution to the final loss as the sublosses' values use different scales. This is to ensure that the values computed by all subloss functions during training remain within a consistent magnitude.

IV. Experimentation

We conduct a comparative analysis of our proposed KSIQA against SOTA IQA methods in terms of accuracy and generalization. Also, we perform ablation studies to rigorously evaluate the effectiveness of our designed individual modules.

A. Experimental Setting

1) Datasets and Usage: We use five widely used IQA datasets, including LIVE [47], CSIQ [48], TID2013 [49], KADID-10k [50], and PIPAL [51], as summarized in Table I. The MOS distributions of these IQA datasets are shown in

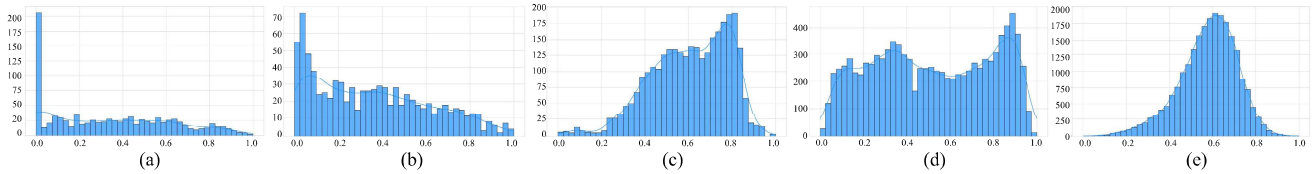


Fig. 6. MOS distributions of IQA datasets. (a) LIVE. (b) CSIQ. (c) TID2013. (d) KADID-10k. (e) PIPAL.

TABLE I
SUMMARY OF IQA DATASETS. "REF." DENOTES THE NUMBER OF REFERENCE IMAGES, "DIST." DENOTES THE NUMBER OF DISTORTED IMAGES, AND "ENV." INDICATES THE DATA COLLECTION ENVIRONMENT

Dataset	Ref.	Dist.	Env.
LIVE [47]	29	779	lab
CSIQ [48]	30	866	lab
TID2013 [49]	25	3,000	lab
KADID-10k [50]	81	10.1k	crowdsourcing
PIPAL [51]	200	23.2k	crowdsourcing

Fig. 6. We use laboratory-based datasets, i.e., LIVE, CSIQ, and TID2013, to evaluate the perceptual relevance of IQA models; and crowdsourcing-based datasets, i.e., KADID-10k and PIPAL, to validate the generalization capability of IQA models. We partition each dataset into training, validation, and test sets using a ratio of 6:2:2 based on the reference images, using the same protocol used in SOTA IQA comparative studies [21]. This ensures that the test and validation data remain undisclosed throughout the training procedure. The validation set assists in selecting the model with optimal performance, while the test set serves to assess the final performance.

2) Evaluation Criteria: We evaluate an IQA model's performance using Pearson's linear correlation coefficient (PLCC) and Spearman's rank-order correlation coefficient (SROCC). PLCC measures the linear correlation between predicted quality scores and ground truth, while SROCC quantifies the extent of monotonic correlation. Higher values of these coefficients indicate better model performance.

3) Implementation: Since we utilize ViT [45] and ResNet [42] that are pretrained on ImageNet [55], we normalize all input images and randomly crop them into a size of 224 $\times$ 224 pixels. We use the outputs of five intermediate blocks $\{0, 1, 2, 3, 4\}$ in ViT, each of which consists of a self-attention module and a feed-forward network. Building upon the principles of DOR-IQA [12], we discretize the quality scores into 30 intervals of equal size within the QAR to obtain a probability distribution. During the training process, data augmentation techniques, including horizontal flipping and random rotation, are applied. The batch size is set to be 16. For optimization, we employ the AdamW optimizer with an initial learning rate $lr = 10^{-4}$ and weight decay of 10^{-5}. The learning rate for each parameter group is determined using a cosine annealing schedule, where $\eta_{\max}$ is set to be the initial learning rate and the number of epochs T_{cur} is set to be 50. Our proposed model, KSIQA, is implemented using PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. It should be noted that throughout the validation and testing phases, we exclusively use the NR-IQA student model, where only the distorted image is used as the input.

4) Choice of Models for Comparison: To facilitate IQA model comparison in a consistent manner, we reproduce the protocols in [13] and [21]. More specifically, for the standard evaluation in Tables II and IV, the results for the FR-IQA models are acquired from [21], whereas the results for NR-IQA models are acquired from [13]. In the cross-dataset validation results of Tables V and VI, the FR-IQA models' results are sourced from [21]. For the NR-IQA models, we implement CVRKD-IQA and MANIQA using their publicly available code to achieve the optimal results for a fair and critical comparison.

5) Statistical Significance Testing: To assess whether the performance differences between IQA models are statistically significant, we perform hypothesis testing following the methodology outlined in [56]. Significance tests are conducted using the designated test set for each IQA dataset, which comprises 20% of the total samples. On the test set, each model yields residual values computed as the difference between predicted and ground-truth quality scores. The performance comparison between two IQA models is based on their respective sets of residuals. If the residual samples satisfy normality assumptions, we apply either a paired t-test (for comparisons between model variants, such as ablation studies) or an independent samples t-test (for comparisons between different IQA models). When normality assumptions are violated, we instead employ nonparametric alternatives: the Wilcoxon signed-rank test (for model variants) or the Mann–Whitney U test (for different IQA models).

B. Comparison to the SOTA

We present a thorough comparison of our proposed model, KSIQA, and SOTA IQA models, including FR-IQA and NR-IQA. Note, we classify the NAR IQA models to the NR-IQA category as they do not require a perfectly aligned reference image. First, we quantitatively assess the effectiveness and perceptual relevance of KSIQA, using three laboratory-based IQA datasets and one crowdsourcing dataset to compare the results to that of other NR-IQA models. As shown in Table II, our KSIQA achieves SOTA performance on three widely used benchmark datasets, including LIVE, CSIQ, and TID2013, and demonstrates highly competitive results on KADID-10k, where it performs comparably to top-performing models, such as MANIQA. The results of the statistical significance testing are shown in Table III, indicating that our proposed model is statistically significantly ($P < 0.05$

TABLE II

PERFORMANCE COMPARISON OF THE PROPOSED KSIQA VERSUS SOTA NR-IQA MODEL ON FOUR STANDARD IQA DATASETS. BOLD ENTRIES INDICATE THE BEST PERFORMANCE

Method	LIVE		CSIQ		TID2013		KADID-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
MEON [52]	0.955	0.951	0.864	0.852	0.824	0.808	0.691	0.604
WaDIQaM [30]	0.955	0.96	0.844	0.852	0.855	0.835	0.752	0.739
DBCNN [53]	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851
TIQA [54]	0.965	0.949	0.838	0.825	0.858	0.846	0.855	0.85
MetaIQA [29]	0.959	0.96	0.908	0.899	0.868	0.856	0.775	0.762
P2P-BM [33]	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.84
HyperIQA [34]	0.966	0.962	0.942	0.923	0.858	0.84	0.845	0.852
TReS [32]	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.915
CVRKD-IQA [17]	0.963	0.951	0.953	0.941	0.911	0.899	0.871	0.862
LIQE [15]	0.972	0.970	0.936	0.938	0.883	0.863	0.863	0.860
Q-align [35]	0.975	0.977	0.961	0.944	0.893	0.891	0.876	0.874
DOR-IQA [12]	0.978	0.977	0.961	0.945	0.901	0.887	0.885	0.883
MANIQA [13]	0.983	0.982	0.968	0.961	0.943	0.937	**0.939**	**0.938**
KSIQA (Ours)	**0.987**	**0.986**	**0.976**	**0.971**	**0.955**	**0.953**	0.923	0.919

TABLE III

RESULTS OF STATISTICAL SIGNIFICANCE TESTING FOR PERFORMANCE COMPARISON BETWEEN KSIQA (OURS) AND OTHER TOP-PERFORMING NR-IQA MODELS, INCLUDING MANIQA, DOR-IQA, AND CVRKD-IQA, ON THE LABORATORY-BASED BENCHMARKS LIVE, CSIQ, AND TID2013. "*" MEANS THAT THE DIFFERENCE IS STATISTICALLY SIGNIFICANT ($P < 0.05$ AT THE 95% CONFIDENCE LEVEL). "-" MEANS THAT THE DIFFERENCE IS NOT STATISTICALLY SIGNIFICANT

Sig	LIVE	CSIQ	TID2013
KSIQA (Ours) vs. MANIQA	*	*	*
KSIQA (Ours) vs. DOR-IQA	*	*	*
KSIQA (Ours) vs. CVRKD-IQA	*	*	*

TABLE IV

PERFORMANCE EVALUATION (GENERALIZATION) OF SOTA IQA MODELS ON THE PIPAL DATASET

IQA Type	Method	Validation		Test	
		PLCC	SROCC	PLCC	SROCC
FR	PSNR	0.269	0.234	0.277	0.249
	NQM [57]	0.364	0.302	0.395	0.364
	UQI [8]	0.505	0.461	0.450	0.420
	SRSIM [58]	0.626	0.529	0.636	0.573
	DISTS [59]	0.634	0.608	0.687	0.655
	IQT [60]	0.840	0.820	0.799	0.790
	FR-teacher (Ours) [21]	**0.865**	**0.852**	**0.828**	**0.822**
NR	Brisque [61]	0.015	0.059	0.087	0.097
	NIQE [18]	0.005	0.115	0.03	0.112
	PI [62]	0.079	0.133	0.123	0.153
	MA [63]	0.129	0.131	0.173	0.224
	SSIM [64]	0.332	0.386	0.377	0.407
	FSIM [65]	0.473	0.575	0.528	0.610
	LIQE [15]	0.585	0.571	0.585	0.596
	Q-align [35]	0.408	0.415	0.414	0.426
	LPIPS-Alex [66]	0.581	0.616	0.584	0.592
	CVRKD-IQA [17]	0.688	0.673	0.672	0.661
	MANIQA [13]	0.715	0.710	0.704	0.740
	KSIQA (Ours)	**0.851**	**0.843**	**0.822**	**0.813**

at the 95% confidence level) better than the top-performing NR-IQA models, CVRKD-IQA, DOR-IQA, and MANIQA, using the statistical method as mentioned above.

TABLE V

PERFORMANCE OF CROSS-DATASET VALIDATION OF SOTA IQA MODELS, USING THE PIPAL DATASET FOR TRAINING, AND LIVE AND TID2013 DATASETS FOR TESTING

Train on		PIPAL			
Test on		LIVE		TID2013	
		PLCC	SROCC	PLCC	SROCC
FR	WaDIQaM [30]	0.837	0.883	0.741	0.698
	RADN [67]	0.878	0.905	0.796	0.747
	FR-teacher (Ours) [21]	**0.911**	**0.920**	**0.804**	**0.763**
NR	TReS [32]	0.643	0.663	0.516	0.563
	CVRKD-IQA [17]	0.807	0.801	0.594	0.586
	MANIQA [13]	0.835	0.855	0.704	0.619
	KSIQA (Ours)	**0.902**	**0.896**	**0.754**	**0.713**

TABLE VI

PERFORMANCE OF CROSS-DATASET VALIDATION OF SOTA IQA MODELS, USING THE KADID-10K DATASET FOR TRAINING, AND LIVE AND TID2013 DATASETS FOR TESTING

Train on		KADID-10k			
Test on		LIVE		TID2013	
		PLCC	SROCC	PLCC	SROCC
FR	PieAPP [10]	0.908	0.919	0.859	0.876
	WaDIQaM [30]	0.940	0.947	0.834	0.831
	LPIPS [66]	0.934	0.932	0.749	0.670
	FR-teacher (Ours) [21]	**0.952**	**0.970**	**0.899**	**0.901**
NR	MANIQA [13]	0.864	0.849	0.745	0.726
	CVRKD-IQA [17]	0.917	0.913	0.733	0.691
	KSIQA (Ours)	**0.941**	**0.945**	**0.856**	**0.853**

To further validate the generalization capability of our KSIQA model, we conduct a comparative study on a challenging crowdsourcing-based IQA dataset, PIPAL, which contains the largest number of reference images as well as the distorted images. Note, we use the publicly available data of the PIPAL dataset for reporting on the results of IQA models. As shown in Table IV, KSIQA and its FR-IQA teacher demonstrate exceptional performance, surpassing their alternative IQA models in the literature. Notably, our KSIQA significantly outperforms

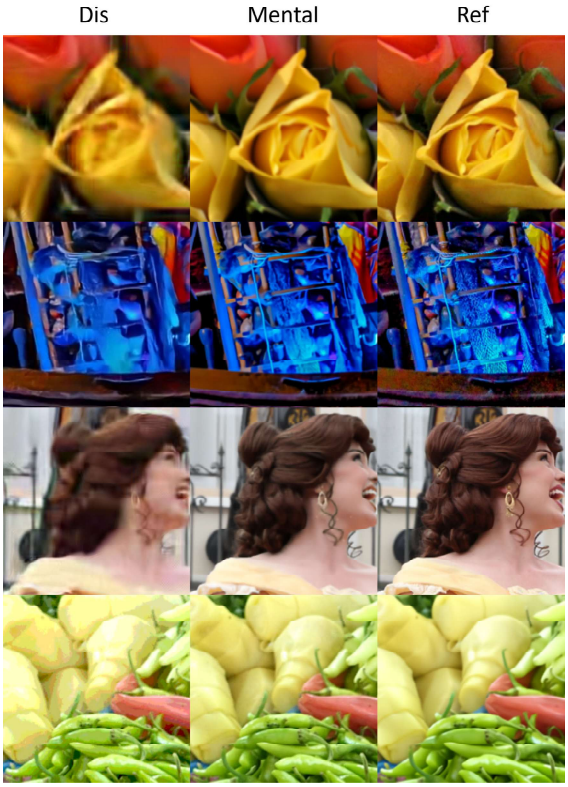


Fig. 7. Schematic of some representative examples of the results of the MIG module for generating the mental imagery. The first column displays the distorted images (Dis), the second column illustrates the mental imagery (mental), and the last column presents the reference images (Ref).

the best-performing NR-IQA model, i.e., MANIQA, with a PLCC improvement of 10.9% and an SROCC improvement of 6.8%. The results collectively demonstrate the KSIQA's capability to effectively address the more complex IQA scenarios of diverse original visual content and a wide spectrum of image distortion types.

To have a more critical evaluation of the generalization capability of our KSIQA model, we conduct a cross-dataset validation, using two laboratory-based IQA datasets, LIVE and TID2013, and two crowdsourcing-based datasets, PIPAL and KADID-10k. More specifically, we train an IQA model on a large-scale in-the-wild dataset (i.e., PIPAL or KADID-10k) and test it on two perception-based datasets (i.e., LIVE and TID2013). As the results shown in Tables V and VI, our proposed KSIQA and its FR-IQA teacher demonstrate a robust generalization ability in comparison to their respective alternative top-ranked models.

C. Ablation Studies

Now, we perform ablation studies to verify the contributions of individual modules contained in the KSIQA framework, using one laboratory-based IQA dataset, TID2013, and one crowdsourcing-based dataset, PIPAL. More specifically, we examine the impact of three core components in the proposed architecture: the mental imagery generation, the KS strategy, and the QAR combined with deep ordinal regression.

TABLE VII

ABLATION STUDY TO VERIFY THE EFFECTIVENESS OF MIG MODULE WITH VERSUS WITHOUT L_2 SUPERVISION, USING THE TID2013 AND PIPAL DATASETS. "*" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT

Method	TID2013		PIPAL	
	PLCC	SROCC	PLCC	SROCC
MIG w/o L_2	0.757	0.723	0.624	0.611
MIG w/ L_2	**0.955***	**0.953***	**0.822***	**0.813***

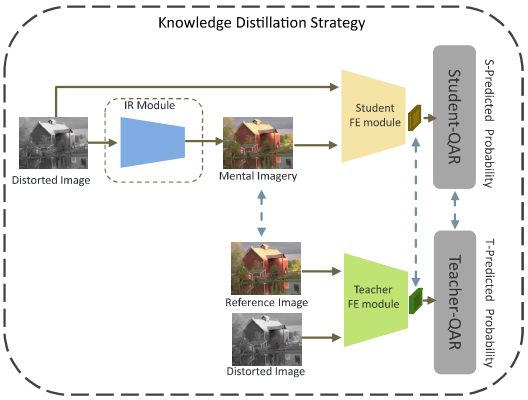


Fig. 8. Schematic of the simulation of a conventional KD (KS) strategy, which can be used to compare with and assess our new KS strategy.

TABLE VIII

ABLATION STUDY TO VERIFY THE EFFECTIVENESS OF THE PROPOSED KS STRATEGY VERSUS KD WITH JOINT TRAINING, USING THE TID2013 AND PIPAL DATASETS. "*" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT

Method	TID2013		PIPAL	
	PLCC	SROCC	PLCC	SROCC
KD-based IQA	0.755	0.721	0.531	0.524
KS-based IQA	**0.955***	**0.953***	**0.822***	**0.813***

1) Effectiveness of Generation of Mental Imagery: To avoid introducing additional image information, such as the HQ NAR images, during the testing phase, we have devised an effective MIG module to learn to generate mental imagery to simulate the visual priors. Without this, the NR-IQA student model cannot effectively discern the distribution differences between HQ and LQ images, hence cannot extract the transferred knowledge from the FR-IQA teacher model. Fig. 7 illustrates some representative examples of the results of the MIG module, showing the effectiveness of supervising the process of generating the mental imagery through learning.

The results shown in Table VII demonstrate the importance of supervising the generation of the mental imagery using the MSE loss. The MIG module for generating the mental imagery of reference without supervision gives poor model performance. It becomes evident that using a loss function to supervise the MIG significantly contributes to achieving good overall performance of the KSIQA model.

2) Effect of KS: The proposed KS technique is new and different from the traditional KD technique. To verify the contribution of KS, we simulate a comparable network based on KD as shown in Fig. 8. We incorporate MSE loss

TABLE IX

ABLATION STUDY TO VERIFY THE EFFECTIVENESS OF THE PROPOSED KS STRATEGY VERSUS THE CONVENTIONAL KD, USING THE TID2013 AND PIPAL DATASETS. "*" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT

Method	TID2013		PIPAL	
	PLCC	SROCC	PLCC	SROCC
KD-based IQA	0.936	0.921	0.785	0.756
KS-based IQA	**0.955***	**0.953***	**0.822***	**0.813***

TABLE X

ABLATION STUDY TO VERIFY THE EFFECTIVENESS OF DO-LOSS VERSUS CONVENTIONAL MSE LOSS, USING THE TID2013 AND PIPAL DATASETS. "*" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT

Method	TID2013		PIPAL	
	PLCC	SROCC	PLCC	SROCC
KSIQA w/ MSE	0.948	0.940	0.801	0.793
KSIQA w/ DO-loss	**0.955***	**0.953***	**0.822***	**0.813***

constraints on the features of the QAR and provide supervision on other feature levels. We compare the best-performing student model obtained from the KD strategy with the student model derived from our proposed KS strategy. The results are listed in Table VIII, indicating that the KS-based model demonstrates superior performance over the KD-based model.

Furthermore, to ensure a fair comparison, we conducted additional experiments where teacher-student models were trained jointly using the KS strategy, in addition to employing the separately trained KD strategy. A comparative experiment between KD (teacher model trained jointly from the student model) and KS is conducted. In this case, a teacher model is derived from a full-reference image quality assessment (FR-IQA) model with a student model being designed for no-reference image quality assessment (NR-IQA); both models are jointly trained. We further explore the challenges and potential issues arising from joint training, and the results are shown in Table IX.

The joint training process in NR-IQA poses the following challenges.

1) The significant divergence in reference information between the teacher and student models presents a substantial obstacle. The teacher model utilizes the authentic reference image, while the student model relies on a proxy reference image generated by a separate network.
2) The presence of the genuine reference image in the teacher model introduces a semantic gap between the two models. Implicit learning of features from the reference image by the teacher model may not be effectively transferred to the student model during KD.
3) Training with mismatched reference information results in suboptimal performance of the NR-IQA student model. The inability to distill knowledge effectively from the FR-IQA teacher model limits the student model's capacity to capture essential features related to IQA.

Therefore, when using joint training, employing the KS strategy rather than the KD in the NR-IQA task proves to be more effective.

The findings illustrate that the KS strategy not only reduces model parameters but also yields better results. This could be attributed to the fact that in the case of shared QAR, the student model can better emulate the generation of highly relevant features from the teacher model, mitigating the need for excessive loss functions, hence avoiding the occurrence of cumulative errors. However, using a KD strategy with multiple loss functions to constrain the QAR's feature map generation for the student network cannot guarantee the exact alignment in features between the student and teacher. This implies that the KS strategy can optimize the features within the shared QAR.

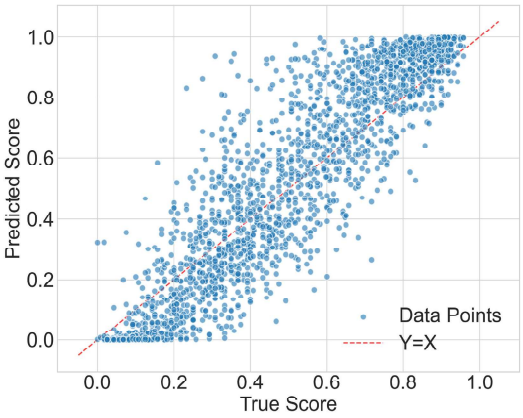


Fig. 9. Scatter plot of predicted scores versus ground-truth scores on the KADID-10k dataset.

3) Effectiveness of Ordinal Loss: We also investigate the impact of employing a QAR with a DO-loss function versus the conventional MSE loss function. The MSE loss overlooks the inherent ordering between different ratings on the perceptual image quality scale, which hinders the model's ability to accurately mimic human judgments of image quality. Based on the new concept of DO-loss, the proposed KSIQA model baserenders the predicted scores into probability distributions, and then utilizes a soft ordinal inference technique to transform these predicted probabilities into a continuous variable representing image quality. As illustrated in Table X, employing the DO-loss yields superior results compared to using the MSE loss for IQA.

D. Distortionwise Performance Analysis

Fig. 9 illustrates the scatter plot of predicted versus ground-truth scores on the KADID-10k dataset, which visually demonstrates their strong correlation.

To evaluate KSIQA's performance across different distortion types, we compute both PLCC and SROCC for all 25 individual distortion categories in the KADID-10k dataset. The detailed results are shown in Table XI. As shown, our proposed KSIQA model demonstrates robust and superior performance across a broad spectrum of distortion types, highlighting its generalization capability. However, we also note a performance drop on several specific distortion types, including Type 7 (color saturation 1), Type 18 (mean shift),

TABLE XI
DISTORTIONWISE PERFORMANCE ANALYSIS OF KSIQA ON KADID-10K DATASET, INCLUDING 25 INDIVIDUAL DISTORTION TYPES

Type	PLCC	SROCC	Type	PLCC	SROCC
1	0.975	0.949	14	0.924	0.925
2	0.953	0.949	15	0.914	0.934
3	0.939	0.940	16	0.899	0.863
4	0.932	0.819	17	0.925	0.753
5	0.926	0.896	18	0.516	0.436
6	0.883	0.873	19	0.951	0.906
7	0.426	0.412	20	0.369	0.449
8	0.924	0.907	21	0.851	0.798
9	0.960	0.923	22	0.850	0.869
10	0.962	0.880	23	0.500	0.548
11	0.849	0.850	24	0.908	0.915
12	0.907	0.896	25	0.591	0.599
13	0.877	0.866			

TABLE XII
PERFORMANCE COMPARISON WITH AUTHENTIC-SPECIFIC SOTA NR-IQA MODELS ON CLIVE AND KONIQ10K DATASETS

Method	CLIVE		KonIQ10k	
	PLCC	SROCC	PLCC	SROCC
Re-IQA [14]	0.854	0.840	0.923	0.914
QPT [68]	0.914	0.895	0.941	0.927
QCN [69]	0.893	0.875	0.945	0.934
KSIQA (Ours)	0.879	0.851	0.926	0.920

Type 20 (noneccentricity patch), and Type 23 (color block), where the correlation coefficients are comparatively lower. This suggests that while the model performs strongly overall, certain complex distortions remain challenging and represent targeted areas for future improvement.

E. Discussion

The novel design of our KS framework, featuring a shared decoder and joint training, merits further discussion concerning potential information leakage and model independence. Both the teacher and student are jointly trained from scratch on the same training set, and the final NR-IQA student model is evaluated on a held-out, unseen test set, ensuring no cross-contamination. The shared decoder acts as a controlled interface for structured knowledge exchange via backpropagation. Rather than causing leakage, it enables deeper semantic alignment between models during training. This approach facilitates collaborative optimization while preserving encoder-level independence. While the shared decoder introduces a tighter coupling between the student and teacher representations, potentially constraining the student's architectural independence, this tradeoff is advantageous. It grounds both models in a shared semantic space, allowing the student to better internalize reference-aware quality reasoning without relying on reference images during inference. Crucially, the student retains a fully independent encoder and is deployed as a standalone NR-IQA model once training is complete.

F. Applicability in Challenging Real-World Scenarios

As discussed in Section IV-B, we have already demonstrated that KSIQA can operate in scenarios where the reference images of the test dataset are not seen/used during training. Once the entire KSIQA model is trained, only the student model is directly used in the testing phases—representing a fully NR application. In this setup, both training and testing datasets follow the traditional IQA paradigm, where pristine reference images are available, and distortions are synthetic. However, in more challenging real-world cases, such as authentically distorted images from datasets, like LIVE Challenge [70] and KonIQ-10k [71], reference images are not available. To adapt KSIQA to these scenarios, we employ a two-stage strategy: *1) pretraining:* train the full KSIQA model (including both full-reference teacher and no-reference student) on large-scale synthetic IQA datasets (e.g., KADID-10k and PIPAL in our experiment) that contain reference images; and *2) fine-tuning:* fine-tune only the NR student model of KSIQA on the authentic IQA datasets that lack reference images. The strategy allows the KSIQA framework to first learn perceptual quality representations from existing synthetic IQA datasets with References and transfer that learned knowledge to authentic, NR settings. Once the KSIQA's student model has internalized perceptual cues from the teacher model, it can independently learn perceptual consistency during fine-tuning on large-scale, authentic, NR IQA datasets. For a fair comparison with SOTA NR-IQA models specifically designed for authentically distorted images, we follow the same fine-tuning protocol implemented in [14], [68], and [69]: "randomly splitting each dataset into training and testing sets with a ratio of 8:2, repeating the process for ten different splits, and reporting the median PLCC and SROCC scores." The results in Table XII show that KSIQA achieves competitive performance compared to SOTA specialized NR-IQA, demonstrating its practical effectiveness and applicability in real-world scenarios. Nevertheless, the current limitation of KSIQA lies in its MIG module, which is simplified for synthetic distortions. A promising direction for future work is to enhance the MIG module to reconstruct more realistic mental imagery from authentically distorted images, for example, by integrating SOTA generative models such as diffusion models.

V. CONCLUSION

In this article, we have presented an innovative framework, namely KSIQA, toward NR-IQA. We have pioneered the design of a deep learning architecture using a KS concept, where an FR-IQA teacher network and an NR-IQA student network share the same regressor and are collaboratively trained. The KS strategy harnesses the power of transferring IQA knowledge from the teacher to the student, while the model parameters are significantly reduced and the learning mechanism is simplified to prevent cumulative error. Also, the proposed NR-IQA student model contains an MIG module to learn to generate mental imagery as a simulated reference. This allows the trained NR-IQA model to take the single distorted image as the input without using any priors of a reference. Finally, our KSIQA framework is equipped with a deep ordinal regression method to better infer a perceptual quality assessment. A comprehensive comparative experiment has been conducted, and results demonstrate the superior performance and robust generalization of the proposed KSIQA

against SOTA IQA methods. Based on this, we conclude that KSIQA is a new class-leading technique for NR-IQA.

REFERENCES

[1] W. Liu, R. Cui, Y. Li, and S. Zhang, "Hybrid-input convolutional neural network-based underwater image quality assessment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1790–1798, Jan. 2025.

[2] Y. Chen, Y. Zhao, L. Cao, W. Jia, and X. Liu, "Learning deep blind quality assessment for cartoon images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6650–6655, Sep. 2023.

[3] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Learning to assess image quality like an observer," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8324–8336, Nov. 2023.

[4] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.

[5] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.

[6] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 1–52, Nov. 2020.

[7] W. Wen et al., "Perceptual quality assessment of virtual reality videos in the wild," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8368–8381, Sep. 2024.

[8] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[9] S. Seo, S. Ki, and M. Kim, "A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2602–2616, Jul. 2021.

[10] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.

[11] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.

[12] H. Wang, Y. Tu, X. Liu, H. Tan, and H. Liu, "Deep ordinal regression framework for no-reference image quality assessment," *IEEE Signal Process. Lett.*, vol. 30, pp. 428–432, 2023.

[13] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1191–1200.

[14] A. Saha, S. Mishra, and A. C. Bovik, "Re-IQA: Unsupervised learning for image quality assessment in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5846–5855.

[15] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14071–14081.

[16] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 3–18.

[17] G. Yin et al., "Content-variant reference image quality assessment via knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 3134–3142.

[18] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[19] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[20] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2013–2026, Dec. 2013.

[21] S. Lao, "Attentions help CNNs see better: Attention-based hybrid image quality assessment network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 1140–1149.

[22] H. Guo, Y. Bin, Y. Hou, Q. Zhang, and H. Luo, "IQMA network: Image quality multi-scale assessment network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 443–452.

[23] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[24] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[25] P. Ye, J. Kumar, and D. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4241–4248.

[26] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2394–2402.

[27] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, "VDTR: Video deblurring with transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 160–171, Jan. 2023.

[28] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 732–741.

[29] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14143–14152.

[30] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[31] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image Video Process.*, vol. 12, no. 2, pp. 355–362, Feb. 2018.

[32] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Oct. 2022, pp. 1220–1230.

[33] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3575–3585.

[34] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3667–3676.

[35] H. Wu et al., "Q-align: Teaching LMMs for visual scoring via discrete text-defined levels," in *Proc. 41st Int. Conf. Mach. Learn.*, 2023, pp. 54015–54029.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[37] N. C. Garcia and P. M. V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 103–118.

[38] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2827–2836.

[39] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[40] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11933–11942.

[41] E. Ben-Baruch, M. Karklinsky, Y. Biton, A. Ben-Cohen, H. Lawen, and N. Zamir, "It's all in the head: Representation knowledge distillation through classifier sharing," 2022, *arXiv:2201.06945*.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

[45] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[46] J. Dai et al., "Deformable convolutional networks," in *Proc. ICCV*, 2017, vol. 1, no. 3, p. 4.

[47] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[48] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[49] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[50] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[51] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 633–651.

[52] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[53] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[54] J. You and J. Korhonen, "Transformer for image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1389–1393.

[55] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[56] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.

[57] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[58] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1473–1476.

[59] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[60] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 433–442.

[61] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Proc. 45th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 723–727.

[62] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2019, pp. 334–355.

[63] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.

[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[65] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[67] S. Shi et al., "Region-adaptive deformable network for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 324–333.

[68] K. Zhao, K. Yuan, M. Sun, M. Li, and X. Wen, "Quality-aware pretrained models for blind image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22302–22313.

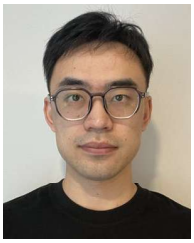
[69] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Blind image quality assessment based on geometric order learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 12799–12808.

[70] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[71] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

Huasheng Wang received the M.S. degree from Dalian University of Technology, Dalian, China, in 2021, and the Ph.D. degree from Cardiff University, Cardiff, U.K., in 2024.

He is currently an Algorithm Engineer with Taobao, Alibaba Group, Hangzhou, China. His research interests include image and video quality assessment and saliency prediction.

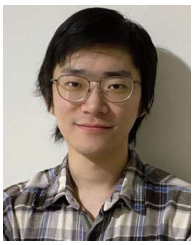


Jiang Liu (Graduate Student Member, IEEE) received the B.Eng. and M.S. degrees from China University of Mining and Technology, Xuzhou, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K.

His interests include action quality assessment, image quality assessment, and saliency prediction.

Hongchen Tan received the Ph.D. degree in computational mathematics from Dalian University of Technology, Dalian, China, in 2021.

He is a Teacher with the College of Future Technology, Dalian University of Technology. His research interests include computer vision.



Jianxun Lou (Member, IEEE) received the B.Eng. degree from Central South University, Changsha, China, in 2018, and the M.S. and Ph.D. degrees from the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K., in 2020 and 2024, respectively.

He is currently a Lecturer at Northeast Electric Power University, Jilin, China. His research interests include visual perception modeling and visual quality assessment.

Xiaochang Liu is currently pursuing the bachelor's degree with the School of Mathematics, Sun Yat-sen University, China.

Her research interests include mathematical modeling and data analytics.

Wei Zhou received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2021. He was with the University of Waterloo, Waterloo, ON, Canada, from 2019 to 2021. He was a Post-Doctoral Fellow at the University of Waterloo. He was a Visiting Scholar at the National Institute of Informatics, Tokyo, Japan, He is an Assistant Professor at Cardiff University, Cardiff, U.K. His research interests span multimedia computing, perceptual image processing, and computational vision.

Dr. Zhou is currently an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Ying Chen (Senior Member, IEEE) received the B.S. degree in applied mathematics and the M.S. degree in electrical engineering and computer science, Peking University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computing and electrical engineering from Tampere University of Technology (TUT), Tampere, Finland, in 2010.

He is currently leading Audiovisual Technology Group, Taobao, Alibaba Group, Hangzhou, China, supporting end-to-end multimedia features and applications within Taobao. His research areas include video coding, image/video restoration and enhancement, image/video quality assessment, and video transmission.

Roger Whitaker is currently a Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research interests include the intersection of machine and human intelligence, including human behavior.

Dr. Whitaker is an Area Editor of *Online Social Networks and Media* (Elsevier) and an Associate Editor of *Social Network Analysis and Mining* (Springer).

Walter Colombo (Member, IEEE) is a Lecturer at Cardiff University, where he previously worked as a Researcher and a Research Software Engineer. His research primarily concerns social computing, intelligence and evolution, and agent-based modeling by exploiting parallel processing at scale. From his Ph.D. years in evolutionary optimization, he has built up extensive expertise for cross-disciplinary research projects spanning computer science, social sciences, and psychology. He supported model development and complex code for a range of projects, including agent-based modeling and artificial intelligence deployed on high-performance computing resources. His recent research interests include interpreting psychological and evolutionary concepts alongside the relation between AI, innovation, and creativity.

Hantao Liu (Member, IEEE) received the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 2011.

He is currently a Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K.