Initialize and Train a Unified Quality Assessment Model for Images/Videos in the Wild

Dingquan Li, Haiqiang Wang, Wei Gao, and Ge Li

***Abstract*—Image/video quality assessment (I/VQA) is critical for various image and video processing tasks, where it plays the role of the 'referee' and the 'coach,' *i.e.*, benchmarking and optimizing the image/video processing systems. For in the wild scenario, with no access to pristine references, blind I/VQA models are more practical but challenging than the opposite ones. Deep neural networks have revolutionized the field of blind I/VQA. However, the performance of trained blind I/VQA networks is highly related to network initialization techniques, which get little attention. In this paper, we present a unified blind I/VQA network by treating images as single-frame videos and systematically study the initialization strategies for this blind I/VQA network. The unified blind I/VQA network consists of two modules: feature extractor and quality regressor. The feature extractor includes the convolutional layers of an image classification network followed by a global spatial average pooling layer. The quality regressor contains a recurrent neural network followed by a global temporal pooling layer. For the IQA task, the feature extractor can be initialized randomly or from an ImageNet-pretrained image classification model, and the quality regressor is generally initialized randomly. For the VQA task, the feature extractor can be initialized randomly, from an ImageNet-pretrained image classification model, or from a trained IQA model. Meanhwile, the quality regressor can be initialized randomly or from the trained IQA model. Experimental results show that basic ResNet architectures, without extra design to accommodate I/VQA tasks, still achieve state-of-the-art performance if the unified models are properly initialized. A better practice to initialize the I/VQA network is to train the network on more relevant datasets, *i.e.*, the initialization strategies, from the best to the worst, are finetuned IQA model initialized, ImageNet-pretrained model initialized, and randomly initialized. We provide the PyTorch implementation at https://github.com/lidq92/IQA4VQA as the supporting information for reproducible scientific research.**

***Index Terms*—Feature extraction, image and video quality assessment, network initialization, quality regression**

I. Introduction

Images and videos currently stand out as the most significant data traffic on the Internet. With the popularity of portable mobile devices, non-professional users can easily capture images and videos to record their life and share them on the Internet. These contents may contain undesired distortions like blurriness caused by camera shake or underexposure artifacts in low-light environments. Image and video quality assessment (I/VQA) can help to detect low-quality videos with various degradation. Furthermore, I/VQA models could be used to benchmark and optimize video processing pipelines with quality restoration techniques included. Given the importance of I/VQA in the field of image and video processing, I/VQA has attracted more and more researchers in recent years [1], [2], [3], [4], [5], [6], [7], [8]. In most cases, the pristine reference is unavailable. Thus blind I/VQA models are required.

In recent years, deep learning has greatly advanced the state of the art in image processing and computer vision, resulting in significant improvements in accuracy and efficiency, such as image denoising, image matting, image classification, object tracking, image segmentation, age estimation, and defect detection. This technology has also been applied in the field of blind I/VQA in the wild, where deep neural networks have shown considerable performance improvements [6], [9], [10], [11]. Existing researches mainly focus on the dataset construction [12], [13], architecture design [14], learning strategies [3], loss function design [15], *etc*. Network initialization gets little attention in training I/VQA networks, although it is an important aspect for model training [16]. Existing I/VQA networks could be initialized randomly or from models trained on other datasets. For example, all the parameters of the IQA network are randomly initialized in [17], and the feature extractor of the IQA network in [18] is initialized from ImageNet-pretrained models. In [19], one branch of the IQA network is initialized randomly and pretrained on a classification task using quality degradation types and degrees, while the other branch is initialized by ImageNet-pretrained weights. For the VQA networks, besides random initialization [20], the ImageNet-pretrained models are used for initializing the feature extractor in [14], [21]. In addition, quality-aware pre-training is performed for initializing the spatial feature extractor in [22] and contrastive self-supervised pre-training is conducted on a large-scale unlabeled video dataset for the same purpose [23].

Besides, there is no unified model specifically-designed for blind I/VQA. Although video quality can be naturally predicted by applying an IQA model on a frame basis, followed by a temporal pooling operation, the performance of such solution on the VQA task is generally poor due to the lack of temporal modelling. And most of current VQA models are not suitable for IQA since they exploit temporal or spatial-temporal information with a hard requirement of at least two video frames. Instead, in this paper, by treating the image as a single-frame video, we present a unified blind I/VQA network with recurrent neural network for temporal modelling, which is applicable to both the IQA task and the VQA task. However,

D. Li is with Network Intelligence Research Department, Peng Cheng Laboratory, Shenzhen, China (e-mail: dingquanli@pku.edu.cn).

H. Wang is with Tencent Inc., Shenzhen, China (e-mail: walleewang@tencent.com).

W. Gao and G. Li are with SECE, Peking University Shenzhen Graduate School, Shenzhen, China (e-mail: gaowei262@pku.edu.cn; geli@ece.pku.edu.cn).

it is still an open question how such a unified network should be initialized to achieve competitive performance on both of IQA and VQA tasks. Therefore, we systematically study the initialization strategies of blind I/VQA networks, and conduct experiments towards recommendations on better practice.

Specifically, to explore the feasibility to unify blind IQA and VQA in a single network, we simply take the well-designed image classification networks [24] for feature extraction, followed by a shallow gated recurrent unit network [25] for quality regression. The deep features from the ImageNet-pretrained model are sensitive to distortion and contain semantic information [26], making them useful for quality estimation. Even for the I/VQA problem under study, we can still benefit from the ImageNet-pretrained image classification model. Thus, we can initialize the feature extractor of the IQA network by ImageNet-pretrained weights instead of random initialization. As for the quality regressor, it is often initialized randomly as there is no prior. After the IQA network training process is finished, we obtain a model well-trained on the IQA dataset. Therefore, for the VQA network, which has the same architecture as the IQA network, its feature extractor can be initialized randomly, by ImageNet-pretrained weights, or by the trained IQA network weights. Meanwhile, the quality regressor of the VQA network can be initialized from the trained IQA network instead of random initialization.

We conduct comprehensive experiments on four I/VQA datasets and draw a conclusion that, regardless of network architecture, a better way to initialize the I/VQA networks is to pretrain the networks on other relevant datasets. That is, random initialization is the worst choice, and a well-trained IQA network is more helpful for initializing the VQA network than the ImageNet-pretrained one. The experimental results also show that the proposed unified blind I/VQA network achieves comparable performance to the state-of-the-art I/VQA models.

To summarize, the contributions of this work lie on the following three aspects:

- This work presents a unified model for I/VQA in the wild by treating the image as a single-frame video.
- This work systematically studies the initialization strategies of the unified blind I/VQA models, and the experimental analysis leads to a recipe of initialization.
- The simple implementation of the unified blind I/VQA model achieves state-of-the-art performance on four in-the-wild I/VQA datasets.

II. Related Work

This section briefly reviews representative and recent advanced researches on blind IQA (BIQA) and blind VQA (BVQA) in two separate subsections. For more details and progresses in the field of I/VQA, please refer to some recent surveys, *e.g.*, [27], [28], [29], [30], [31].

A. Blind Image Quality Assessment

The research on general (not distortion-specific) BIQA has been in a bottleneck until the 2010s when it encountered two crucial turning points. The first turning point was the creation of BIQA models based on natural scene statistics (NSS) or dictionary learning in 2012 [32], [33]. The second turning point was the application of deep learning technology to BIQA in 2014 [17].

NSS models assume that original undistorted natural images have strong statistical regularity, but the existence of distortion will break this regularity. Therefore, the quality of an image can be predicted by measuring the displacement of statistics caused by distortion. BRISQUE [32] first calculates the mean subtracted contrast normalized (MSCN) coefficients of the image in the spatial domain, whose distributions are close to the generalized Gaussian distribution (GGD). The existence of distortion causes the MSCN coefficient distribution to shift and the statistical characteristics of the coefficient distribution to change. Besides, the distribution of the product of the adjacent MSCN coefficients can be modeled by the asymmetric GGD (AGGD), and distortion also changes the statistical characteristics of these distributions. Therefore, the BRISQUE model uses GGD to fit the MSCN coefficient and AGGD to fit the product of the adjacent MSCN coefficients in four directions, respectively. According to the parameters fitted by GGD and AGGD, 36 NSS features are obtained, which are regressed to the subjective quality scores. NIQE [34] uses the same 36 NSS features as BRISUQE. However, the NIQE model does not need to use the support vector regressor (SVR) for quality prediction. Instead, it uses a multivariate Gaussian model (MVG) to fit these NSS features. Then, the quality of an image is indicated by the distance of MVG with regard to a high-quality image set. ILNIQE [35] extends the NIQE with additional statistical characteristics of gradient, color, *etc*. In addition to considering NSS in the spatial domain, there are other BIQA models based on NSS in transform domains, such as the wavelet domain [36] and the discrete cosine transform domain [37].

BIQA models based on dictionary learning extract quality-related features from images by constructing a dictionary/codebook. CORNIA [33] first uses an unsupervised learning approach to construct an unlabelled codebook through K-means clustering, then obtains quality-related features of the test image blocks based on the codebook through soft-threshold assignment and max-pooling, and finally predicts image quality through linear SVR. HOSA [38] exploits high-order statistics aggregation, requiring only a small codebook. QAC [39] is another representative model based on dictionary learning. It first performs quality-related clustering, that is, clustering image blocks according to the scores given by a full-reference IQA model into several quality levels. Then, it uses the centroids of the clustering as the codebook to infer the quality scores of all image blocks and finally converts all the block scores into the quality scores of the image through weighted average pooling.

In 2013, a supervised filter learning model for joint feature extraction and quality prediction was proposed, which used a backpropagation-based algorithm for model optimization [40]. This model bridged traditional machine learning and modern deep learning techniques. The same research team subsequently replaced the model's filters with convolutional neural networks (CNNs) to improve it, producing the first

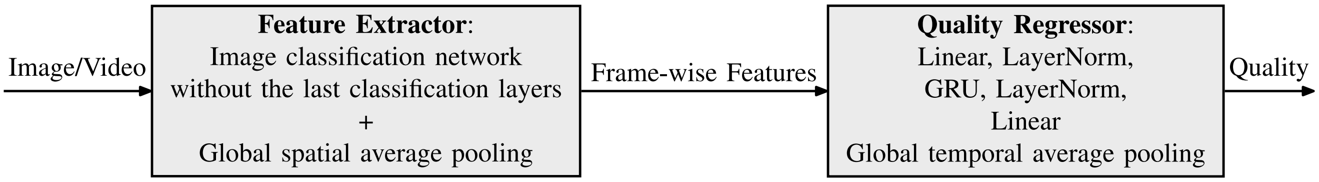

Image/Video
Feature Extractor:
Image classification network
without the last classification layers
+
Global spatial average pooling
Frame-wise Features
Quality Regressor:
Linear, LayerNorm,
GRU, LayerNorm,
Linear
Global temporal average pooling
Quality

Fig. 1. Flowchart of the proposed unified blind I/VQA framework, where an image is treated as a single-frame video.

deep learning-based BIQA model, *i.e.*, CNNIQA [17]. After that, network-based learnable computational structures have revolutionized the field of BIQA [41], [42], [13], [3]. Along this line, the loss function plays a vital role in training BIQA models. A natural way is to use the Minkowski metric by formulating BIQA as a regression task. BIQA can also be viewed as a problem of learning-to-rank [43] to infer relative quality. The cross-entropy loss and the fidelity loss can be used for pairwise learning-to-rank, while the cross-entropy loss over permutation probabilities [44] and the norm-in-norm loss [15] have been successfully adopted by listwise learning-to-rank. Recently, researchers have turned to focus on unified optimization strategies for allowing a single BIQA model to learn from multiple IQA datasets simultaneously [3], continual learning of BIQA models for overcoming subpopulation shift [45], [46], [47], and adversarial attacks of BIQA models [48], [49].

B. Blind Video Quality Assessment

For a long time, BVQA researches mainly focused on developing distortion-specific models based on distortion analysis, such as compression distortion, transmission error, shakiness, spatially correlated noise, and video scaling. However, there are few studies on general NR-VQA models. Universal NR-VQA models need to work on quality evaluation of videos with various types of distortion, which makes it challenging to analyze specific distortion in model design. A straightforward way to blindly estimate video quality is by separately predicting frame quality scores with a BIQA model and then temporally pooling the scores. A comparative evaluation of different temporal pooling strategies is shown in [50]. However, the performance is generally unsatisfactory due to no or limited temporal modelling. Several previous studies have attempted to address this issue. For example, V-BLIINDS [51] is a BIQA model that exploits spatio-temporal NSS feature of frame differences and motion coherency characteristics. VIIDEO [52] is a completely-blind VQA model that explores intrinsic statistical regularities of natural videos to evaluate video quality by quantifying the distortions-induced statistics irregularities. TLVQM [53] extracts two levels of handcrafted features, including high-complexity spatial features computed on sparse frames for quantifying spatial distortions and low-complexity temporal features calculated per frame for measuring temporal variations. CNN-TLVQM [54] further enhances TLVQM with a trained IQA network for spatial feature extraction. VIDEVAL [55] selects suitable handcrafted features from previous I/VQA models to handle diverse and authentic distortions, which serve as the inputs of an SVR to output video quality. RAPIQUE [56] combines the handcrafted NSS features and deep CNN features for BVQA.

Benefiting from distortion-sensitive and content-aware features of image classification networks, deep BVQA models emerged and have achieved excellent performance. VSFA [21] extracts spatial features from ImageNet-pretrained ResNet50 and then exploits a gated recurrent unit (GRU) network to model the temporal relationship between the spatial features of video frames. RIRNet [57] utilizes recurrent-in-recurrent networks to prompt accurate motion perception for BVQA by incorporating the speed-tuning property of neurons. In [22], quality-aware pre-training is performed to extract compelling spatial features. In [10], an end-to-end spatial feature extractor is trained to extract quality-aware spatial features from sparse frames, combined with pre-extracted temporal features from dense frames to infer video quality. Recently, an efficient end-to-end BVQA model, named FAST-VQA is proposed with the concept of fragment sampling [6].

III. The Unified Blind I/VQA Network

In this section, we present a unified blind I/VQA network and discuss different initialization strategies for its optimization. I/VQA generally includes two stages: feature extraction and quality regression. Correspondingly, our blind I/VQA network consists of the feature extractor and quality regressor (See Fig. 1).

The key for our unified solution for both BIQA and BVQA is treating the image as a single-frame video, and we must design a network that can accept the single-frame video as input, *i.e.*, no frame difference, motion, or optical flow from two adjacent frames is explored. Although this kind of network design may weaken the temporal modelling of the BVQA task, we can resort to gated recurrent unit (GRU) [25] for modelling the relationship between frames. For simplification, here we adopt well-designed image classification networks [24], [58] for feature extraction. Specifically, we drop the last classification layers of the image classification network and add a global spatial average pooling layer to aggregate information from the activation maps. As such, we obtain frame-wise features that serve as the inputs of the quality regressor. Thanks to the capability of the feature extractor, one can design a pretty simple quality regressor. In our case, we use a Linear layer for dimension reduction, a two-layers GRU for feature propagation along the temporal axis. This is followed by another Linear layer for frame-wise quality prediction and a global temporal average pooling layer for the final quality prediction of the input image/video. In addition, layer normalization (LayerNorm) [59] is used to speed up the training process.

TABLE I

INSTANCES OF UNIFIED BLIND I/VQA NETWORKS WITH BASIC RESNET ARCHITECTURES. INSIDE THE BRACKETS ARE THE SHAPE OF A RESIDUAL BLOCK, AND OUTSIDE THE BRACKETS IS THE NUMBER OF STACKED BLOCKS ON A STAGE. "C=32" SUGGESTS GROUPED CONVOLUTIONS WITH 32 GROUPS. DOWNSAMPLING IS PERFORMED BY CONV3_1, CONV4_1, AND CONV5_1 WITH A STRIDE OF 2. THE NONLINEARITY AND NORMALIZATION ARE OMITTED FOR BREVITY

stage	layer name	ResNet18 [24]	ResNet34 [24]	ResNet50 [24]	ResNeXt101 (32x8d) [58]
feature extractor	conv1	7×7, 64, stride 2			
		3×3 max pool, stride 2			
	conv2_x	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, C{=}32 \\ 1\times1, 256 \end{bmatrix}\times3$
	conv3_x	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, C{=}32 \\ 1\times1, 512 \end{bmatrix}\times4$
	conv4_x	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 1024 \\ 3\times3, 1024, C{=}32 \\ 1\times1, 1024 \end{bmatrix}\times23$
	conv5_x	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 2048 \\ 3\times3, 2048, C{=}32 \\ 1\times1, 2048 \end{bmatrix}\times3$
	gsap	global spatial average pooling (GSAP)			
quality regressor	fc1	256-d fully-connected layer			
	gru	two-layer GRU with hidden size 64			
	fc2	1-d fully-connected layer			
	gtap	global temporal average pooling (GTAP)			

A. Network Architecture

We can represent a colored image or video as a tensor $\mathbf{x} \in \mathbb{R}^{T\times 3\times H_0\times W_0}$, where T is the number of frames, 3 represents the color channels (red, green, blue), and $H_0\times W_0$ is the spatial resolution of the image or video. For the case of an image, which can be considered as a single-frame video, we have $T = 1$. We can then input this tensor into the convolutional layers of a well-designed image classification network f with parameters $\mathbf{w}$ to produce output, *i.e.*,

$$\mathbf{x}_{\text{maps}} = f(\mathbf{x}; \mathbf{w}). \tag{1}$$

The tensor $\mathbf{x}_{\text{maps}}$ contains C output activation maps with dimensions $H \times W$ for a total of T frames. After applying a global spatial average pooling (GSAP) layer, these maps are transformed into frame-wise features $\mathbf{x}_{\text{feat}} \in \mathbb{R}^{T\times C}$, *i.e.*,

$$\mathbf{x}_{\text{feat}} = \text{GSAP}(\mathbf{x}_{\text{maps}}). \tag{2}$$

Then, we proceed to estimate the frame-wise quality scores from these frame-wise features. We resort to GRU for temporal modelling (We also apply GRU to the image or single-frame video for a unified model, although there is no temporal modelling in such situation). Since high dimensional features would introduce difficulty in GRU training, we consider a linear layer before GRU for dimension reduction. After temporal modelling with GRU, we simply use a linear layer to map the hidden features to frame-wise quality scores. LayerNorm is considered for accelerating network optimization. Overall, we adopt the sequential structure g ([Linear, LayerNorm, GRU, LayerNorm, Linear]) with parameters $\mathbf{v}$ for frame-wise quality prediction, *i.e.*,

$$\mathbf{q} = g(\mathbf{x}_{\text{feat}}; \mathbf{v}). \tag{3}$$

For simplicity, the quality score Q of the input image/video can be obtained by global temporal average pooling (GTAP) of the frame-wise quality scores $\mathbf{q} = [q_1, \cdots, q_T] \in \mathbb{R}^{T\times 1}$, *i.e.*,

$$Q = \text{GTAP}(\mathbf{q}) = (q_1 + \cdots + q_T)/T. \tag{4}$$

To have a clear look, Table I presents several instances of the unified I/VQA network with basic ResNet architectures.

B. Network Initialization

Network initialization is the first step of the training procedure. It is an essential aspect of training deep neural networks. We intend to systematically study different initialization strategies for training unified blind I/VQA networks. If no trained weights are available, the default random initialization algorithms in the popular deep learning framework (*e.g.*, PyTorch [60]) are good choices. However, trained weights should be a better starting point than the random initialized ones. With this consideration, we compare different initialization strategies of the feature extractor and the quality regressor.

For the blind IQA problem, our feature extractor is borrowed from the well-designed image classification network. The parameters of the network could be randomly initialized $\mathbf{w} = \mathbf{w}_{\text{random}}$, or initialized from the ImageNet-pretrained weights $\mathbf{w} = \mathbf{w}_{\text{ImageNet}}$. As for the quality regressor, we consider the default random initialization $\mathbf{v} = \mathbf{v}_{\text{random}}$.

For the blind VQA problem, the feature extractor can be initialized randomly $\mathbf{w} = \mathbf{w}_{\text{random}}$, from the ImageNet-pretrained weights $\mathbf{w} = \mathbf{w}_{\text{ImageNet}}$, or from the above trained

IQA weights $\mathbf{w} = \mathbf{w}_{\text{IQA}}$. Besides random initialization $\mathbf{v} = \mathbf{v}_{\text{random}}$, the quality regressor can also be initialized by trained IQA weights $\mathbf{v} = \mathbf{v}_{\text{IQA}}$.

C. Loss Function

For network training, we adopt the norm-in-norm loss [15] due to its effectiveness on fast convergence. Specifically, for a batch of B images/videos, we denote the predicted quality scores and the ground truth scores as $\{Q_i^{\text{pred}}\}_{i=1}^{B}$ and $\{Q_i^{\text{gt}}\}_{i=1}^{B}$, respectively. Then, the loss in the batch is computed as follows.

$$\ell = \frac{1}{2B}\sum_{i=1}^{B}\left|\frac{Q_i^{\text{pred}} - \mu^{\text{pred}}}{\sigma^{\text{pred}}} - \frac{Q_i^{\text{gt}} - \mu^{\text{gt}}}{\sigma^{\text{gt}}}\right|, \tag{5}$$

where $\mu^{\text{pred}}/\mu^{\text{gt}}$ and $\sigma^{\text{pred}}/\sigma^{\text{gt}}$ are the mean and standard deviation of objective/subjective quality scores, respectively. They are defined by the following equations.

$$\mu^{\text{pred}} = \frac{1}{B}\sum_{i=1}^{B} Q_i^{\text{pred}}, \tag{6}$$

$$\mu^{\text{gt}} = \frac{1}{B}\sum_{i=1}^{B} Q_i^{\text{gt}}, \tag{7}$$

$$\sigma^{\text{pred}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(Q_i^{\text{pred}} - \mu^{\text{pred}}\right)^2}, \tag{8}$$

$$\sigma^{\text{gt}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(Q_i^{\text{gt}} - \mu^{\text{gt}}\right)^2}. \tag{9}$$

IV. Experiments and Results

In this section, we conduct extensive experiments to study the performance of the proposed unified architecture and the impact of the aforementioned network initialization strategies in the training process, as well as performance comparison with existing methods on blind I/VQA tasks.

A. Experimental Setting

We perform experiments on four representative datasets: KonIQ-10k [13] and CLIVE [61] for IQA, as well as KoNViD-1k [12] and LIVE-VQC [62] for VQA, respectively.

For training and evaluating IQA models, we follow the experimental setting of KonCept512 [13]. Specifically, 10,073 images from KonIQ-10k are divided into 7,058 training images, 1,000 validating images, and 2,015 testing images, respectively. We train the IQA model on the 7,058 training images and save the best model in terms of Spearman's Rank-Order Correlation Coefficient (SROCC) on the 1,000 validation images. Then, we report intra-dataset prediction performance using the SROCC on the 2,015 testing images. The cross-dataset evaluation is performed on CLIVE, which contains 1,162 images. All the images are resized to 664×498 before being fed into the network.

For training and evaluating VQA models, we use the same experimental setting of VSFA [21]. To be specific, 1,200 videos from KoNViD-1k are split into 720, 240, and 240 videos for training, validating, and testing, respectively. We ensure no overlap among training, validation, and testing sets to avoid data leakage. We conduct the splitting procedure ten times and report the average SROCC. We use the IQA model weights trained on KonIQ-10k as $\mathbf{w}_{\text{IQA}}$. The model trained on the first split is used for cross-dataset evaluation on LIVE-VQC (containing 585 videos). For the reduction of time complexity, a video is equally divided into 16 groups along the temporal axis, and only the first frame of each group is sampled.

We adopt image classification networks as I/VQA feature extractors. We evaluate not only ResNet-like architectures including ResNeXt101 [58], ResNet50, ResNet34, and ResNet18 [24], but also non-ResNet-like architectures such as GoogleNet [63], VGG16 [64], and AlexNet [65]. Quality regressor first reduces the frame-wise features to a dimension of 256, and the hidden size is set to 64 in the two-layer GRU.

For network optimization, we train the model using Adam optimizer [66] with a batch size of 8 for a total of 30 epochs. We use an initial learning rate 10^{-4} and apply a step learning rate schedule. Besides, we use a tenth of the initial learning rate, *i.e.*, 10^{-5}, for parameters initialized from trained weights. To support reproducible scientific research, we will provide our PyTorch implementation as the supporting information for reproducible scientific research.

B. Results and Analysis

We analyze the results of the IQA and VQA tasks in two separate paragraphs.

IQA: Table II and Table III demonstrate the intra- and cross-dataset performance comparisons in terms of SROCC for IQA networks initialized with different strategies, respectively[1]. We have the following observations. First, initialization with ImageNet-pretrained weights is better than random initialization, with huge SROCC gains, especially when the parameters in the feature extractor ($\mathbf{w}$) are frozen. Second, freezing parameters would constrain the capability of the model to learn good feature representations. This is validated by comparing between the 2nd and the 5th columns, as well as the 3rd and the 6th columns. Third, a good initialization with weights pretrained on a large-scale dataset can compensate for the domain gap between different tasks. This is evidenced by the fact that frozen ImageNet-pretrained feature extractor (the 3rd column) outperforms randomly-initialized learnable feature extractor (the 5th column). Fourth, with an ImageNet-pretrained initialization (the 6th column), a ResNet-like network with more layers can improve the performance at the cost of increased computational complexity. But this is not the case for random initialization (the 5th column). We doubt that the network encounters difficulty in the training process when there are more layers randomly initialized. Finally, results on non-ResNet-like architectures lead to the same observations.

We show the SROCC curve on the test set with different initialization strategies for ResNeXt101 in Fig. 2a,

[1] All the reported values shown in the paper are rounded to three decimal places.

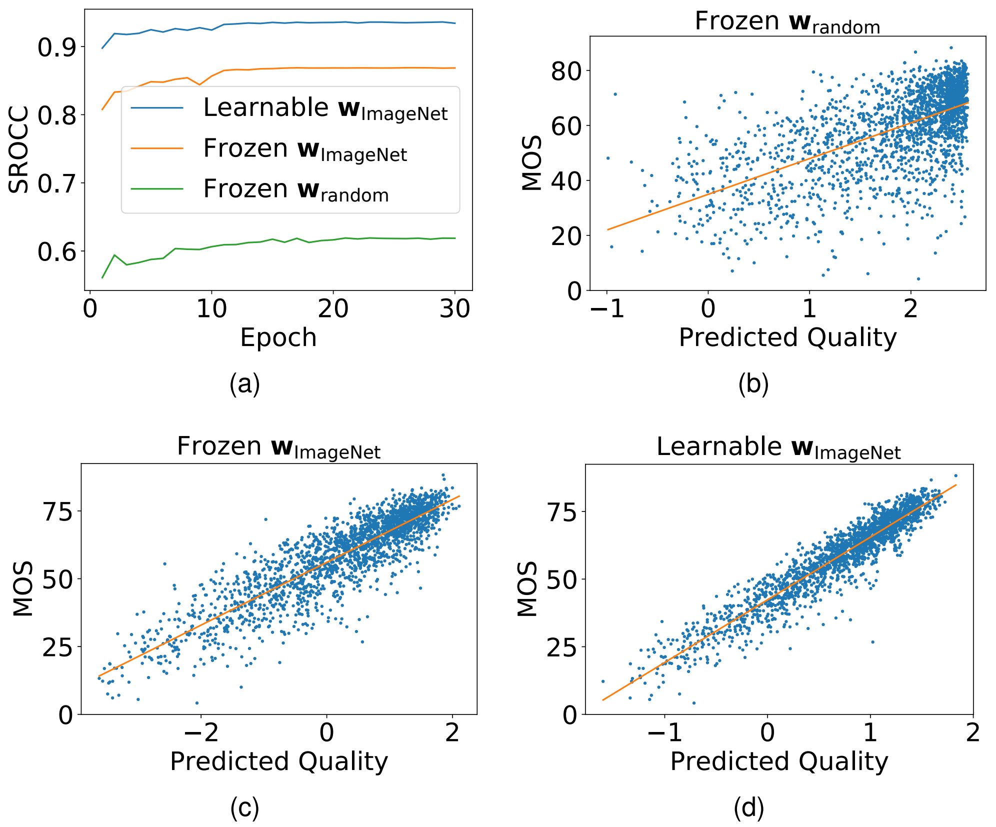
TABLE II
INTRA-DATASET PERFORMANCE COMPARISON ON KONIQ-10K FOR IQA NETWORKS INITIALIZED WITH DIFFERENT STRATEGIES

Architecture of $\mathbf{w}$	$\mathbf{w}$ is frozen, and initialized with		Gain	$\mathbf{w}$ is learnable, and initialized with		Gain
	$\mathbf{w}_{\text{random}}$	$\mathbf{w}_{\text{ImageNet}}$		$\mathbf{w}_{\text{random}}$	$\mathbf{w}_{\text{ImageNet}}$	
ResNeXt101	0.619	0.869	+0.250	0.786	0.935	+0.150
ResNet50	0.598	0.874	+0.276	0.783	0.920	+0.137
ResNet34	0.561	0.819	+0.257	0.776	0.908	+0.132
ResNet18	0.596	0.829	+0.233	0.786	0.905	+0.119
GoogleNet	−0.104[1]	0.815	+0.920	0.783	0.898	+0.115
VGG16	0.641	0.852	+0.210	0.788	0.914	+0.127
AlexNet	0.650	0.848	+0.198	0.743	0.885	+0.142

[1] GoogleNet with frozen random $\mathbf{w}$ leads to the worst result since the network depth is very deep, and there is no residual connection for helping optimization.

TABLE III
CROSS-DATASET PERFORMANCE COMPARISON ON CLIVE FOR IQA NETWORKS INITIALIZED WITH DIFFERENT STRATEGIES

Architecture of $\mathbf{w}$	$\mathbf{w}$ is frozen, and initialized with		Gain	$\mathbf{w}$ is learnable, and initialized with		Gain
	$\mathbf{w}_{\text{random}}$	$\mathbf{w}_{\text{ImageNet}}$		$\mathbf{w}_{\text{random}}$	$\mathbf{w}_{\text{ImageNet}}$	
ResNeXt101	0.496	0.751	+0.256	0.641	0.829	+0.188
ResNet50	0.437	0.761	+0.324	0.646	0.803	+0.157
ResNet34	0.417	0.711	+0.294	0.662	0.794	+0.132
ResNet18	0.452	0.689	+0.236	0.665	0.754	+0.089
GoogleNet	−0.072	0.660	+0.732	0.628	0.774	+0.146
VGG16	0.494	0.715	+0.221	0.647	0.784	+0.137
AlexNet	0.465	0.718	+0.253	0.639	0.765	+0.126



(a)

(b)

(c)

(d)

Fig. 2. Subfigure (a) shows the SROCC curve with regard to the training epoch on the KonIQ-10k test set. The other three subfigures (b-d) are scatter plots between the mean opinion score (MOS) and the predicted quality with different initialization strategies.

TABLE IV
INTRA-DATASET PERFORMANCE COMPARISON ON KONVID-1K FOR VQA NETWORKS INITIALIZED WITH DIFFERENT STRATEGIES

Frozen $\mathbf{w}$ / Learnable $\mathbf{v}$	$\mathbf{w}_{\text{random}}$	$\mathbf{w}_{\text{ImageNet}}$	$\mathbf{w}_{\text{IQA}}$	Gain ($\mathbf{w}_{\text{ImageNet}}$ vs $\mathbf{w}_{\text{random}}$)	Gain ($\mathbf{w}_{\text{IQA}}$ vs $\mathbf{w}_{\text{ImageNet}}$)
$\mathbf{v}_{\text{random}}$	0.390	0.758	0.821	+0.367	+0.063
$\mathbf{v}_{\text{IQA}}$	0.417	0.784	0.830	+0.367	+0.047
Gain ($\mathbf{v}_{\text{IQA}}$ vs $\mathbf{v}_{\text{random}}$)	+0.026	+0.026	+0.010	+0.394	+0.073

TABLE V
CROSS-DATASET PERFORMANCE COMPARISON ON LIVE-VQC FOR VQA NETWORKS INITIALIZED WITH DIFFERENT STRATEGIES

Frozen $\mathbf{w}$ / Learnable $\mathbf{v}$	$\mathbf{w}_{\text{random}}$	$\mathbf{w}_{\text{ImageNet}}$	$\mathbf{w}_{\text{IQA}}$	Gain ($\mathbf{w}_{\text{ImageNet}}$ vs $\mathbf{w}_{\text{random}}$)	Gain ($\mathbf{w}_{\text{IQA}}$ vs $\mathbf{w}_{\text{ImageNet}}$)
$\mathbf{v}_{\text{random}}$	0.431	0.683	0.738	+0.252	+0.055
$\mathbf{v}_{\text{IQA}}$	0.417	0.681	0.740	+0.264	+0.059
Gain ($\mathbf{v}_{\text{IQA}}$ vs $\mathbf{v}_{\text{random}}$)	−0.014	−0.002	+0.002	+0.250	+0.058

which consolidates the above observations. We can observe that the frozen random feature extractor underperforms the frozen ImageNet-pretrained feature extractor. Unfreezing the ImageNet-pretrained feature extractor, enabling end-to-end learning, leads to a remarkable performance gain. It is also interesting to observe that the ImageNet-pretrained weights updates with end-to-end joint training (*i.e.* learnable $\mathbf{w}$) can achieve good performance within a single epoch. Since the norm-in-norm loss is used, the curve fitting between the subjective and objective quality scores can be treated as a linear line [15]. The other three subfigures (b-d) in Fig. 2 show the scatter plots between the mean opinion score (MOS) and the predicted quality on KonIQ-10k test set, where large divergence is observed in the case of the frozen random initialization, and the dots distribute closely around the fitted line for the case of ImageNet-pretrained initialized feature extractor.

VQA: Table IV and Table V show the intra- and cross-dataset performance comparisons for VQA networks (ResNet50) with different initialization strategies, respectively. Again, for the feature extractor (column comparison), regardless of how the quality regressor is initialized, ImageNet-pretrained initialization brings huge gains over random initialization. Furthermore, IQA-finetuned initialization provides further improvements, *i.e.*, finetuned IQA models are better feature extractors for VQA. For the quality regressor (row comparison), IQA-trained initialization is superior to random initialization in the intra-dataset setting, no matter the initialization strategy adopted in the feature extractor. However, in the cross-dataset setting, improvement is found only for the best feature extractor (the IQA-finetuned one). Overall, IQA-trained initialization for the whole VQA network achieves the best performance, *i.e.*, 0.830 for KoNViD-1k and 0.740 for LIVE-VQC. We show the SROCC values on KoNViD-1k under different train-val-test splits in Fig. 3, where we can see that IQA-trained initialization helps the optimization of both the feature extractor and the quality regressor.

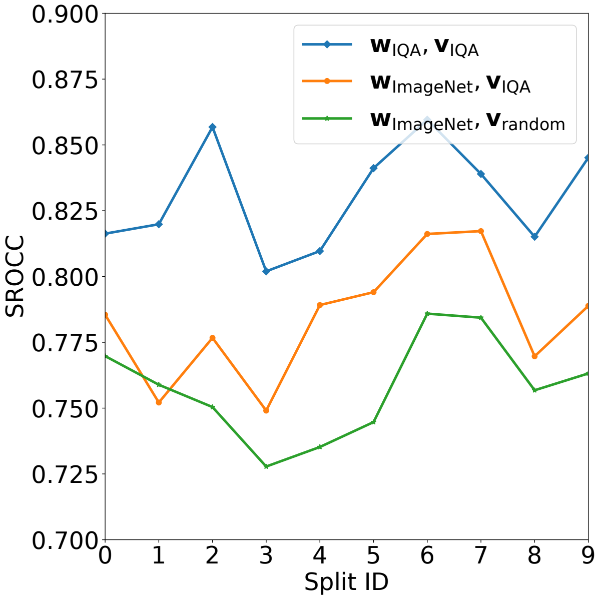


Fig. 3. SROCC on the test set of KoNViD-1k under different splits.

C. Performance Comparison with State-of-the-art Methods

Table VI shows the performance comparison between our method and several representative I/VQA methods on four representative datasets (KonIQ-10k, CLIVE, KoNViD-1k, and LIVE-VQC). The compared methods include ten IQA methods (BRISQUE [32], CORNIA [33], CNNIQA [17], HOSA [38], DeepBIQ [67], KonCept512 [13], HyperIQA [42], LinearityIQA [15], UNIQUE [3], DCNet [68]) and seven VQA methods (TLVQM [53], VSFA [21], CNN-TLVQM [54], VIDEVAL [55], RAPIQUE [56], AGM-VQA [14], Li22 [22]). On the one hand, although the IQA methods perform well on the IQA task, their performance on the VQA task is terrible. On the other hand, most VQA methods (except for VSFA) are not applicable to the IQA task since the corresponding models cannot accept the single-frame video as input. With the consideration of temporal modelling, VQA methods perform much

TABLE VI

Performance comparison with representative blind I/VQA methods. KonIQ-10k and KoNViD-1k columns indicate Intra-dataset results. CLIVE and LIVE-VQC columns indicate cross-dataset results. In each column, the best result is in bold. 'X' means that the I/VQA model at that row is not applicable to the corresponding column's I/VQA dataset. '-' means that the result is not reported from previous work

I/VQA Model		KonIQ-10k	CLIVE	KoNViD-1k	LIVE-VQC
IQA	BRISQUE (TIP 2012)	0.705	0.561	0.654	0.448
	CORNIA (CVPR 2012)	0.780	0.621	0.705	0.372
	CNNIQA (CVPR 2014)	0.572	0.465	-	-
	HOSA (TIP 2016)	0.805	0.628	0.721	0.316
	DeepBIQ (SIViP 2018)	0.872	0.742	-	-
	KonCept512 (TIP 2020)	0.921	0.825	-	-
	HyperIQA (CVPR 2020)	0.906	0.785	-	-
	LinearityIQA (ACM MM 2020)	**0.938**	**0.836**	-	-
	UNIQUE (TIP 2021)	0.896	0.786	-	-
	DCNet (ACM MM 2022)	0.910	0.805	-	-
VQA	TLVQM (TIP 2019)	X	X	0.759	0.572
	VSFA (ACM MM 2019)	-	-	0.794	0.593
	CNN-TLVQM (ACM MM 2020)	X	X	0.820	0.711
	VIDEVAL (TIP 2021)	X	X	0.770	0.592
	RAPIQUE (OJSP 2021)	X	X	0.805	0.627
	AGM-VQA (TMM 2022)	X	X	0.818	0.681
	Li22 (TCSVT 2022)	X	X	0.835	0.695
	GST-VQA (TCSVT 2022)	X	X	0.814	0.680
	DisCoVQA (TCSVT 2023)	X	X	**0.847**	**0.782**
I/VQA	Proposed	0.935	0.829	0.830	0.740

better than the IQA methods on the VQA task. Our method is capable of assessing the quality of both images and videos, which also achieves state-of-the-art (SOTA) performance or is comparable to the SOTA methods. It is worth noting that there is no other extra design on the network architecture or motion information considered in our method. This verifies the importance of network initialization strategies in blind I/VQA tasks.

V. Conclusion

A good initialization is always preferable since it can provide more robust training, less data requirement, faster convergence, and better performance. We have presented a unified blind I/VQA network and systematically studied initialization strategies for optimizing it. We conclude that a good practice for initializing I/VQA models is to initialize the parameters from well-trained models in a relevant task if available. The proposed unified I/VQA network can achieve comparable performance to the state-of-the-art methods with proper initialization. The limitation of this work is that the feature extractor of the VQA network is frozen due to resource constraints. In the future, we would like to design a new VQA network capable of doing end-to-end (video-to-quality) training, with a goal to verify that a good initialization also helps network training when both the quality regressor and the feature extractor are learnable. Besides, visual attention plays an important role in human perception, and we intend to add attention module to improve the performance of the unified I/VQA network.

References

[1] G. Yue, C. Hou, T. Zhou, and X. Zhang, "Effective and efficient blind quality evaluator for contrast distorted images," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2733–2741, 2019.

[2] Q. Jiang, W. Zhou, X. Chai, G. Yue, F. Shao, and Z. Chen, "A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9784–9796, 2020.

[3] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[4] J. Xu, W. Zhou, Z. Chen, S. Ling, and P. Le Callet, "Binocular rivalry oriented predictive autoencoding network for blind stereoscopic image quality measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[5] M. Wang, Y. Huang, J. Lin, W. Xie, G. Yue, S. Wang, and L. Li, "Quality measurement of screen images via foreground perception and background suppression," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[6] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *European Conference on Computer Vision*, 2022, pp. 538–554.

[7] X. Yang, F. Li, L. Li, K. Gu, and H. Liu, "Study of natural scene categories in measurement of perceived image quality," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[8] Y. Liu, X. Yin, Y. Wang, Z. Yin, and Z. Zheng, "HVS-based perception-driven no-reference omnidirectional image quality assessment," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[9] L. Shen, B. Zhao, Z. Pan, B. Peng, S. Kwong, and J. Lei, "Channel recombination and projection network for blind image quality measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[10] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for UGC videos," in *ACM International Conference on Multimedia*, 2022, pp. 856—-865.

[11] C. Yang, P. An, and L. Shen, "Blind image quality measurement via data-driven transform-based feature enhancement," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[12] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *International Conference on Quality of Multimedia Experience*, 2017, pp. 1–6.

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

[14] X. Guan, F. Li, Y. Zhang, and P. C. Cosman, "End-to-end blind video

quality assessment based on visual and memory attention modeling," *IEEE Transactions on Multimedia*, 2022.

[15] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *ACM International Conference on Multimedia*, 2020, pp. 789–797.

[16] D. Mishkin and J. Matas, "All you need is a good init," *International Conference on Learning Representations*, 2016.

[17] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[18] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Generalizable no-reference image quality assessment via deep meta-learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1048–1060, 2021.

[19] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[20] J. Chen, H. Wang, M. Xu, G. Li, and S. Liu, "Deep neural networks for end-to-end spatiotemporal video quality prediction and aggregation," in *IEEE International Conference on Multimedia and Expo*, 2021, pp. 1–6.

[21] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[22] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5944–5958, 2022.

[23] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Contrastive self-supervised pre-training for video quality assessment," *IEEE Transactions on Image Processing*, vol. 31, pp. 458–471, 2022.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[26] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?" *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2019.

[27] X. Yang, F. Li, and H. Liu, "A survey of DNN methods for blind image quality assessment," *IEEE Access*, vol. 7, pp. 123 788–123 806, 2019.

[28] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[29] D. Li, T. Jiang, and M. Jiang, "Recent advances and challenges in video quality assessment," *ZTE Communications*, vol. 17, no. 1, pp. 3–11, 2019.

[30] S. Athar, Z. Wang, and Z. Wang, "Deep neural networks for blind image quality assessment: Addressing the data challenge," *arXiv preprint arXiv:2109.12161*, 2021.

[31] A. Antsiferova, S. Lavrushkin, M. Smirnov, A. Gushchin, D. S. Vatolin, and D. Kulikov, "Video compression dataset and benchmark of learning-based video-quality metrics," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022, pp. 1–12.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[33] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[35] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[36] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[37] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[38] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[39] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.

[40] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 987–994.

[41] X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.

[42] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[43] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma, "FRank: A ranking method with fidelity loss," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 383–390.

[44] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.

[45] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[46] J. Liu, W. Zhou, X. Li, J. Xu, and Z. Chen, "LIQA: Lifelong blind image quality assessment," *IEEE Transactions on Multimedia*, 2022.

[47] R. Ma, H. Luo, Q. Wu, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Remember and reuse: Cross-task blind image quality assessment via relevance-aware incremental learning," in *ACM International Conference on Multimedia*, 2021, pp. 5248–5256.

[48] W. Zhang, D. Li, X. Min, G. Zhai, G. Guo, X. Yang, and K. Ma, "Perceptual attacks of no-reference image quality models with human-in-the-loop," *arXiv preprint arXiv:2210.00933*, 2022.

[49] J. Korhonen and J. You, "Adversarial attacks against blind image quality assessment models," in *ACM International Conference on Multimedia Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 3–11.

[50] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *IEEE International Conference on Image Processing*, 2020, pp. 141–145.

[51] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[52] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

[53] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[54] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *ACM International Conference on Multimedia*, 2020, pp. 3311–3319.

[55] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[56] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 425–440, 2021.

[57] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-in-recurrent network for video quality assessment," in *ACM International Conference on Multimedia*, 2020, pp. 834–842.

[58] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[59] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[61] D. Ghadiyaram and A. C. Bovik, “Massive online crowdsourced study of subjective and objective picture quality,” *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[62] Z. Sinno and A. C. Bovik, “Large-scale study of perceptual video quality,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2019.

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[64] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[65] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

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

[68] Z. Zhou, Y. Xu, R. Xu, and Y. Quan, “No-reference image quality assessment using dynamic complex-valued neural model,” in *ACM International Conference on Multimedia*, 2022, pp. 1006–1015.