Opinion Unaware Image Quality Assessment via Adversarial Convolutional Variational Autoencoder

Ankit Shukla
Bennett University
Greater Noida, India
ankitshukla0165@gmail.com

Avinash Upadhyay
Bennett University
Greater Noida, India
avinres@gmail.com

Swati Bhugra
Indian Institute of Technology Delhi
New Delhi, India
swati6102200ece@gmail.com

Manoj Sharma
Bennett University
Greater Noida, India
mksnith@gmail.com

Abstract

Image quality assessment is a challenging computer vision task due to the lack of corresponding reference (pristine) images. This no-reference bottleneck has been tackled with the utilisation of subjective mean opinion scores (MOS) termed as supervised blind image quality assessment (BIQA) methods. However, inaccessible opinion score scenarios limits their applicability. To relieve these limitations, we propose to employ reconstruction based learning trained only on pristine images. This permits an implicit distribution learning of pristine images and the deviation from this learned feature distribution is subsequently utilised for unsupervised image quality assessment. Specifically, an adversarial convolutional variational auto-encoder framework is employed with KL divergence, perceptual and discriminator loss. With state-of-the-art results on four benchmark datasets, we demonstrate the effectiveness of our proposed framework. An ablation study has also been conducted to highlight the contribution of each module i.e. loss and quality metric for an efficient unsupervised BIQA.

1. Introduction

The acquisition, compression, transmission, and storage of digital images inevitably introduces noise/distortions in images. These distortions directly affect the reliability of subsequent image processing. For instance, the accuracy of disease diagnosis depends on the quality of medical images [48]. Thus, current computer vision research has focused on image quality assessment (IQA) as a crucial pre-processing task. The objective of IQA is to automatically quantify image quality consistent with human assessments [50].

Existing IQA methods are primarily categorised into three classes (1) Full-Reference (FR), (2) Reduced-Reference (RR), and (3) No-Reference (NR) methods [48]. This categorisation is based on the availability of a reference (pristine) image with the corresponding distorted image. FR and RR methods have shown promising performance with complete and partial reference image information respectively [50]. In real-world scenarios, reference images are often unavailable, such as for authentically distorted images. Thus, NR-IQA permits broad applicability in contrast to FR and RR methods.

No-Reference IQA methods quantify perceptual quality without relying on reference (pristine) images. These NR-IQA methods are also termed as blind image quality assessment (BIQA) methods. A majority of the existing BIQA methods utilise a supervised regression model trained on distorted images and the corresponding mean opinion scores (MOS) [48], where MOS is a subjectively generated score based on the perceptual quality of an image. These supervised methods are also termed as opinion aware BIQA methods. With the advancement in deep learning, current opinion aware methods are based on the utilisation of convolutional neural networks (CNNs) [2, 4, 60]. The optimal training of deep learning models requires large training samples with MOS. But the cost of subjective annotations is time-consuming and requires multiple assessments. Thus, these opinion aware BIQA methods lack generalizability [50]. To mitigate the dependence on large annotated data, unsupervised (opinion unaware) BIQA methods have been introduced. These methods do not rely on expensive subjective scores and may provide better applicability. Thus, this paper focuses on opinion unaware BIQA methods.

In this context, we propose to employ reconstruction based strategy for quality-aware feature learning from pristine images. Generative Adversarial Network (GAN) [11] based frameworks have already been proposed in this context. For example, a quality-aware GAN was developed to generate a hallucinated reference conditioned on the distorted image [22]. Similarly, authors in [25] proposed to generate the primary content of a distorted image. Another study [38] presented a restorative adversarial network to reconstruct the input distorted images. It is to be noted that these aforementioned methods belong to the supervised/opinion aware BIQA paradigm. That is they rely on subjectively generated mean opinion scores. In contrast, the proposed framework is unsupervised/opinion unaware BIQA method trained only on pristine images that permits an implicit rich feature distribution learning.

Recently, authors in [41,66] proposed data augmentation strategies for generating positive and negative pairs to be utilised in a contrastive self-supervised paradigm. Specifically, this pretext task permits the learning of quality aware features without relying on MOS. However, the quality prediction based on these learnt features still relied on MOS, formulated as a supervised regression model. In contrast, we propose to utilise the deviation from the learned rich feature distribution for unsupervised image quality assessment. Another study [62] employed a pairwise learning-to-rank loss for extracting quality aware features. This loss formulation also relied on mean opinion score at the training stage. In contrast, the proposed training paradigm is not constrained by specific distortion types, datasets, or subjectively generated MOS during training. Similar to our work, few studies [34,51,58] have attempted to model multivariate Gaussian (MVG) based on statistical features response from patches of only pristine images. Although, low-dimensional their performance is highly dependent on expert knowledge. In addition, due to the diverse degradation in natural images these features are also sensitive to one distortion type and may not be applicable to other distortions. However, the proposed reconstruction based strategy permits an implicit rich feature distribution learning with no prior assumptions. The contributions of this paper are as follows:

- We propose to utilise reconstruction based learning via adversarial variational autoencoder network. Trained only on pristine images, the proposed framework is unconstrained with respect to specific distortion types, datasets, or MOS.

- A designed combination of loss functions permits content-agnostic and quality-aware feature learning. The deviation from this learned feature distribution is utilised for unsupervised/opinion unaware BIQA method at the inference stage.

- An ablation study has been conducted to show the effectiveness of the proposed modules with respect to loss and image quality metrics.

The rest of the paper is organised as follows: Section 2 gives an overview of the relevant BIQA literature, and Section 3 presents the proposed methodology. Experiments and results analysis are included in Section 4. Finally, section 5 concludes the paper.

2. Related work

Blind image quality assessment (BIQA) methods are broadly categorised as (1) Supervised BIQA methods that utilise opinion scores in addition to the distorted images also termed as opinion-aware BIQA methods, (2) Weakly supervised BIQA methods that derive opinion scores from existing full reference IQA methods and (3) Unsupervised BIQA methods that only utilises distorted images also termed as opinion-unaware BIQA methods. In this section, we review BIQA methods based on the aforementioned categorisation.

2.1. Supervised BIQA Methods

These methods rely on image quality scores along with distorted images. In this context, few studies proposed to utilise hand-crafted features that encapsulate the repeating patterns of natural scenes termed as Natural Scene Statistics (NSS) extracted from (a) spatial domain, (b) discrete cosine transform domain, and (c) wavelet domain. Blind/referenceless image spatial quality evaluator (BRISQUE) [33], blind image integrity notator (BLIIDNS) [39], BLIIDNS-II [40], and blind image quality index (BIQI) [35] are widely adopted methods belonging to this category. In contrast, few methods proposed to encapsulate the edge distribution of image gradients such as gradient magnitude map and the Laplacian of the Gaussian response (GMLOG) [52], blind structural degradation (BSD) [20] and no-reference structural and luminance (NRSL) [21]. Authors in [56] proposed to automatically learn codebook representation (CORNIA) from raw images and a support vector regressor was employed to develop a learned model with codebook representation and subjective scores. Distortion identification frameworks have also been proposed with the underlying assumption that distortions significantly modify the image's statistical characteristics. For example, Distortion Identification based Image Verity and INtegrity Evaluation index (DIIVINE) was proposed in [36] that performs distortion identification followed by image quality evaluation. DeepBIQ [2] utilised multiple sub-regions features from fine-tuned CNNs, followed by average pooling of these regions for image quality assessment. Similarly, authors in [4,60] also proposed two-stage CNN-based frameworks for synthetic and authentic distortions. Transformer-based image quality assessment methods have

been investigated in [9]. In contrast to these deep-learning networks, the proposed framework does not rely on image quality scores.

2.2. Weakly Supervised BIQA Methods

These methods rely on Full-reference (FR) methods to derive pseudo subjective scores for image quality assessment. For example, authors [53] used the FR method presented in [12] to generate pseudo labels, which were subsequently used in support vector regression. Similarly, reciprocal rank fusion (RRF) was used in [55] to train the image quality assessment method on pseudo labels. In another study [49], authors used different FR methods for each distortion and combined these distortion-specific scores to predict the overall image quality. However, these methods suffers from the limitations of the adopted FR models for generating labeled training dataset.

In addition to pseudo scores, methods based on pseudo-ranking order have also been proposed. In [23], authors generated ranked image pairs by introducing various distortion levels in the reference images to train a Siamese network. In contrast to an efficient back-propagation strategy proposed in the previous study, authors in [27, 28] proposed a loss function and perceptual uncertainty index for optimal training of the Siamese network. Specifically, pseudo-quality scores in these studies were automatically generated via FR models such as Multiscale structural similarity (MS-SSIM), visual information fidelity (VIF), and gradient magnitude similarity deviation (GSMD). In the majority of these methods, the pseudo-ranking scores were generated from the reference image with the same distortion type but varying distortion levels. Thus, these labeled training datasets do not capture the pseudo-ranking between different distortions.

2.3. Unsupervised BIQA Methods

In contrast to the previously mentioned image quality methods, few methods have been explored that alleviate the limitation posed by the dependence on image quality scores. For instance, authors in [34] developed a natural image quality evaluator (NIQE) that learns a multivariate Gaussian (MVG) model on natural scene statistics (NSS) features extracted from patches of pristine natural images. The image quality score is then computed via distance between the MVG model of the investigated image and the learned pristine image MVG model. This method was extended in [58] as Integrated Local NIQE (ILNIQE) by introducing additional statistical features i.e. color, gradient, and Log-Gabor filter response. However, they generate image quality scores that assign more weights to the salient patches of the investigated image based on a pre-trained CNN. Wu et al. [51] derived statistical features from binary patterns of local image structures (method denoted as LPSI) for BIQA. Based on the empirical studies that highlight the unimodal assessment of the aforementioned studies, authors in [24] introduced structure, naturalness, and perception quality based degradations for BIQA. In contrast to the previously mentioned methods that compute the distance between features of investigated image patches to a corpus of pristine images, self-supervised frameworks have also been explored. In [30] the authors formulated the pretext task as predicting distortion types and its distortion levels for feature learning (method denoted as CONTRIQUE). It is to be noted that the investigated images in authentically distorted images show variations in contents and degradation types, this instance-level discrimination may limit the learning of quality-aware features. A similar framework was also utilised in [31] on synthetically generated images. Another study [5] employed patch prediction as a pretext task for synthetically distorted images (method denoted as SPIQ).

In addition, methods based on specific distortion types have also been proposed. Authors in [13] introduced a multi-step IQA framework that systematically aggregates distortion-specific single-quality metrics with different distortion effects. In another study [65], the distortion parameters of singly and multiply distorted images were predicted based on natural scene statistics (NSS) features. In contrast, authors in [32] proposed to generate a reference image based on the introduction of severest distortion and compared it with the investigated image for quality prediction.

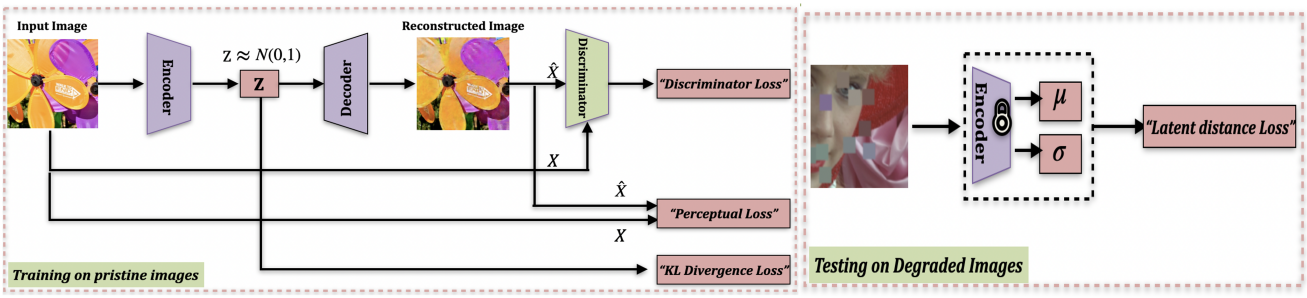
It is evident that most opinion-aware BIQA methods have been proposed for distortion-specific or synthetic distortion-type scenarios. However, due to the diverse degradation characteristics, these studies are limited in terms of their applicability in authentically distorted scenarios.

3. Methodology

The workflow of the proposed framework is shown in Figure 1. The framework comprises of two stages that accomplish (1) pristine data distribution learning and (2) effectively using the learned distribution from the first stage to perform unsupervised BIQA. Specifically, the first stage focuses on learning rich feature distribution of pristine images based on high-quality image reconstruction. And the second stage focuses on quality metric generation of the degraded/distorted test image based on the learned pristine feature distribution. We elucidate these stages in the following subsections.

3.1. Unsupervised Latent Space Learning

The first stage learns to generate pristine images based on Convolutional Variational Autoencoder Generative Adversarial Network (CVAE-GAN) [14]. It combines adver-



(a) Training workflow of the proposed framework. (b) Testing workflow of the proposed framework

Figure 1. (a) Training workflow of the proposed framework. Here, X denotes the input pristine image, $\hat{X}$ denotes the reconstructed image and (b) Testing workflow of the proposed framework. Here, μ and σ are the mean and standard deviation of the test image.

sarial learning [11] with the unsupervised learning of convolutional variational autoencoders.

CVAE-GAN comprises of two main components: a Convolutional Variational Autoencoder (CVAE) and a discriminator. CVAE is a variant of autoencoder [64] with a probabilistic model consisting of an encoder and a decoder. The encoder generates a low-dimensional representation termed as the "latent code" or "latent representation" of the input images. This latent representation is then fed to the decoder to generate an image similar to the input image. Specifically, the encoder is trained to map the input data with the latent space distributed according to a probability distribution (for example Gaussian distribution). The second component of the CVAE-GAN model is the discriminator network.

We propose to train CVAE-GAN with only pristine (high-quality) images to learn its latent space representation. While the CVAE learns to reconstruct pristine images, the discriminator network takes the reconstructed image as input and outputs a probability score belonging to the pristine distribution. The two networks CVAE and discriminator are trained in an adversarial manner. Where, the CVAE attempts to fool the discriminator by generating high-quality images and the discriminator's task is to distinguish between the original pristine images and the generated images. Kullback-Leibler Divergence (KLD) loss [6] is used to enforce the latent distribution belonging to the pristine distribution. The Kullback-Leibler divergence [17] is formulated as:

$$L_{KLD} = KL(q(z|x;\phi)||p(z)) = \frac{1}{2}\sum_{j=1}^{J}(1 + log(\sigma_j^2) - \mu_j^2 - \sigma_j^2) \quad (1)$$

where, J is the dimensionality of the latent space, μ_j and σ_j are the mean and standard deviation of the j^{th} dimension of the encoder distribution, and $p(z)$ is a standard normal distribution. We also utilised VGG loss [45] to reconstruct perceptually enhanced images. Reconstruction loss based on VGG loss is given by:

$$L_{per}(x;\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}((f_i - \hat{f}_i)^2) \quad (2)$$

where, f_i is the i^{th} VGG feature of the input image, and $\hat{f}_i$ is the corresponding feature in the reconstructed image. The loss for the CVAE is given by:

$$L_{CVAE}(x;\theta,\phi) = L_{per}(x;\theta,\phi) + KL(q(z|x;\phi)||p(z)) \quad (3)$$

where x is the input image, θ and ϕ are the parameters of the CVAE, z is the latent variable, $q(z|x;\phi)$ is the encoder distribution, $p(z)$ is the prior distribution, $L_{per}(x;\theta,\phi)$ is the perceptual loss, and $KL(q(z|x;\phi)||p(z))$ is the Kullback-Leibler divergence between the encoder and prior distributions. To summarise, the training loss of the CVAE-GAN can be written as:

$$L_{tot}(x,z;\theta_g,\theta_d) = E_{x\sim p_{data}(x)}[log(D(x;\theta_d))] + E_{z\sim p(z)}[log(1 - D(CVAE(z;\theta_g);\theta_d))] \quad (4)$$

where x is a real image, z is a random noise vector given by CVAE, θ_g and θ_d are CVAE and discriminator networks' parameters. The discriminator loss ensures high quality-aware feature learning and also mitigate content-based learning.

3.2. Unsupervised Image Quality Assessment

The first stage of the proposed framework permits feature distribution learning of only the pristine images. We then propose to utilise the learned weights of the encoder at the inference stage (shown in Figure 1(b)) for unsupervised BIQA. This hypothesis was motivated by the empirical observation shown in Figure 2. Figure 2 shows the scatter plot of latent representations on CLIVE [8] and KonIQ-10k [15] datasets. t-SNE algorithm has been employed to reduce the dimensionality of these latent representation and project it to a 2D space for qualitative analysis. Specifically, the figure shows that the latent representation of pristine images lies close to the learned distribution whereas, the latent representations of distorted/noisy images shows separability. Thus, the deviation from the learned feature rich distribution of the test image provides a pseudo indicator of the noise/degradations, thus facilitating unsupervised BIQA. To

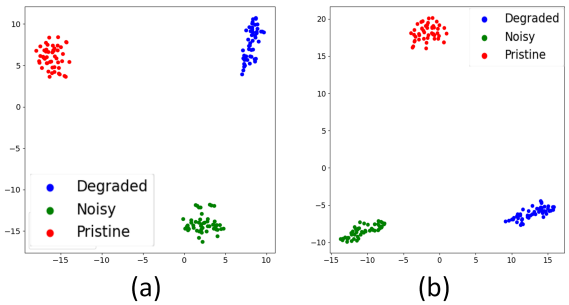


Figure 2. t-SNE visualization of the latent representations from (a) CLIVE [8] dataset (b) FLIVE [43] dataset.

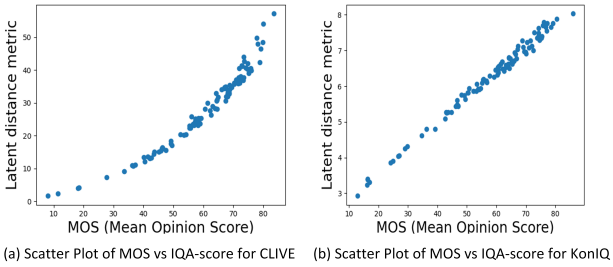


Figure 3. Correlation of latent distance metric with Mean Opinion Score. (a) CLIVE dataset and (b) KonIQ dataset.

quantify this deviation, we employ latent distance metric (S_Z) proposed in [1], formulated as:

$$S_Z(\mu_{hd}, \mu_t, \Sigma_{hd}, \Sigma_t) = \sqrt{(\mu_{hd} - \mu_t)^T (\Sigma_{hd} + \Sigma_t)^{-1} (\mu_{hd} - \mu_t)} \quad (5)$$

Here, μ_{hd} and Σ_{hd} are the parameters of the learned distribution, and μ_t and Σ_t are the parameters of the test data distribution.

The latent distance metric computes the distance between the latent representation and a reference point in the latent space. The reference point is obtained by averaging the latent representation of pristine images in each dataset. This metric reflects the quality deviation of the test image from the pristine quality at the inference stage. To summarise, the second stage permits unsupervised BIQA based on the learned encoder and latent space representation of pristine images without relying on subjective MOS.

3.3. Implementation Details

In this study, we investigated the combinations of VGG, KLD, and discriminator loss to learn rich features based on the reconstruction of pristine images. Pre-trained VGG-19 network [3] was employed to extract the features from the intermediate layers to calculate the VGG loss [16]. An Adam optimizer with a learning rate of 0.001 and a batch size of 8 was used to train CVAE-GAN. The implementations used PyTorch framework with RTX 2080 Ti and Tesla V100 GPUs.

4. Experiments & Results

4.1. Datasets

It is to be noted that the training of the proposed framework rely only on the pristine images, this permits the utilisation of pristine images from multiple datasets. Specifically, we employed 140K pristine images from the KADIS [19] dataset and 4744 pristine images from the Waterloo exploration [26] dataset to learn the distribution of high-quality image features.

The proposed framework was evaluated on LIVE [44], CLIVE [8], FLIVE [43], KADID-10k [19], KonIQ-10k [15] and SPAQ [7] datasets. Among these datasets, LIVE [44] and KADID-10k [19] are legacy datasets containing synthetic distortions. LIVE contains 779 distorted images generated from 29 reference images with five types of distortion: JPEG compression, JPEG2000 compression, Gaussian blur, Gaussian noise, and fast fading. KADID-10k [19] contains 10,125 distorted images generated from 81 reference images with 25 types of distortion at five levels of severity.

KonIQ-10k [15], CLIVE [8], FLIVE [43] and SPAQ [7] are 'In the Wild' datasets containing authentic distortions. The KonIQ-10k [15] dataset consists of 10,073 natural images and is rated by 1,459 human observers using a crowdsourcing platform. CLIVE [8] consists of 1,162 natural images and FLIVE contains 982 images with different distortion types such as JPEG compression, Gaussian blur, white noise, and bit error. The SPAQ [7] dataset consists of 11,125 smartphone-captured natural images rated by 4,876 observers. In the KonIQ-10k and SPAQ datasets, the subjective scores are generated using an absolute category rating (ACR) method and then converted to MOS. In contrast, distorted images in the LIVE dataset are scored using a double-stimulus continuous quality scale (DSCQS) method.

4.2. Results

In this subsection, we firstly show the comparative analysis with state-of-the-art BIQA methods on different datasets. To highlight that the proposed framework is unconstrained with respect to specific distortion types, an additional comparative analysis was also conducted with different distortions. Lastly, ablation studies are included to show the contribution of quality metrics and loss functions in our proposed framework. The proposed framework was evaluated on IQA metrics (1) SRCC [46] and (2) PLCC [37]. These metrics measure the correlation between the predicted quality scores i.e. the latent distance metric (in our study) and MOS.

4.2.1 Comparative analysis on Benchmark Datasets

This analysis was conducted on 5 datasets: LiVE [44] KADID-10K [19], CLIVE [8], SPAQ [7] and KonIQ-10k

Table 1. Comparative analysis of the proposed framework with state-of-the-art BIQA models based on median SRCC and PLCC across ten sessions. '*' denotes the values reported in the original paper, BRISQUE [33] and CORNIA [56] results are reported from CONTRIQUE [31], remaining results are from LIQE [63].

Dataset	LIVE [44]		KADID-10k [19]		CLIVE [8]		KonIQ-10k [15]		SPAQ [7]	
Methods	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE [34]	0.908	0.904	0.389	0.442	0.446	0.507	0.415	0.438	0.697	0.685
ILNIQE [59]	0.887	0.894	0.565	0.611	0.469	0.518	0.509	0.534	0.719	0.654
Ma19 [29]	0.922	0.923	0.465	0.501	0.336	0.405	0.360	0.398	-	-
BRISQUE [33]	0.939	0.935	0.528	0.567	0.608	0.629	0.665	0.681	0.809	0.817
CORNIA [56]	0.947	0.950	0.516	0.558	0.629	0.671	0.780	0.795	0.709	0.725
CONTRIQUE* [31]	0.960	0.961	0.934	0.937	0.845	0.857	0.894	0.906	0.914	0.919
PaQ2PiQ [57]	0.544	0.558	0.403	0.448	0.732	0.755	0.722	0.716	-	-
KonCept [15]	0.673	0.619	0.503	0.515	0.778	0.799	0.911	0.924	-	-
MUSIQ [18]	0.837	0.818	0.572	0.584	0.785	0.828	0.915	0.937	0.917	0.921
HyperIQA [47]	0.966	0.968	0.872	0.869	0.855	0.878	0.900	0.915	0.915	0.918
TreS [10]	0.965	0.963	0.881	0.879	0.846	0.877	0.907	0.924	-	-
UNIQUE [61]	0.961	0.952	0.884	0.885	0.854	0.884	0.895	0.900	-	-
LIQE* [63]	0.970	0.951	0.930	0.931	**0.904**	0.910	0.919	0.908	-	-
Re-IQA* [42]	0.970	**0.971**	-	-	0.840	0.854	0.914	0.932	-	-
QPT-ResNet50* [67]	0.610	0.677	-	-	0.894	**0.914**	**0.927**	**0.941**	0.925	0.927
Proposed	**0.976**	0.970	**0.936**	**0.941**	0.862	0.871	0.909	0.921	**0.931**	**0.942**

[15]. Specifically, the proposed framework was evaluated on all five datasets across ten sessions and the median SRCC and PLCC values are reported in Table 1. Among the selected BIQA methods: NIQE [34], ILNIQE [59], and Ma19 [29] are unsupervised/opinion unaware methods whereas the remaining methods belong to supervised BIQA paradigm. For comparison, we used SRCC and PLCC values of the state-of-the-art models either from [63] or from the original paper.

It is evident in Table 1, that the performance of opinion-unaware methods varies across all datasets. Similarly, supervised BIQA methods trained on large dataset or a combination of datasets do not show generalized performance on datasets with synthetic and authentic distortions. In contrast, the proposed framework equipped with an efficient training strategy i.e. trained on a large pristine image dataset outperforms opinion-unaware methods on both authentic distortions and synthetic distortions. In comparison with other supervised methods that either rely on opinion scores at some stage or are trained with both distorted images and the corresponding pristine images, we show comparable performance. Thus, highlighting the effectiveness of the training paradigm to learn quality-aware features and its generalisation applicability on both authentic and synthetic noise.

4.2.2 Comparative analysis on Different Distortions

We also conducted the performance evaluation of our framework with different distortion types namely, "White Noise", "Gaussian Blur", "JPEG2000 compression", JPEG compression, and "Bit Errors in JPEG2000 Stream" from LIVE [44]. The performance is contrasted with opinion-unaware methods: NIQE [34], QAC [54], and IL-NIQE [59] based on SRCC evaluation metric. This quantitative evaluation is shown in Table 2 and qualitative evaluation on various distortion types is depicted in Figure 4. Figure 4 highlights the contrast between latent representations of differently distorted images.

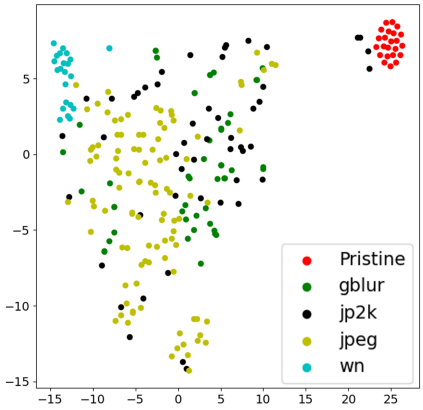


Figure 4. t-SNE visualisation of the latent representations from LIVE [44] dataset.

To summarise, the aforementioned comparative analysis show that the proposed framework can handle diverse and complex distortions in natural images for image quality assessment. To accomplish this task, the learned compact and regularised latent representation encodes the quality-related

Table 2. Comparative analysis of the proposed framework with state-of-the-art opinion-unaware BIQA methods based on SRCC metric.

Distortion types	NIQE [34]	QAC [54]	IL-NIQE [59]	Proposed
White Noise	0.9718	0.9511	0.9807	**0.9885**
Gaussian Blur	0.9328	0.9134	0.9153	**0.9516**
JPEG2000 compression	0.9186	0.8621	0.8939	**0.9234**
JPEG compression	0.9412	0.9362	0.9418	**0.9618**
Bit Errors in JPEG2000 Stream	0.8635	0.8231	0.8327	**0.8961**

Table 3. Ablation study with respect to Training Loss. Median SRCC and PLCC across ten sessions are reported.

Dataset	LIVE [44]		KADID-10K [19]		CLIVE [8]		KonIQ-10K [15]		SPAQ [7]	
Methods	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
MSE + KLD	0.853	0.892	0.817	0.824	0.719	0.791	0.723	0.868	0.863	0.893
VGG + KLD	0.927	0.943	0.901	0.918	0.826	0.852	0.816	0.907	0.917	0.931
MSE + KLD+ Disc	0.951	0.957	0.916	0.919	0.847	0.859	0.854	0.915	0.921	0.937
VGG + KLD+ Disc	**0.976**	**0.970**	**0.934**	**0.932**	**0.862**	**0.871**	**0.909**	**0.921**	**0.931**	**0.942**

features of the pristine images. This facilitates the computation of image quality scores that are highly correlated with subjective quality scores.

4.2.3 Ablation Study with respect to Training Loss

We conducted an ablation study to show the contribution of the proposed combination of VGG loss, KLD loss, and Discriminator loss. In this study, we first trained two variants of CVAE, one uses the combination of Mean Squared Error loss and KLD loss while the other uses a combination of VGG loss (perceptual loss) and KLD loss. The aforementioned variants of CVAE-GAN was further trained on an additional discriminator loss. We conducted the evaluation across ten sessions and report the median SRCC and PLCC in Table 3.

MSE loss encourages the CVAE to produce images with minimum pixel-wise differences. However, blur and artifacts are present in the reconstructed images, resulting in low SRCC and PLCC across all datasets. VGG loss captures the perceptual and semantic information of the image, thus using the VGG loss instead of MSE loss generates realistic images that preserve the content and style of the image. The effect of VGG loss is clearly shown in the performance improvement of the CVAE for all datasets. The use of discriminator loss introduces diversity and realism to the reconstructed images that may not be captured with VGG loss alone that is primarily focused on perceptual quality and mitigating the role of content in the image. It is clear from Table 3 that the proposed combination of VGG loss, KLD loss, and discriminator loss provides the best performance on both synthetic and realistic distortions. Figure 5 shows a qualitative comparison of the generated pristine images with different loss functions. In the figure, three sample images with different distortion types are used to qualitatively and quantitatively compare the reconstruction efficacy of different training paradigms. For quantitative comparison of the reconstructed images, PSNR and SSIM are employed. It is evident that the proposed framework produces the best reconstruction results. MSE+KLD loss tends to produce blurry images that lose some details and textures. Whereas, VGG+KLD loss tends to produce sharper images that preserve some details and textures, but also introduce some artifacts and color shifts. The VGG+KLD+Discriminator loss tends to produce realistic images with restored details and textures.

4.2.4 Ablation Study with respect to Quality Metric

Table 4 shows the performance evaluation of the proposed framework wth respect to various image quality metrics. Specifically, we compare the reconstruction error metric (REM) and latent distance metric (LDM). It is to be noted that in the first stage, the discriminator has been trained to differentiate between low and high-quality images in an adverserial manner. Thus, we also propose to utilise the discriminator prediction score (DSM) [68] to quantify the noise/degradation (i.e. quality). It is evident that LDM achieves better performance than REM and DSM on all datasets showing its efficacy to measure the image quality from the learned latent representation. Figure 3 shows the correlation of latent distance metric computed from the learned distribution with MOS.

5. Conclusion

In this study, we presented a novel training strategy based on an adversarial variational autoencoder (CVAE-GAN) for opinion unaware BIQA. Trained only on pristine images, the proposed framework is unconstrained with respect to specific distortion types, datasets, and MOS. A designed combination of loss functions permits content-agnostic and quality-aware feature learning. The deviation from this learned feature distribution is utilised for unsuper-

Table 4. Ablation study with respect to Quality Metric. Median SRCC and PLCC across ten sessions are reported.

Image quality metric	KonIQ [15]		CLIVE [8]		FLIVE [43]		SPAQ [7]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Reconstruction metric	0.681	0.795	0.653	0.741	0.419	0.387	0.852	0.873
Discriminator metric	0.883	0.902	0.847	0.823	0.536	0.491	0.919	0.907
Latent distance metric	**0.909**	**0.921**	**0.862**	**0.871**	**0.571**	**0.638**	**0.931**	**0.942**

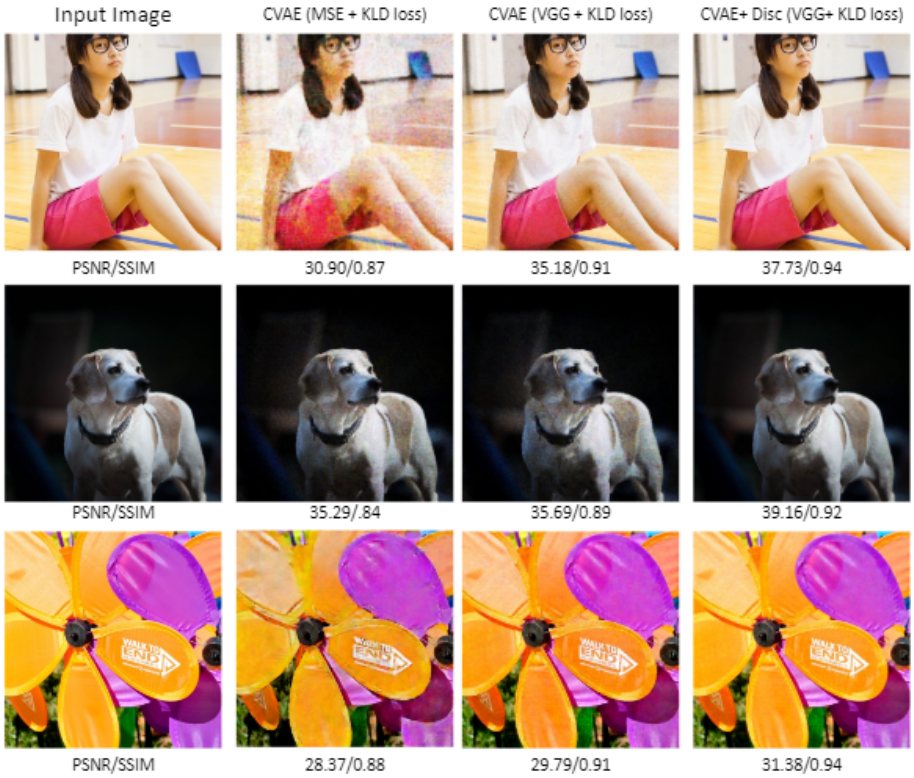


Figure 5. Ablation study with respect to different loss functions. PSNR/SSIM metric is reported with the reconstructed image.

vised/opinion unaware BIQA method at the inference stage.

Based on the objective evaluation of the distorted images, we highlighted the superior performance of our proposed framework compared to state-of-the-art unsupervised BIQA on four authentically distorted benchmark datasets. Thus, demonstrating the effectiveness of our proposed design with respect to loss and quality metrics. Our work delves into utilising the learned latent space of widely available pristine image, permitting unsupervised/opinion aware quality assessment applicable in real world scenarios.

References

[1] Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan. No reference opinion unaware quality assessment of authentically distorted images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2459–2468, 2023. 5

[2] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12:355–362, 2018. 1, 2

[3] Tiago Carvalho, Edmar RS De Rezende, Matheus TP Alves, Fernanda KC Balieiro, and Ricardo B Sovat. Exposing computer generated images by eye's region classification via transfer learning of vgg19 cnn. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 866–870. IEEE, 2017. 5

[4] Luigi Celona and Raimondo Schettini. Cnn-based image quality assessment of consumer photographs. In *London Imaging Meeting*, volume 2020, pages 129–133. Society for Imaging Science and Technology, 2020. 1, 2

[5] Pengfei Chen, Leida Li, Qingbo Wu, and Jinjian Wu. Spiq: A self-supervised pre-trained model for image quality assessment. *IEEE Signal Processing Letters*, 29:513–517, 2022. 3

[6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 4

[7] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 5, 6, 7, 8

[8] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016. 4, 5, 6, 7, 8

[9] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 3

[10] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 6

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 4

[12] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia*, 18(6):1098–1110, 2016. 3

[13] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting*, 60(3):555–567, 2014. 3

[14] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample, 2020. 3

[15] 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*, 29:4041–4056, 2020. 4, 5, 6, 7, 8

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[17] James M. Joyce. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. 4

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 6

[19] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 789–797, 2020. 5, 6, 7

[20] Qiaohong Li, Weisi Lin, and Yuming Fang. Bsd: Blind image quality assessment based on structural degradation. *Neurocomputing*, 236:93–103, 2017. 2

[21] Qiaohong Li, Weisi Lin, Jingtao Xu, and Yuming Fang. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia*, 18(12):2457–2469, 2016. 2

[22] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 732–741, 2018. 2

[23] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017. 3

[24] Yutao Liu, Ke Gu, Yongbing Zhang, Xiu Li, Guangtao Zhai, Debin Zhao, and Wen Gao. Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):929–943, 2019. 3

[25] Jupo Ma, Jinjian Wu, Leida Li, Weisheng Dong, and Xuemei Xie. Active inference of gan for no-reference image quality assessment. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[26] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017. 5

[27] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017. 3

[28] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P Simoncelli. Blind image quality assessment by learning from multiple annotators. In *2019 IEEE international conference on image processing (ICIP)*, pages 2344–2348. IEEE, 2019. 3

[29] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P Simoncelli. Blind image quality assessment by learning from multiple annotators. In *2019 IEEE international conference on image processing (ICIP)*, pages 2344–2348. IEEE, 2019. 6

[30] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 3

[31] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using synthetic images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 93–102, 2022. 3, 6

[32] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia*, 20(8):2049–2062, 2017. 3

[33] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 2, 6

[34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2, 3, 6, 7

[35] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters*, 17(5):513–516, 2010. 2

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

[37] Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895. 5

[38] Hongyu Ren, Diqi Chen, and Yizhou Wang. Ran4iqa: Restorative adversarial nets for no-reference image quality assessment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2

[39] Michele A Saad, Alan C Bovik, and Christophe Charrier. A dct statistics-based blind image quality index. *IEEE Signal Processing Letters*, 17(6):583–586, 2010. 2

[40] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 2

[41] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 2

[42] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 6

[43] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 5, 8

[44] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 5, 6, 7

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4

[46] Charles Spearman. The proof and measurement of association between two things. 1961. 5

[47] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 6

[48] Zhihua Wang, Zhi-Ri Tang, Jianguo Zhang, and Yuming Fang. Toward a blind image quality evaluator in the wild by learning beyond human opinion scores. *Pattern Recognition*, page 109296, 2023. 1

[49] Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on image processing*, 29:7414–7426, 2020. 3

[50] Leyuan Wu, Xiaogang Zhang, Hua Chen, Dingxiang Wang, and Jingfang Deng. Vp-niqe: An opinion-unaware visual perception natural image quality evaluator. *Neurocomputing*, 463:17–28, 2021. 1

[51] Qingbo Wu, Zhou Wang, and Hongliang Li. A highly efficient method for blind image quality assessment. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 339–343. IEEE, 2015. 2, 3

[52] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014. 2

[53] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 995–1002, 2013. 3

[54] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 995–1002, 2013. 6, 7

[55] Peng Ye, Jayant Kumar, and David Doermann. Beyond human opinion scores: Blind image quality assessment based on synthetic scores. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4241–4248, 2014. 3

[56] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012. 2, 6

[57] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 6

[58] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 2, 3

[59] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 6, 7

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

[61] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 6

[62] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 2

[63] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 6

[64] Yifei Zhang. A better autoencoder for image: Convolutional autoencoder. In *ICONIP17-DCEC. Available online: http://users. cecs. anu. edu. au/Tom. Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58. pdf (accessed on 23 March 2017)*, 2018. 4

[65] Yi Zhang and Damon M Chandler. Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation. *IEEE Transactions on Image Processing*, 27(11):5433–5448, 2018. 3

[66] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023. 2

[67] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023. 6

[68] Yunan Zhu, Haichuan Ma, Jialun Peng, Dong Liu, and Zhiwei Xiong. Recycling discriminator: Towards opinion-unaware image quality assessment using wasserstein gan. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 116–125, New York, NY, USA, 2021. Association for Computing Machinery. 7