No Reference Opinion Unaware Quality Assessment of Authentically Distorted Images

Nithin C Babu*[1], Vignesh Kannan*†[2], and Rajiv Soundararajan[1]
[1]Indian Institute of Science, Bengaluru, India
[2]Mercedes-Benz Research and Development India
nithinc@iisc.ac.in, vigneshkannan1696@gmail.com, rajivs@iisc.ac.in

Abstract

The quality assessment (QA) of camera captured authentically distorted images is important on account of its ubiquitous applications and challenging due to the lack of a reference. While there exists a plethora of supervised no reference (NR) image QA (IQA) algorithms, there is a need to study unsupervised or opinion unaware algorithms on account of their superior generalization performance. We explore self-supervised learning (SSL) for the feature design on authentically distorted images to predict quality without training on human labels. While SSL on synthetic distortions has recently shown promise, there is a need to enrich the feature learning on authentic distortions. The key challenge in achieving this is in the learning of quality sensitive features with mitigated content dependence. We design a self-supervised contrastive learning approach which only requires positives and introduce a content separation loss by estimating a bound on the mutual information between the features learnt and the content information. We show on multiple authentically distorted datasets that our self-supervised features can predict image quality by comparing with a corpus of pristine images and achieve state-of-the-art performance.[§]

1. Introduction

Image quality assessment (IQA) is an important tool in benchmarking and comparing different camera captured authentically distorted images. Particularly, the setting of No Reference (NR) IQA becomes important in this scenario where a reference pristine image is not available for comparison. NR IQA algorithms are typically designed in a learning framework by training on human opinion scores. However, human labels are hard to obtain for training owing to the need to conduct large scale subjective studies, which are cumbersome. Further, exploring the full capability of deep learning for NR IQA becomes challenging owing to the need for a large number of annotated images. Thus, we focus on the problem of unsupervised or opinion unaware NR IQA for authentically distorted images without training on human opinion scores.

Although supervised models trained with human labels on a single large dataset are beginning to show promising results in cross-dataset experiments [36, 34, 7, 27], as cameras, sensors and algorithms keep evolving, the distortions that one encounters also keep evolving with time. Conducting large scale studies continuously to design and update supervised methods is expensive. Thus, there is a need to study opinion unaware quality methods in parallel for better scalability and easier model updates.

Perhaps, the most successful methods for unsupervised NR IQA such as NIQE [22] and IL-NIQE [35] are based on natural scene statistics (NSS). While NSS based features have been successful for several distortions and they capture important aspects of quality, such approaches have not yet achieved satisfactory performance on authentically distorted images [10, 34, 6, 30]. Surprisingly, deep features trained for image classification have been shown to contain quality relevant information and can be trained to predict perceptual quality. Nevertheless, it is important to explore how deep networks can be trained to learn features that capture distortions in authentically distorted images more explicitly and without human supervision. The goal of our work is to explore the quality feature learning for authentically distorted images and use them to predict quality without the need of human label supervision in any step.

In this work, we consider the unsupervised NR IQA problem through self-supervised feature learning. Thus no human labels are involved in any step. This is an important approach and a path towards attaining robust no reference

*Equal contribution
†Work done by the author while at the Indian Institute of Science
[§]https://github.com/nithincbabu7/iqa-ContentSep

IQA performance on various datasets. Recently, there has been some work on self-supervised quality feature learning for synthetic and authentic distortions [19, 3]. However, the feature learning on authentically distorted images can be further improved. Self-supervised quality feature learning on authentically distorted images is challenging because multiple aspects such as content and quality can change among different authentically distorted images. For example, in contrastive learning, the identification of negatives that vary only in quality is challenging on authentically distorted images. Thus, the self-supervised learning of quality features on authentically distorted images is non-trivial.

We adopt a two-stage approach of feature learning to enrich features pre-trained on synthetic datasets using authentically distorted images. When we take patches from different authentically distorted images for self-supervised learning, there can be multiple variations such as those in quality and content. To more accurately learn quality related features from image patches, we adopt a contrastive learning method that not only discriminates image patches, but also minimizes an estimate of the dependence of features on content related variations through a mutual information bound.

While the dependence of features on content for IQA has been exploited for improving their performance in supervised NR IQA [27], we believe that such dependence can lead to a performance loss in unsupervised NR IQA approaches that compare features with a corpus of pristine images [22]. In supervised NR IQA, the availability of large annotated datasets implicitly ensures effective use of the content information to predict visual quality. On the other hand, it is not clear how the content information can be related to quality without supervision. Thus, there is a need to explicitly mitigate the content dependence of quality features in unsupervised NR IQA of authentically distorted images.

To summarize, our main contributions are as follows:

- We present a two-stage self-supervised feature learning approach with different learning methodologies on synthetic and authentically distorted images.
- While learning features on authentically distorted images, we only consider positives due to the difficulty in obtaining negatives that vary in quality alone.
- We introduce a mutual information based loss function while learning on authentic distortions to mitigate the dependence of features on content and enrich the learning of quality representations.
- We introduce a contrastive likelihood loss to optimize the variational approximation computed while estimating the bound on mutual information.
- We show that our self-supervised features can be used to make perceptually consistent image quality predictions without training on any human opinion scores.

2. Related Work

Supervised NR IQA: One of the most successful approaches to supervised NR IQA is based on NSS features [21, 25, 23, 32] and modeling of the human visual system [9]. BRISQUE [21], BLIINDS [25] and DIIVINE [23] represent a few popular examples of methods inspired from such an approach. While NSS based methods capture several synthetic distortions, their performance has suffered on authentically distorted camera captured images. With the emergence of deep learning, several researchers have studied end-to-end trained [2, 17, 12] and pre-trained deep networks for NR IQA with some modifications [27, 36, 7, 37, 34, 13]. The latter approach has been reasonably successful for authentically distorted images. Hyper IQA [27] adopts a hyper network to model the image semantics for NR IQA while DB-CNN [36] presents a two-stream approach to capture both synthetic and authentic distortions. The role of transformers to process the pretrained deep features for NR IQA has also been explored [7]. MetaIQA [37] explores the meta learning on synthetic distortions to quickly adapt the quality model to authentic distortions.

Weakly Supervised NR IQA: Another class of NR IQA methods such as [18, 16, 15] rely on weak supervision by making use of existing Full Reference (FR) and NR IQA metrics. Ma *et al.* [18] use multiple metrics and assign an associated reliability for each of the annotators. A CNN is finally trained to estimate quality by optimizing for consistency with the annotators. DipIQ [16] first generates a large number of quality discriminable image pairs using FR measures and then uses a pairwise learning algorithm in tandem with perceptual uncertainty levels to learn an opinion unaware IQA metric. RankIQA [15] trains a Siamese network to rank images between which the relative quality is known based on relative distortion levels. However the above methods cannot be used in the context of authentically distorted images where neither a reference is available nor the distortion levels are known.

Self-supervised/Unsupervised feature learning for NR IQA: One of the earliest approaches for learning quality features without human labels was designed in CORNIA [33]. A dictionary learning approach was adopted to learn quality aware features. More recently, self-supervised feature learning methods have been explored for NR IQA. CONTRIQUE [19] learns image features by predicting the distortion types and levels as a pretext task while performing instance discrimination on authentically distorted images. However, since both content and quality can change while discriminating instances, the learning of quality features can get impacted. The same approach is also investigated on synthetically generated images through the dead leaves model [20]. SPIQ [3] adopts a patch prediction framework to learn contrastive features on synthetically distorted images. In this method, the inefficiency in patch prediction can impact the feature learning process.

Unsupervised NR IQA: The NIQE [22] and IL-NIQE [35] formulations represent examples of methods that compute a distance between NSS features to a corpus of pristine natural image patches as a quality index. This represents an unsupervised approach for NR IQA without having to train the quality features on human scores. While NIQE works with NSS features, IL-NIQE enriches the features by adding other quality-aware features such as gradient features, log-Gabor filter responses and color statistics. The goal of our work is to show that by learning a richer set of features, one can use this unsupervised approach to predict perceptual quality without the need for human labels in any step.

3. Method

We propose a two step approach for unsupervised NR-IQA of authentically distorted images. We first learn quality features on a large corpus of synthetically distorted images and then fine tune the learning on authentic distortions. CONTRIQUE [19] jointly learns features on synthetic distortions using distortion labels and uses authentic distortions by deploying an instance discrimination framework [31]. However, such a method does not clearly address the learning of quality aware features on authentically distorted images. For example, if every sample in the data is assigned as a different class during instance discrimination, a heavy content bias can overpower the learning procedure and mitigate the learning of quality aware features. Therefore, we pre-train features on synthetically distorted images using M-SCQALE [11] and introduce a novel method to fine-tune these features and mitigate content bias on authentically distorted images. Our pre-training framework is chosen to be consistent with our fine-tuning framework.

3.1. Synthetic Data Pretraining

We provide an overview of M-SCQALE [11] used for pre-training before describing our contributions in the next subsection. M-SCQALE is a multi-view contrastive learning framework for IQA where the goal is to learn features that discriminate positive and negative pairs of views. In particular, a positive pair of views is chosen as a pair of large patches from the same image to capture the global image quality features while a negative pair is chosen from different distorted versions of the same image. While M-SCQALE was designed for low light image QA, we pretrain using this framework for several synthetic distortions as described in Section 4.2.1. Further, we only pre-trained a single scale and did not observe much improvement using multiple scales in M-SCQALE when integrated with our contributions in authentic fine-tuning. We observe that M-SCQALE requires both positives and negatives for learning, which are hard to design for authentically distorted images.

3.2. Authentic Fine-Tuning with Content Separation

Overview: As discussed in Section 1 and 3.1, there is a challenge in determining the positives and negatives for contrastive learning on authentically distorted images. This motivates us to explore contrastive learning on authentically distorted images without using negatives to fine-tune the quality representations. Several recent contrastive learning methods such as BYOL [8] and SimSiam [4] show excellent performance without the need of negatives altogether. BYOL uses two feature encoders such that the weights of the alternate encoder are updated as a momentum based moving average of the main feature encoder. The work by Chen *et al.* [4] does away with the need of even having an alternate momentum update based encoder and uses the idea of stop gradients to prevent collapsing solutions. We employ the SimSiam [4] framework to enable quality aware feature learning without using negatives by drawing patches from the same image as positives.

However patches taken from the same image can bias the model towards learning the correlation between the content and disturb the quality awareness of the pre-trained features. The quality features are ideally not supposed to be sensitive to the content of the image but only to the distortions. While the role of content dependence has been explored in supervised NR IQA [27, 7], we believe that content dependence can impact performance in unsupervised NR IQA approaches based on computing distances to a corpus of pristine images [22]. This motivates us to disentangle the content information from the learned features to mitigate the impact of content bias. We achieve this by minimizing a bound on the mutual information between the learnt features and the image content. We describe our entire framework as follows and in Figure 1.

Fine-Tuning Setup: We sample N images in a minibatch denoted by $\{I_1, I_2, \cdots, I_N\}$. Each sample is randomly divided into either vertical or horizontal halves and the largest square patch from each half is chosen. For any image I_n, $n \in \{1, 2, \ldots, N\}$, let $C_1^n(.)$ and $C_2^n(.)$ be the functions which crop large non-overlapping patches from I_n and resize them to a size of $M \times M$. Let $x_1^n = C_1^n(I_n)$ and $x_2^n = C_2^n(I_n)$ denote the two augmented views/positives drawn from a sample I_n. $f(.)$ here denotes the feature encoder with weights initialized from the synthetically pretrained network. We denote the prediction MLP head by $h(.)$. Let $z_1^{(n)} = f(x_1^n)$, $z_2^{(n)} = f(x_2^n)$, $p_1^{(n)} = h(f(x_1^n))$ and $p_2^{(n)} = h(f(x_2^n))$. The loss that is used to update the network weights is given by [4]

$$\mathcal{L}_c = \sum_{n=1}^{N} \frac{D(p_1^{(n)}, \text{sg}(z_2^{(n)}))}{2} + \frac{D(p_2^{(n)}, \text{sg}(z_1^{(n)}))}{2}. \quad (1)$$

The function $D(.)$ denotes the negative of normalized co-

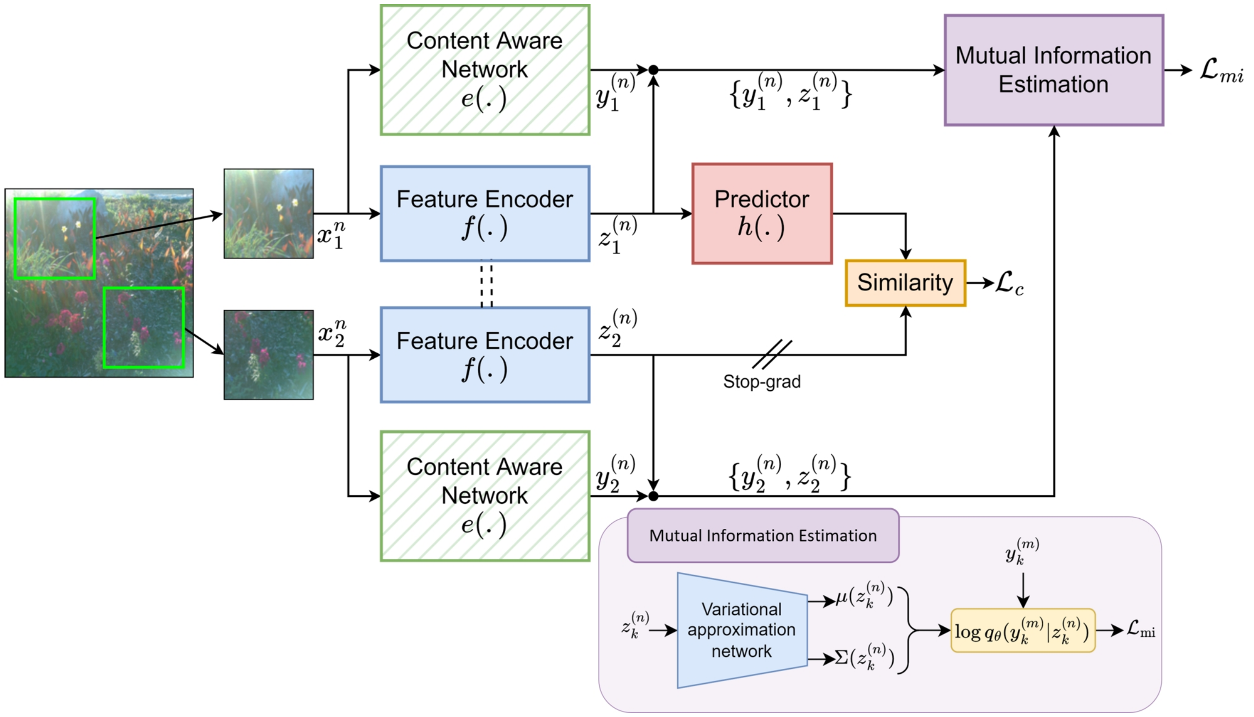


Figure 1: Block diagram describing our framework for authentic fine-tuning with content separation.

sine similarity. The sg(.) term indicates a stop gradient operation, which ensures that the gradient from $z_2^{(n)}$ does not flow back to the feature encoder on x_2^n to prevent collapsing solutions [4]. The gradients from $p_2^{(n)}$ flow back to the feature encoder through the second term in the loss function. It acts in a vice versa fashion for x_1^n.

Mitigation of Content Dependence: We introduce another loss term to mitigate the dependence of the learnt features on content information. Let $e(.)$ denote a CNN which extracts content aware information from an image. Let the content aware features be denoted as $y_1^{(n)} = e(x_1^n)$ and $y_2^{(n)} = e(x_2^n)$. To disentangle content from our feature learning we try to minimize the mutual information $I(Y_k; Z_k)$ for $k = \{1, 2\}$ where Y_k and Z_k are the random variables with $y_k^{(n)}$ and $z_k^{(n)}$ as the corresponding samples.

Contrastive Log-ratio Upper Bound (CLUB): To minimize mutual information, we adopt the Contrastive Log-ratio Upper Bound (CLUB) on mutual information [5], estimated as

$$\mathcal{L}_{mi_k} = \frac{1}{N}\sum_{n=1}^{N}\Bigg[\log q_\theta\left(y_k^{(n)}|z_k^{(n)}\right) - \frac{1}{N}\sum_{m=1}^{N}\log q_\theta\left(y_k^{(m)}|z_k^{(n)}\right)\Bigg]. \tag{2}$$

The above equation uses a variational distribution $q_\theta(Y_k|Z_k)$ that approximates $p(Y_k|Z_k)$. The distribution $q_\theta(Y_k|Z_k)$ is modeled as a neural network parametrized by θ whose output describes the parameters of the conditional distribution. Further, the conditional distribution is modeled as an independent multivariate Gaussian, whose parameters need to be predicted. The neural network takes $z_k^{(n)}$ as input and predicts the mean $\mu(z_k^{(n)})$ and variance $\sigma^2(z_k^{(n)})$ of Y_k given $Z_k = z_k^{(n)}$. According to [5], the neural network parameters should be updated such that $q_\theta(Y_k, Z_k)$ is similar to the intractable joint distribution, $p(Y_k, Z_k)$, than to the product of the marginals, $p(Y_k)p(Z_k)$. The authors facilitate this by minimizing $\text{KL}(p(Y_k, Z_k)\|q_\theta(Y_k, Z_k))$ over θ, which is the same as maximizing the log-likelihood of $q_\theta(Y_k|Z_k)$, implemented as maximizing

$$\mathcal{L}_{\theta_k} = \frac{1}{N}\sum_{n=1}^{N}\log q_\theta(y_k^{(n)}|z_k^{(n)}). \tag{3}$$

Contrastive Likelihood Loss for Variational Approximation: For the parameters θ of the variational approximation network to be effective in minimizing the mutual information, we need $q_\theta(Y_k, Z_k)$ to be similar to the joint distribution $p(Y_k, Z_k)$ than to the product of marginals $p(Y_k)p(Z_k)$. The minimization of $\text{KL}(p(Y_k, Z_k)\|q_\theta(Y_k, Z_k))$ ensures that the distributions $p(Y_k, Z_k)$ and $q_\theta(Y_k, Z_k)$ are similar, but need not guarantee that $p(Y_k)p(Z_k)$ and $q_\theta(Y_k, Z_k)$ are dissimilar. Therefore, we propose to minimize $\mathrm{KL}(p(Y_k, Z_k) \| q_\theta(Y_k, Z_k))$ and maximize $\mathrm{KL}(p(Y_k)p(Z_k) \| q_\theta(Y_k, Z_k))$ through the following optimization

$$\begin{aligned}\min_{\theta} \, [&\mathrm{KL}(p(Y_k, Z_k) \| q_\theta(Y_k, Z_k)) \\ &-\mathrm{KL}(p(Y_k)p(Z_k) \| q_\theta(Y_k, Z_k))] \, .\end{aligned} \tag{4}$$

Following [5], the above optimization problem can be implemented as a maximization of the following loss function over θ,

$$\begin{aligned}\mathcal{L}_{\theta_k} = \frac{1}{N} \sum_{n=1}^{N} \Bigg[&\log q_\theta(y_k^{(n)} | z_k^{(n)}) \\ &- \frac{1}{N} \sum_{m=1}^{N} \log q_\theta(y_k^{(m)} | z_k^{(n)}) \Bigg] .\end{aligned} \tag{5}$$

The steps for obtaining Equation (5) from (4) are explained in detail in the Supplementary Material.

The loss functions used in the equations (2) and (5) are exactly the same. In one case, the loss is minimized by updating the feature encoder parameters and in the other case the loss is maximized by updating the variational approximation network. The final content dependence loss $\mathcal{L}_{mi}$ is the average of $\mathcal{L}_{mi_k}$ over $k = \{1, 2\}$. Similarly, the variational approximation network is updated using $\mathcal{L}_\theta$, which is again an average of $\mathcal{L}_{\theta_k}$ over k. The overall multi-task loss for updating the feature encoder $f(.)$ and the predictor $h(.)$ is

$$\mathcal{L} = \mathcal{L}_c + \lambda_{mi} \mathcal{L}_{mi}, \tag{6}$$

where λ_{mi} is a hyperparameter used to scale $\mathcal{L}_{mi}$. We employ warm-start [1] on $\mathcal{L}_{mi}$ by adding it to the contrastive loss function $\mathcal{L}_c$ after some iterations of training the feature encoder with the contrastive loss. The variational approximation network is updated alternately with the feature encoder updates at each iteration of the learning process.

3.3. Quality Prediction

Similar to [11], we replace the NSS features with our deep features in the popular completely blind quality prediction framework NIQE [22]. Let (μ_r, Σ_r) correspond to the Gaussian model parameters learnt on the quality features of a set of sharp and colorful pristine patches and (μ_d, Σ_d) correspond to that of the input image patches. The quality score for each test image is predicted using

$$Q = \sqrt{\left((\mu_r - \mu_d)^T \left(\frac{\Sigma_r + \Sigma_d}{2} \right)^{-1} (\mu_r - \mu_d) \right)}. \tag{7}$$

We use the method described above for evaluating and comparing different features learnt using various feature learning methods.

4. Experiments

4.1. Databases

We evaluate the performance of different unsupervised NR IQA methods on four authentically distorted datasets namely CLIVE [6], KONIQ [10], FLIVE [34] and CID [30]. CLIVE [6] contains a total of 1,162 images captured using multiple mobile devices. These images contain a diverse mix of distortions such as noise, blur, under-exposure, over-exposure etc. KONIQ [10] contains a total of 10,073 images with various distortions such as noise, JPEG compression artifacts, motion blur, over-saturation etc. These images were sampled from the YCC100M [28] dataset. CID [30] contains 473 images, with camera captured distortions such as blur, noise, under-enhancement, over-enhancement etc. FLIVE [34] contains a total of 40,000 images and 120,000 patches. We use only the 40,000 images from FLIVE and not the patches for all our experiments. Images with different sizes and aspect ratios with a mix of real world distortions make it a challenging dataset for IQA.

4.2. Implementation Details

4.2.1 Synthetic Pretraining

We use 840,000 images from the KADIS [14] dataset for the synthetic pretraining stage. We use 4 distorted versions per scene and 16 number of scenes per mini-batch. We train for 5 epochs using Adam optimizer and a learning rate of 0.01. We set temperature parameter as $\tau = 0.1$. We use ResNet-50 as our feature encoder $f(.)$.

4.2.2 Authentic Fine-Tuning

We use 10,000 randomly sampled images from the AVA [24] dataset such that there is no overlap with the images used for testing. During this fine-tuning stage, we only fine-tune the first convolutional layer and last bottleneck layer of $f(.)$. The prediction MLP head $h(.)$ takes an input of size 2048 from global average pooling output of $f(.)$. $h(.)$ has a hidden layer of size 512 with batch normalization applied to it and outputs a feature of size 2048.

We use a ResNet-50 pre-trained on ImageNet as our content-aware network $e(.)$. We tap the final 1000-dimensional softmax output as the content feature y. We do not update the parameters of the network $e(.)$. The variational approximation network approximates the conditional distribution of content feature y given the encoded feature output z. The network takes z as input through an Exponential Linear Unit (ELU) activation function with its α parameter set to 1.0. The ELU output is passed through two parallel MLP blocks, each with a hidden layer of size 1000. One of the MLP blocks outputs the mean, and the other outputs the logarithm of variance of the content features y. The

Dataset	CLIVE [6]		KONIQ [10]		FLIVE [34]		CID [30]	
Method	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE [22]	0.46	0.48	0.53	0.54	0.21	0.29	0.23	0.22
IL-NIQE [35]	0.44	0.49	0.51	0.53	0.22	0.27	0.31	0.40
CORNIA* [33]	0.07	0.07	0.04	0.02	0.05	0.13	0.27	0.29
CONTRIQUE* [19]	0.38	0.42	0.63	0.61	0.26	0.29	**0.74**	**0.76**
Proposed	**0.51**	**0.52**	**0.65**	**0.64**	**0.30**	**0.33**	0.64	0.66

Table 1: Performance analysis. * denotes opinion unaware version as explained in Section 4.3 for fair comparison.

approximations of the mean and variance obtained from the MLP blocks are used to obtain the conditional distribution, which is further used in the loss functions $\mathcal{L}_{mi}$ and $\mathcal{L}_{\theta}$.

We use a batch size of $N = 16$ for updating the parameters of $f(.)$, $h(.)$ and the variational approximation network. We use learning rates of 0.001 and 0.0001 for $f(.)$ and $h(.)$ respectively. We choose $\lambda_{mi} = 1000$ in Equation (6) and set the learning rate for the variational approximation network to be 10^{-7}. We use a patch resize of $M = 256$ for stability during training as authentic datasets have images with varied resolutions. We fine-tune on the sampled images from the AVA [24] dataset for 5 epochs with the $\mathcal{L}_{mi}$ term being added only after the first epoch. We use an RTX 2080 Ti GPU and PyTorch framework for all our experiments.

4.2.3 Quality Prediction

We use the same set of 125 images as used by NIQE [22] to learn the pristine MVG model in Equation (7). We use patches of size $R = 96$ similar to NIQE [22]. As FLIVE [34] has images with multiple resolutions, we resize each image to 512×512 before employing the quality prediction step.

4.3. Performance Comparisons and Analysis

We compare against popular unsupervised (opinion unaware) NR-IQA methods such as NIQE [22] and IL-NIQE [35]. We also compare with the unsupervised feature learning method CORNIA [33] and the self-supervised feature learning method CONTRIQUE [19]. Both these methods are used to extract features and these features are used in the prediction framework described in Section 3.3 for a fair comparison in an unsupervised setting. We did not compare with SPIQ [3], as the code and the pre-trained model are unavailable and several implementation details are not provided. The same set of 10,000 images which are used in our authentic fine-tuning stage are used to build the dictionary for CORNIA [33]. We use the pretrained model made available by the authors of CONTRIQUE [19] to evaluate the CONTRIQUE features. Since our goal is opinion unaware NR IQA for authentically distorted images, we do not compare with other supervised NR IQA methods.

We employ the Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) to evaluate the performance of different methods. The SRCC and PLCC are computed between the set of ground truth quality scores and the predictions for all images in each dataset as none of the compared methods require training on quality labels. Before computing the PLCC, the predicted scores are passed through a non-linearity [26].

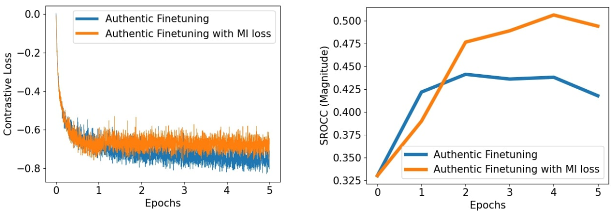
We observe from Table 1 that our method outperforms other comparable methods on CLIVE [6] and FLIVE [34] with significant margins. We also see improvements on the KONIQ [10] dataset. We observe that CONTRIQUE [19] performs better than our method on CID [30]. We also observe that CORNIA [33] performs quite poorly on CLIVE [6], KONIQ [10] and FLIVE [34] but performs better than NIQE [22] on the CID [30] dataset. We note that the structure of the CID [30] dataset is very different as compared to the other three, since there are only around eight unique scenes. Although there are several distorted versions, the variety of scenes is limited contributing to a reduced diversity in this dataset. The performance on FLIVE [34] is typically lower when compared with other datasets, which can be expected as it is a very challenging dataset even for the supervised methods. Nevertheless, we achieve the best results among all other unsupervised NR methods.

4.4. Ablations

Strength of different components: We evaluate the strength of each of our proposed components in terms of the SRCC performance on authentically distorted datasets. In particular, we evaluate the performance of the pretrained features on the synthetic dataset, and the impact of fine-tuning on the authentic dataset with and without the mutual information based cost function. We see from Table 2 that each component has its own merit. Adding the fine-tuning on authentic datasets pushes up the performance on CLIVE [6], FLIVE [34] and KONIQ [10]. The addition of the content dependence loss term in the authentic fine-tuning part

Synthetic	Authentic	MI	CLIVE [6]	KONIQ [10]	FLIVE [34]	CID [30]
✓	✗	✗	0.33	0.60	0.18	0.62
✓	✓	✗	0.44	0.63	0.29	0.61
✓	✓	✓	**0.51**	**0.65**	**0.30**	**0.64**

Table 2: Strength of adding different components in terms of SRCC Performance.



(a) Contrastive Loss

(b) Test set SRCC

Figure 2: A comparison of training and test performance curves when fine-tuning with and without MI loss term. The contrastive loss term and the test set performance increase after introducing the MI loss term at the first epoch.

improves the performance further on all the datasets.

Content Dependence Loss based Regularization: We qualitatively analyze the effect of adding the content dependence loss through the Mutual Information (MI) term in Equation (2) during the authentic fine-tuning procedure in Figure 2. We obtain the curve of the contrastive loss term $\mathcal{L}_c$ at each iteration, and the SRCC performance on CLIVE [6] dataset over each epoch of training. Introducing the MI loss in the training procedure after the first epoch affects the contrastive loss curve as we can see an increase in the loss after the first epoch in Figure 2a. Further, Figure 2b shows an improvement in the test set accuracy over the course of training when using the content dependence loss. The improvement in the SRCC curve and the deterioration in the contrastive loss curve are indicators of regularization and improved generalization performance of the model after adding the content dependence loss.

Analysis of Features for Different Configurations: We analyze the t-Distributed Stochastic Neighbor Embedding (t-SNE) [29] plot of the feature outputs from three different configurations of our algorithm in Figure 3. We do this for the synthetically pretrained model and the fine-tuned model with and without the content dependence loss term. We take all samples from CLIVE [6] dataset for this analysis. We assign each image to five coarse quality bins which are color coded. We observe that as compared to the synthetically pretrained model, the vanilla fine-tuned model without content dependence loss gives larger separation between the samples of the dataset. The fine-tuned model with content dependence loss is able to separate the different quality bins even better thereby enabling better quality prediction.

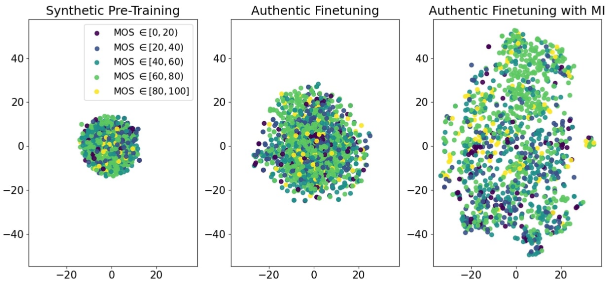


Figure 3: t-SNE plot showing the spread of feature predictions when different components of our method are used. We observe a larger separation between quality bins as each component is added.

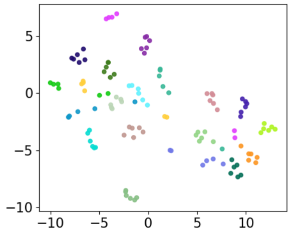


Figure 4: t-SNE plot of the content aware features. Same color implies same content with different distortions. Samples with same color tend to cluster together.

Analysis of Content Features: We evaluate the effectiveness of the content-aware features from $e(.)$ in capturing the content relevant information. We qualitatively evaluate the effectiveness of $e(.)$ by conducting a t-SNE analysis on the content features of a synthetic dataset. We randomly choose 20 reference images from the KADIS [14] dataset and analyze the t-SNE plots of the content aware features of the reference images and their distorted versions. Figure 4 shows a plot of content features reduced to two dimensions using t-SNE. Each color corresponds to a reference image. We observe different distorted versions from the same reference clustering together, showing us that the content feature captures rich content-relevant information. However that being the case, this also points out that the content features are not very sensitive to the distortions or quality. Thus, learning quality features to share minimal information with these content features helps mitigate the content bias in the quality features.

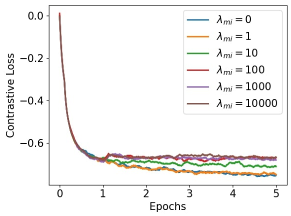


Figure 5: Impact of choosing different values of λ_{mi} in the contrastive loss curve.

λ_{mi}	SRCC
0	0.42
1	0.41
10	0.47
100	0.47
1000	0.49
10000	0.47

Table 3: SRCC performance variation on CLIVE with respect to λ_{mi}

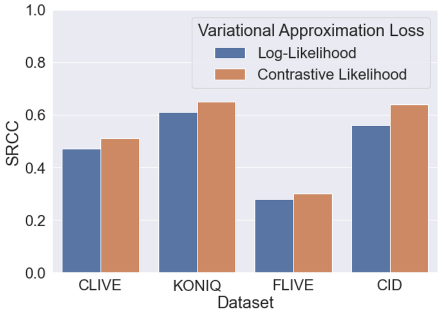


Figure 6: The SRCC performance when using log likelihood vs contrastive likelihood formulations in the variational approximation network

Impact of varying hyperparameters: We show experiments with respect to changing the scaling value (λ_{mi}) of the MI loss. We plot the variation of $\mathcal{L}_c$ loss curve for different values of λ_{mi} in Figure 5. However, we plot a smoothed version of the contrastive loss for a better visualization of the variations. Table 3 shows the variation in SRCC performance on CLIVE[6] with respect to the variation in λ_{mi}. We report the performance at the end of five epochs for each of the runs for fair comparison. We observe that as λ_{mi} increases, the jump in $\mathcal{L}_c$ that corresponds to the regularization becomes larger. As expected we find that increasing λ_{mi} up to a certain value improves SRCC performance. However once the regularization becomes too aggressive (at $\lambda_{mi} = 10000$) we find that there is drop in the SRCC performance.

Impact of Contrastive Likelihood in Variational Approximation: We also compare the performance of our method when using the contrastive likelihood loss in Equation (5) for training the variational approximation network, as compared to using a log-likelihood based loss in Equation (3). We see from Figure 6 that we get consistent improvement in the performance when using the contrastive likelihood loss over the likelihood loss.

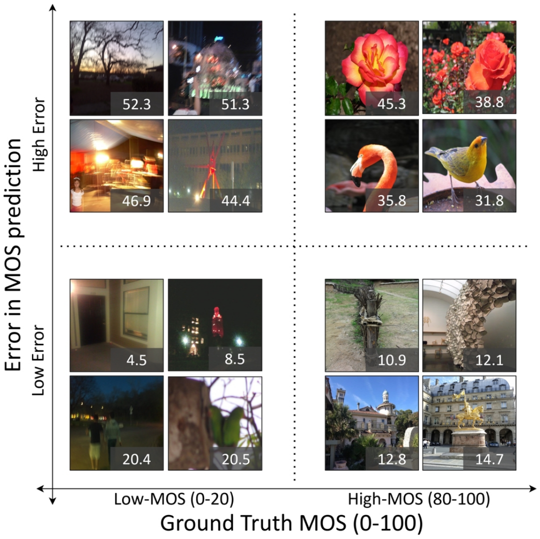


Figure 7: Qualitative analysis of the predicted scores. The MOS prediction error is indicated on each image.

4.5. Qualitative Analysis

We now present a qualitative analysis of how our quality prediction model performs. We compute the absolute error in ground truth quality (or mean opinion score (MOS)) prediction by computing a non-linear fit between our quality prediction using Equation (7) and MOS on the CLIVE [6] dataset. Figure 7 shows the images corresponding to different relationships between the error and MOS. For low-MOS images, we observe that the model can predict the quality of dark images well while the error is higher for images with a combination of distortions, such as low light and motion blur. Most of the images where our model fails for high-MOS images seem to have some aesthetic attribute while it performs well on images that do not have such attributes.

5. Conclusion

Our two-stage self-supervised feature learning presents a novel framework for unsupervised NR-IQA of authentically distorted images with state-of-the-art performance on multiple authentically distorted datasets. We infer that the mitigation of content bias in unsupervised feature learning of quality aware features has an integral part to play. Our mutual information minimization based formulation and contrastive likelihood based optimization effectively address this content bias and give significant improvements. This is particularly important in the context of unsupervised NR IQA where it is not clear how to use the content based features without labels for supervision.

Acknowledgement: This work was supported in part by a grant from the Department of Science and Technology, Government of India, under grant CRG/2020/003516.

References

[1] Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in Neural Information Processing Systems*, 33:3884–3894, 2020.

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

[3] 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.

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[5] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[6] Deepti Ghadiyaram and Alan Conrad Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25:372–387, 2016.

[7] 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 3209–3218, 2022.

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[9] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, 2014.

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

[11] Vignesh Kannan, Sameer Malik, and Rajiv Soundararajan. Quality assessment of low light restored images: A subjective study and an unsupervised model. *arXiv preprint arXiv:2202.02277*, 2022.

[12] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016.

[13] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C. Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017.

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

[15] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1040–1049, 2017.

[16] 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.

[17] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[18] 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, 2019.

[19] 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.

[20] 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.

[21] Anish Mittal, Anush K. Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21:4695–4708, 2012.

[22] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a ”completely blind” image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

[23] 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.

[24] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[25] 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.

[26] 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.

[27] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3664–3673, 2020.

[28] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth,

and Li-Jia Li. Yfcc100m. *Communications of the ACM*, 59(2):64–73, Jan 2016.

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[30] Toni Virtanen, Mikko Nuutinen, Mikko Vaahteranoksa, Pirkko Oittinen, and Jukka Häkkinen. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 24(1):390–402, 2015.

[31] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] 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.

[33] 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, 2012.

[34] 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.

[35] 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.

[36] 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, 2020.

[37] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14143–14152, Jun. 2020.