Quality-aware Pre-trained Models for Blind Image Quality Assessment

Kai Zhao†, Kun Yuan†, Ming Sun, Mading Li and Xing Wen
Kuaishou Technology
{zhaokai05,yuankun03,sunming03,limading,wenxing}@kuaishou.com

Abstract

Blind image quality assessment (BIQA) aims to automatically evaluate the perceived quality of a single image, whose performance has been improved by deep learning-based methods in recent years. However, the paucity of labeled data somewhat restrains deep learning-based BIQA methods from unleashing their full potential. In this paper, we propose to solve the problem by a pretext task customized for BIQA in a self-supervised learning manner, which enables learning representations from orders of magnitude more data. To constrain the learning process, we propose a quality-aware contrastive loss based on a simple assumption: the quality of patches from a distorted image should be similar, but vary from patches from the same image with different degradations and patches from different images. Further, we improve the existing degradation process and form a degradation space with the size of roughly $2 \times 10^7$. After pre-trained on ImageNet using our method, models are more sensitive to image quality and perform significantly better on downstream BIQA tasks. Experimental results show that our method obtains remarkable improvements on popular BIQA datasets.

1. Introduction

With the arrival of the mobile internet era, billions of images are generated, uploaded and shared on various social media platforms, including Twitter, TikTok, etc [30]. As an essential indicator, image quality can help these service providers filter and deliver high-quality images to users, thereby improving Quality of Experience. Therefore, huge efforts [12,20,22,24,52,73] have been devoted to establishing an image quality assessment (IQA) method consistent with human viewers. In real-world scenarios, there usually exists no access to the reference images and the quality of reference images is suspicious. Thus, blind IQA (BIQA) methods are more attractive and applicable, despite full-reference IQA has achieved prospective results [33].

†Equal contribution.
propose to train a model on synthetic images with artificial degradation and then regress the model onto small-scale target BIQA datasets [43, 73, 89]. However, images generated by rather simple degradation process with limited distortion types/levels are far from authentic. Further, synthetic images are regularly distorted from limited pristine images of high quality, so the image content itself only has a marginal effect on the image quality. Yet in real-world scenarios, the image quality is closely related to its content, due to viewers’ preferences for divergent contents [38, 62].

Self-supervised learning (SSL) or unsupervised learning is another potential choice to overcome the problem of lacking adequate training data, for its ability to utilize an amount of unlabeled data. Such technique is proven to be effective in many common computer vision tasks [2]. However, different from models for these tasks which mainly focus on high-level information, representations learned for BIQA should be sensitive to all kinds of low-level distortions and high-level contents, as well as interactions between them. There are few research focused on such area in the literature, let alone a deep SSL designed for BIQA with state-of-the-art performance [44].

In this work, we propose a novel SSL mechanism that distinguishes between samples with different perceptual qualities, generating Quality-aware Pre-Trained (QPT) models for downstream BIQA tasks. Specifically, we suppose the quality of patches from a distorted image should be similar, but vary from patches from different images (i.e., content-based negative) and the same image with different degradations (i.e., degradation-based negative). Moreover, inspired by recent progress in image restoration, we introduce shuffle order [84], high-order [70] and a skip operation to the image degradation process, to simulate real-world distortions. In this way, models pre-trained on ImageNet are expected to extract quality-aware features, and boost downstream BIQA performances. We summarize the contributions of this work as follows:

- We design a more complex degradation process suitable for BIQA. It not only considers a mixture of multiple degradation types, but also incorporates with shuffle order, high-order and a skip operation, resulting in a much larger degradation space. For a dataset with a size of $10^7$, our method can generate more than $2 \times 10^{14}$ possible pairs for contrastive learning.

- To fully exploit the abundant information hidden beneath such an amount of data, we propose a novel SSL framework to generate QPT models for BIQA based on MoCoV2 [8]. By carefully designing positive/negative samples and customizing quality-aware contrastive loss, our approach enables models to learn quality-aware information rather than regular semantic-aware representation from massive unlabeled images.

- Extensive experiments on five BIQA benchmark datasets (sharing the same pre-trained weight of QPT) demonstrate that our proposed method significantly outperforms other counterparts, which indicates the effectiveness and generalization ability of QPT. It is also worth noting that the proposed method can be easily integrated with current SOTA methods by replacing their pre-trained weights.

2. Related Work

In this section, we provide a brief review of major works on BIQA and recent advances in contrastive learning and image degradation modeling.

2.1. Blind Image Quality Assessment

Before the rise of deep learning, natural scene statistics (NSS) theory dominates the realm of BIQA, which assumes that pristine natural images obey certain statistics distribution and various distortions will break such statistical regularity [59]. Based on this theory, various hand-crafted features are proposed in different domains, including spatial [47, 48], frequency [49, 54] and gradient [85]. Meanwhile, some learning-based methods [46, 78] have also been preliminarily explored. Typically, they estimate the subjective quality using support vector regression and features learned by visual codebooks [77].

In recent years, a variety of deep learning-based methods have been studied for BIQA and significantly boost the in-the-wild performance [21, 28, 38, 53, 67]. As a pioneer work, a shallow network [31] consisting of only three layers is built to solve BIQA in an end-to-end manner. After that, later works naturally expand the BIQA model by deepening the network depth [4, 42] or utilizing more effective building blocks [60, 91]. Recently, transformer-based BIQA methods [32, 76, 80, 92] are sprouting up, based on the assumption that transformer can compensate the missing ability of CNNs to capture non-local information [16].

Except for elaborate model design, a few research are devoted to solving the chief obstacle in BIQA, the paucity of labeled data. Some of them try to make full use of existing supervisory signals, such as rank learning [37, 39], multi-task learning [17] and mixed-dataset training [62, 90]. Meanwhile, others turn to large-scale pre-training [32, 42, 82], which generates a large number of distorted images labeled by specifications of degradation process [89] or quality scores estimated from FR models [43].

In comparison, our proposed method is based on contrastive learning without needing any explicit supervisory signal, thereby fully utilizing plenty of real-world images.

2.2. Self-supervised Learning

SSL, as a form of unsupervised learning, is used to learn a good data representation for the downstream true purpose...
A very straightforward idea to conduct such learning is minimizing the difference between a model’s output and a fixed target, such as reconstructing the input pixels [69] or predicting the pre-defined categories [15, 86]. Motivated by the successes of masked language modeling in natural language processing, masked image modeling [26, 75] is becoming a hot trend in computer vision.

Unlike the above methods, the goal of contrastive learning is to learn such an embedding space, where similar samples are assembled while dissimilar pairs are repelled [30]. Specifically, contrastive learning generally involves two aspects: pretext tasks and training objectives. Owing to the flexibility of contrastive learning, a wide range of pretext tasks have been proposed, such as colorization [86, 87], multi-view coding [64], and instance discrimination [27, 74]. When it comes to training objectives, the major trend is the move from only one positive and one negative sample [10, 55] to multiple positive and negative pairs [18, 25, 66]. Especially, InfoNCE [66] has become the most popular choice [5, 9, 27].

Notably, all of the above works can be classified as semantic-aware pre-training, because they encourage views (augmentations) of the same image to have a similar representation while ignoring the changes of perceived image quality. In this work, we redesign a quality-aware pretext task for BIQA, which will be discussed in Sec. 3.3.

2.3. Image Degradation Modelling

Most of the previous works focus on several classical distortion types [56, 57, 72, 83] to extract distortion-specific features. As the importance of BIQA is rising, more distortion types have been developed and applied to synthesize distorted images [40]. DipIQ [41] summarizes common distortion operations, such as noise, blur and compression, and further subdivides each with five levels. DB-CNN [89] introduces additional five distortion types. Recently, many works [21, 34, 91] point out that the above methods are too simple to simulate real-world images with complex and mixed distortions, due to the synthesized image only with a specific distortion type and level. In this work, we combine multiple degradation tricks [70, 84] proposed in the field of image restoration to form a much larger degradation space and generate distorted images that are more realistic.

3. Method

To excavate the potential of the “pre-training and fine-tuning” paradigm on BIQA for better performance, we present QPT models by pre-training on ImageNet. In Sec. 3.1, we will briefly review some related SSL frameworks. In Sec. 3.2, we will detail the degradation process for generating distorted images that contain quality-related information. In Sec. 3.3, we will give the definition of quality-aware pretext task and optimization objective.

3.1. Reviewing the Framework of SSL

SSL enables learning representations from orders of magnitude more data, which addresses the obstacle for BIQA in establishing large-scale annotated datasets. SSL methods can be roughly divided into generative-based [26, 68, 75] and contrastive-based [7, 23, 27]. Among them, MAE [26], as a generative-based method, develops an asymmetric encoder-decoder architecture for image reconstruction. MoCo [27], as a contrastive-based method, builds a dynamic dictionary with a queue and a moving-averaged encoder for contrastive learning. Relatively, contrastive learning is more suitable for BIQA, for which can easily measure quality ranking between samples. Due to the fact that the quality of an image is related to various complicated factors [65, 71], a large number of images are needed to ensure the diversity of negative samples [7]. Empirically, we choose MoCo, since it provides a memory-efficient framework to establish a large and consistent set.

In brief, MoCo builds a dictionary using a momentum queue to enlarge available size for comparison, where samples are progressively replaced with incoming mini-batch ones. And keys in the dictionary are updated through a momentum-based moving average of the encoder network, maintaining consistency with query samples. MoCo adopts the instance-level discrimination task [74] as its pretext task (i.e., a query matches a key if they are encoded views of the same image). Then, a contrastive loss, named InfoNCE, is adopted to distinguish the positive key from negative ones in a softmax manner, which can be noted as:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{N} \exp(q \cdot k_i / \tau)},$$

where $q$ is an encoded query, $k_i$ is the $i$-th encoded key of a dictionary consisting of $N$ samples, and $k_+$ is a single key that matches $q$. And $\tau$ is a temperature hyper-parameter.

3.2. Degradation Space

To obtain utilizable quality-related information in self-supervised scenarios, a simple way is to generate image pairs using controllable degradations manually. Before designing specific degradation types, several observations need to be mentioned: First, there exist many factors affecting the perceptual quality of an image (e.g., content, distortions, compression). For example, an image with meaningless content is often classified as a low-quality one [36]. Also, distortions (e.g., blurriness [13], noise [58] and sharpness [81]) introduced during the image production phase and compression artifacts introduced by transcoding or transmission, will degrade the overall quality. Second, the aforementioned factors, as a complicated composition, are always involved in practical scenarios [1, 58]. An image may go through various processes of shooting, editing,
compression, and transmission, resulting in more complicated distortions. These scenarios increase the difficulty of BIQA tasks. Thus a good pre-trained model for IQA should take these into account, covering diverse distortions using suitable degradations as large and realistic as possible.

Based on these observations, we design the degradation space from the perspective of individual operations and their compositions. For the first observation, we introduce a variety of degradation types to simulate real distortions, which can be divided into three categories: (1) geometric deformation, which simulates distortions introduced during editing process or adaptation on various display equipment, including 4 operations of scale jitter [32], horizontal flip, down-sampling and up-sampling [50]; (2) color change [63], which can be caused by lightness, chroma, and hue differences in the shooting or codec process, including 2 operations of color jitter and grayscale; (3) texture adjustment, which can be captured from environmental interference or transmission, including 3 operations of adding noise, fuzzification [19] and JPEG compression [11]. Given an image \( y \), the process can be formulated as:

\[
x = D(y; w, p),
\]

where \( D \) denotes the function of degradation. \( w \) represents hyper-parameters of the degradation (e.g., scale size, location, brightness intensity, etc.). And \( p \in \{0, 1\} \) indicates whether this operation is performed. For the second observation, we propose a sequence, consisting of randomly chosen degradations, for more complicated types (e.g., a list of \{scale jitter, color jitter, adding noise\}), where each operation can be skipped, and the order can be shuffled:

\[
x = D^o(y) = D_o(\cdots D_1(y; w_1, p_1) \cdots ; w_o, p_o),
\]

where \( o \) is the number of selected degradations. Furthermore, to simulate multiple transfers, we adopt a high-order degradation type, involving multiple repeated processes of \( D^o(x) \), where each process performs with the same procedure but different hyper-parameters. Empirically, we employ a degenerate approach with at most two orders, as it could resolve most real cases while keeping simplicity. The illustration of the overall process is shown in Fig. 2.

The degradation space formed by our operations is much larger than those in existing super-resolution (SR) methods [70,84], which reconstruct high-resolution images from low-resolution ones. Since BIQA does not require pixel-level aligned pairs like SR, more operations like geometric deformation/color change can be used. Moreover, SR requires that the source images should be high-quality and the distributions of target and source images should be as consistent as possible. These factors limit the number of available compositions. Without these restrictions, in our method, the size of degradation space can be expanded substantially. Theoretically, for a space containing 9 degradation types, the number of discrete compositions is \( 2 \times \sum_{i=1}^{9} C_0^i \times A_i \approx 2 \times 10^7 \).

### 3.3. Quality-aware Pretext Task

Different from the classical semantic-aware pretext task (i.e., treating each image instance as a distinct class) [3,74], we propose a new quality-aware pretext task that distinguishes between samples with different perceptual qualities.

Suppose we have \( N \) images \( y^{(1)} \cdots y^{(N)} \). As shown in Fig. 3, each image \( y^{(n)} \) will be differently augmented by \( K \) degradation compositions, obtaining different “views” \( x^{(n)}_1, \cdots, x^{(n)}_K \) that own consistent content but with different perceptual qualities. Then we extract a “query” (or “anchor”) and a “key” patch that originates from the same view \( x^{(n)}_k \) but with different locations, denoted \( x^{(n)}_{k,i} \) and \( x^{(n)}_{k,j} \) respectively. The query and key are considered as a positive pair due to their consistent content and degrada-
Figure 3. The framework of Quality-aware Pretext Task (QPT). First, images are augmented by diverse degradations. Second, patches will be extracted from various views to form patch pairs. Pairs generated from the same view but different locations are treated as positive pairs. And other pairs are noted as negative pairs, including degradation-based and content-based. Then, the patches are transformed into encoded features by a network. Last, these feature pairs are constrained using a Quality-aware Contrastive Loss (QC-Loss).

Table 1. Comparison of two types of pretext tasks, where $\tilde{k}$ represents a degradation/augmentation type different from $k$, $\tilde{n}$ is an image different from $n$, and $*$ denotes an arbitrary value.

<table>
<thead>
<tr>
<th>Pretext task</th>
<th>Positive Pair</th>
<th>Negative Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>semantic-aware</td>
<td>$(x_{k,i}^{(n)}, x_{k,j}^{(n)})$</td>
<td>$(x_{k,i}^{(n)}, x_{k,i}^{(\tilde{n})})$</td>
</tr>
<tr>
<td>quality-aware</td>
<td>$(x_{k,i}^{(n)}, x_{k,j}^{(n)})$</td>
<td>$(x_{k,i}^{(n)}, x_{k,i}^{(n)}), (x_{k,i}^{(n)}, x_{k,i}^{(\tilde{n})})$</td>
</tr>
</tbody>
</table>

4. Experiments

4.1. Datasets and Evaluation Criteria

Datasets. Our method is evaluated on five public BIQA datasets, including BID [12], CLIVE [20], KonIQ10K [29], SPAQ [17] and FLIVE [79]. BID is a blur image dataset that contains 586 images with realistic blur distortions (e.g., out-of-focus, simple motion, complex motion blur). CLIVE consists of 1,162 images with diverse authentic distortions captured by mobile devices. KonIQ10K contains 10,073

for computing the dot-product similarity. Then our method can be optimized using a Quality-aware Contrastive Loss (QC-Loss), which is formulated as follows:

$$
\mathcal{L}_{QC} = -\beta \cdot \log \sum_{k=1}^{K} \frac{\exp(f_{k,i}^{(n)} \cdot f_{k,i}^{(n)}/\tau)}{\sum_{k \neq k'}^{K} \exp(f_{k,i}^{(n)} \cdot f_{k',j}^{(n)}/\tau)} - \log \sum_{k=1}^{K} \frac{\exp(f_{k,i}^{(n)} \cdot f_{k,i}^{(\tilde{n})}/\tau)}{\sum_{k \neq k'}^{K} \exp(f_{k,i}^{(n)} \cdot f_{k',j}^{(\tilde{n})}/\tau)},$$

where the first part denotes distinguishing degradation-based negative pairs, and the second part is for content-based negative pairs. $\beta$ is a coefficient for balancing.

Let $\mathcal{F}$ denote the transformation of the encoder network. Given an input patch $x_{k,i}^{(n)}$, the encoded feature can be noted as $f_{k,i}^{(n)} = \mathcal{F}(x_{k,i}^{(n)})/||\mathcal{F}(x_{k,i}^{(n)})||$, after a $L_2$ normalization...
images which are selected from YFCC-100M. And the selected images cover a wide and uniform range of distortions in terms of quality indicators such as brightness, colorfulness, contrast, noise, sharpness, etc. SPAQ consists of 11,125 images captured by different mobile devices, covering a large variety of scene categories. FLIVE is the largest in-the-wild IQA dataset by far, which contains 39,810 real-world distorted images with diverse contents, sizes and aspect ratios. Details are given in Tab. 2.

**Evaluation criteria.** Pearson’s Linear Correlation Coefficient (PLCC) and Spearman’s Rank-Order Correlation Coefficient (SRCC) are selected as criteria to measure the accuracy and monotonicity, respectively. They are in the range of [0, 1]. A larger PLCC means a more accurate numerical fit with MOS scores. A larger SRCC shows a more accurate ranking between samples. For all datasets, following [21, 32], we split the dataset into a 80% training set and a 20% testing set randomly.

### 4.2. Implementation Details

Our experiments are performed using PyTorch [51], and are all conducted on 8 Nvidia V100 GPUs. Following the classical pre-training and fine-tuning paradigm, we run the whole experiments in two stages:

**Pre-training.** We inherit most settings from MoCoV2 [8], while modify the pretext task and degradation process for the quality-aware purpose. Specifically, we use a ResNet50 [28] with an extra two-layer MLP as the encoder and train it on ImageNet. We use SGD optimizer with weight decay of 0.0001, momentum of 0.9 and a mini-batch size of 256. The initial learning rate is 0.03 and multiplied by 0.1 at 120 and 0.0001, momentum of 0.9 and a mini-batch size of 256. We use AdamW optimizer with weight decay of 0.01 and mini-batch size of 64. The Learning rate is initialized with 0.0002, and decayed by the cosine annealing strategy. For BID and CLIVE, experiments are trained for 100 epochs. And for other three dataset, the number is 200. By default, we select last epoch for weight initialization and evaluation. During training stage, we take five crops (i.e., four corners and a center patch) from an image, and average their corresponding prediction scores to get the final prediction score. To reduce the randomness in training/test set splitting, following [21, 60, 91], we repeat train/test procedures 10 times for each dataset, and report the median SRCC and PLCC results.

**Fine-tuning.** Following the scheme from [62], we resize the smaller edge of images to 340 while maintaining aspect ratio, and randomly sample and horizontally flip patches with size 320x320 pixels. We use AdamW optimizer with weight decay of 0.01 and mini-batch size of 64. The Learning rate is initialized with 0.0002, and decayed by the cosine annealing strategy. For BID and CLIVE, experiments are trained for 100 epochs. And for other three dataset, the number is 200. By default, we select last epoch for weight initialization and evaluation. During training stage, we take five crops (i.e., four corners and a center patch) from an image, and average their corresponding prediction scores to get the final prediction score. To reduce the randomness in training/test set splitting, following [21, 60, 91], we repeat train/test procedures 10 times for each dataset, and report the median SRCC and PLCC results.

### 4.3. Comparison with SOTA BIQA Methods

We report the performance of SOTA BIQA methods in Tab. 3. Generally, our method pushes the current SOTA results a big step forward (using the same pre-trained weight generated by QPT), improving the best results to 0.8875 (+3.0%) of SRCC and 0.9109 (+4.6%) of PLCC in BID, 0.8947 (+4.0%) of SRCC and 0.9141 (+2.4%) of PLCC in CLIVE, 0.9271 (+1.1%) of SRCC and 0.9413 (+1.3%) of PLCC in KonIQ10K, 0.9250 (+0.8%) of SRCC and 0.9279 (+0.7%) of PLCC in SPAQ, and 0.5746 (+0.9%) of SRCC and 0.6748 (+1.4%) of PLCC in FLIVE. Notably, these results are achieved using an ordinary architecture of ResNet-50, showing the power of utilizing QPT models. Besides, we also report the results of naive ResNet-50 (i.e., using ImageNet supervised pre-training).

Specifically, QPT models surpass traditional methods (e.g., NIQE [48], ILNIQE [85], BRISQUE [47] and BMPRI [46]) that rely on hand-crafted features in large margins. Compared with earlier CNN-based methods (e.g., CNN-IQA [31], WaDiQaM-NR [4] and SFA [38]), QPT shows the advantages of big data pre-training. For HyperIQA [60] that utilized a self-adaptive hyper network architecture to estimate content understanding, perception rule learning and quality predicting, QPT-ResNet50 obtains higher results without fine-designed architectures. Compared with current SOTA Transformer-based methods (e.g., MUSIQ [32] and TReS [21]), QPT models still outperform them, showing the effectiveness of quality-aware pretext training.

Due to the flexibility of QPT, it can be easily integrated with current SOTA methods by replacing their pre-trained weights. We reproduced two SOTA methods (i.e., HyperIQA and TReS) based on their official open-source code, giving the results in Tab. 4. The results of the comparison are based on a consistent dataset partitioning and training hyper-parameters, following official implementation. It can be seen that QPT further boosts SOTA methods effectively (e.g., a even better result with 0.8982 of SRCC in BID can be achieved), showing satisfactory generalization ability. We further give the cross-dataset evaluation in Tab. 5. Compared with the most competitive methods, QPT takes the lead in three scenarios.

Tab. 2. Details of selected BIQA datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Resolution</th>
<th>MOS Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID [12]</td>
<td>586</td>
<td>480P~2112P</td>
<td>0~5</td>
</tr>
<tr>
<td>CLIVE [20]</td>
<td>1,162</td>
<td>500P~640P</td>
<td>0~100</td>
</tr>
<tr>
<td>KonIQ10K [29]</td>
<td>10,073</td>
<td>768P</td>
<td>0~100</td>
</tr>
<tr>
<td>SPAQ [17]</td>
<td>11,125</td>
<td>1080P~4368P</td>
<td>0~100</td>
</tr>
<tr>
<td>FLIVE [79]</td>
<td>39,810</td>
<td>160P~700P</td>
<td>0~100</td>
</tr>
</tbody>
</table>

3Results in FLIVE follow an aligned train/test split with CONTRIQUE. When using random 80%/20% splits, SRCC is 0.5746, PLCC is 0.6748.
Table 3. Performance of existing SOTA methods and the proposed QPT models on five in-the-wild BIQA datasets. The “⋆” means missing corresponding results in the original paper. The best and second-best results are bolded and underlined. With the inherent learned knowledge in QPT models, a widely-used ResNet50 can outperform existing SOTA methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>BID</th>
<th>CLIVE</th>
<th>KonIQ10K</th>
<th>SPAQ</th>
<th>FLIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRCC PLCC</td>
<td>SRCC PLCC</td>
<td>SRCC PLCC</td>
<td>SRCC PLCC</td>
<td>SRCC PLCC</td>
</tr>
<tr>
<td>NIQE [48]</td>
<td>0.4772</td>
<td>0.4731</td>
<td>0.4536</td>
<td>0.4676</td>
<td>0.5260</td>
</tr>
<tr>
<td>ILNIQE [85]</td>
<td>0.4946</td>
<td>0.4538</td>
<td>0.4531</td>
<td>0.5114</td>
<td>0.5029</td>
</tr>
<tr>
<td>BRISQUE [47]</td>
<td>0.5736</td>
<td>0.5401</td>
<td>0.6005</td>
<td>0.6211</td>
<td>0.7150</td>
</tr>
<tr>
<td>BMPRI [46]</td>
<td>0.5154</td>
<td>0.4583</td>
<td>0.4868</td>
<td>0.5229</td>
<td>0.6577</td>
</tr>
<tr>
<td>CNNIQQA [31]</td>
<td>0.6163</td>
<td>0.6144</td>
<td>0.6269</td>
<td>0.6008</td>
<td>0.6852</td>
</tr>
<tr>
<td>WaDiQaM-NR [4]</td>
<td>0.6526</td>
<td>0.6359</td>
<td>0.6916</td>
<td>0.7304</td>
<td>0.7294</td>
</tr>
<tr>
<td>SFA [38]</td>
<td>0.8202</td>
<td>0.8253</td>
<td>0.8037</td>
<td>0.8213</td>
<td>0.8882</td>
</tr>
<tr>
<td>DB-CNN [89]</td>
<td>0.8450</td>
<td>0.8590</td>
<td>0.8443</td>
<td>0.8624</td>
<td>0.8780</td>
</tr>
<tr>
<td>HyperIQA [60]</td>
<td>0.8544</td>
<td>0.8585</td>
<td>0.8546</td>
<td>0.8709</td>
<td>0.9075</td>
</tr>
<tr>
<td>PaQ-2-PiQ [79]</td>
<td>- -</td>
<td>0.840</td>
<td>0.850</td>
<td>- -</td>
<td>0.870</td>
</tr>
<tr>
<td>CONRTIQUE* [44]</td>
<td>- -</td>
<td>0.854</td>
<td>0.890</td>
<td>- -</td>
<td>0.896</td>
</tr>
<tr>
<td>UNIQUE* [90]</td>
<td>0.858</td>
<td>0.873</td>
<td>0.854</td>
<td>0.890</td>
<td>0.916</td>
</tr>
<tr>
<td>MUSIQ* [32]</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>0.916</td>
</tr>
<tr>
<td>ResT* [21]</td>
<td>- -</td>
<td>0.846</td>
<td>0.877</td>
<td>- -</td>
<td>0.915</td>
</tr>
<tr>
<td>ResNet50</td>
<td>0.8423</td>
<td>0.8621</td>
<td>0.8527</td>
<td>0.8807</td>
<td>0.9121</td>
</tr>
<tr>
<td>QPT-ResNet50</td>
<td>0.8875</td>
<td>0.9109</td>
<td>0.8947</td>
<td>0.9141</td>
<td>0.9271</td>
</tr>
</tbody>
</table>

4.4. Ablation Studies

Impact of data amount for QPT. To evaluate the effectiveness of the quality-aware pretext task, we use 20%, 50% and 100% percentages of the ImageNet dataset, analyzing the impact on the number of used data for downstream results. As given in Tab. 6, the larger the number used in the pre-training process, the better the performance on the downstream IQA datasets. Still, it is quite a surprise that our method obtains SOTA results with only 50% of the data. Furthermore, training larger-scale data is an open question and can be investigated in future work, where QPT may benefit from JFT-300M [61].

Impact of encoder capacity for QPT. To verify the effect of model capacity on feature modeling, we select two series of networks for analysis, including the widely-used CNN architectures of ResNet and an emerging Transformer-based architectures of MobileViTV2 [45]. As shown in Tab. 7, increasing the model size can help improve the ability of feature representation, resulting in better downstream performance.
performance. It should be pointed out that large models will also increase the time consumption of the pre-training phase (i.e., ResNets of \( \{ 53h, 61h, 75h \} \), MBViTV2s of \( \{ 66h, 92h, 109h \} \)). We did not use larger networks out of tradeoffs, which can be tried in future work.

**Impact of different compositions of negative samples.** To evaluate the effectiveness of negative samples and the proposed QC Loss, we give the results in Tab. 8. The highest result is achieved when both types of negative samples are present. And the absence of inter-sample negative pairs leads to a larger performance drop. This is due to the fact that compared with intra-sample negative pairs (i.e., \( K-1 \)), the number of inter-sample ones (i.e., \( N-K \)), which contain both different content and degradation types, is larger.

**Ablation on the degradation space.** Experiments are performed to evaluate the impact of different degradations and compositions. In Fig. 4, we give the results under different compositions of degradation types in a first-order manner. Trying all combinations is impractical due to the huge degradation space. Experiments are performed in a coarse-grained manner. The results show that there is a clear effect difference between the degradation types of a fixed form. More flexible and larger space are needed to cover different downstream scenarios. We further evaluate the strategy of composition as shown in Fig. 2. In Tab. 9, a combination of the three strategies yielded the best results.

**4.5. Comparisons with other pretext tasks**

We further compare QPT with other pretext methods, including train-from-scratch (i.e., \( \text{w/o} \) pre-training weights), supervised and MoCo. We extract features from the encoder output for both fine-tuning and linear probing evaluations.

**Linear probing evaluation.** In this way, the weights of encoder are frozen for feature extraction directly. As shown in Tab. 10, QPT obtains the best performance, surpassing MoCo by 8.75% of SRCC in BID and 8.17% in CLIVE.

**End-to-end fine-tuning evaluation.** Linear probing has been a popular protocol in the past few years. However, it misses the opportunity of pursuing strong but nonlinear features which is indeed a strength of deep learning [26]. We also give the fine-tuning results in Tab. 11. QPT still achieves the best results. Generally, higher results can be obtained than linear probing. These results further prove the effectiveness of QPT over instance-level discrimination or supervised classification pretext tasks in IQA scenarios.

**5. Conclusion**

In this paper, we proposed the QPT to generate pre-trained models for downstream BIQA tasks, alleviating the obstacle of insufficient annotated data. By introducing diverse degradation types and compositions, we construct a degradation space that contains \( 2 \times 10^7 \) possible degradations, to simulate more complicated and realistic distorted images. To constrain the learning process, we proposed the QC-Loss that treats patch pairs extracted from the same degraded images as positive ones. And other pairs are noted as negative ones, including degradation-based and content-based ones. After pre-trained on ImageNet using QPT, BIQA models obtain significant improvements on five BIQA benchmark datasets. Moreover, QPT can be easily integrated with current SOTA methods by replacing their pre-trained weights, showing good generalization ability.
References


