

Learning Generalizable Perceptual Representations for Data-Efficient No-Reference Image Quality Assessment

Suhas Srinath* Shankhanil Mitra* Shika Rao Rajiv Soundararajan

Indian Institute of Science
Bengaluru, India 560012

{suhass12, shankhanilm, shikarao, rajivs}@iisc.ac.in

Abstract

No-reference (NR) image quality assessment (IQA) is an important tool in enhancing the user experience in diverse visual applications. A major drawback of state-of-the-art NR-IQA techniques is their reliance on a large number of human annotations to train models for a target IQA application. To mitigate this requirement, there is a need for unsupervised learning of generalizable quality representations that capture diverse distortions. We enable the learning of low-level quality features agnostic to distortion types by introducing a novel quality-aware contrastive loss. Further, we leverage the generalizability of vision-language models by fine-tuning one such model to extract high-level image quality information through relevant text prompts. The two sets of features are combined to effectively predict quality by training a simple regressor with very few samples on a target dataset. Additionally, we design zero-shot quality predictions from both pathways in a completely blind setting. Our experiments on diverse datasets encompassing various distortions show the generalizability of the features and their superior performance in the data-efficient and zero-shot settings.

1. Introduction

The increasing number of imaging devices, including cameras and smartphones, has significantly increased the volume of images captured, edited, and shared on a global scale. As a result, there is a necessity to assess the quality of visual content to enhance user experience. Image Quality Assessment (IQA) is generally divided into two categories: full-reference (FR) and no-reference (NR) IQA. While FR-IQA relies on pristine reference images for quality assessment, NR-IQA is more relevant and challenging due to the absence of clean references for user-captured images.

Most successful NR IQA methods are deep-learning based, and require a large number of images with human opinion scores for training. As imaging systems evolve, the distortions also evolve, making it difficult to keep creating large annotated datasets for training NR IQA models. This motivates the study of limited data or data-efficient NR IQA models which can be trained on a target IQA application (or database) with limited labels. Such an approach works best if the learned quality representations can generalize well across different distortion types for various IQA tasks. These representations can then be mapped to quality using a simple linear model [16, 25] using the limited labels on a target application. Further, it is desirable that we learn these representations without requiring any human annotations of image quality. The goal of our work is to learn generalizable image quality representations through self-supervised learning to design data-efficient NR models for a target IQA application.

In this regard, while DEIQT [22] studies the data-efficient IQA problem, they train the entire network with millions of parameters, which still requires a reasonable number of labeled training images. On the other hand, recent methods such as CONTRIQUE [16], Re-IQA [25], and QPT [4] focus on self-supervised contrastive learning to learn quality features, which can potentially yield superior performance with limited labels. However, these methods do not consider that images with different distortions could have the same quality, thereby limiting the generalizability of their image features across varied distortions.

Our main contribution is the design of Generalizable Representations for Quality (**GRepQ**) that can predict quality by training a simple linear model with few annotations. We present two sets of features, one to capture the local low-level quality variations and another to predict quality using the global context. To capture low-level quality features, we propose a quality-aware contrastive learning strategy guided by a perceptual similarity measure between distorted versions of an image. In particular, we bring similar-quality images closer in the latent space irrespective of their

*Equal contribution.

§<https://github.com/suhas-srinath/GRepQ>

distortion types. This is achieved by assigning a weight based on a similarity measure between every pair of distorted versions of an image. Our strategy enables the learning of generalizable quality representations invariant to distortion types.

We also leverage the generalization capabilities of large vision-language models for extracting high-level quality information. Notably, the versatile CLIP [23] model can be applied to zero-shot quality prediction [28], although a lack of task-specific fine-tuning limits it. While LIQE [42] finetunes CLIP by integrating scene and distortion information, it requires large-scale training with human labels. We overcome these limitations through a novel unsupervised finetuning of CLIP. We achieve this by segregating images of higher and lower quality into groups using antonym text prompts and employing a group-contrastive loss with respect to the prompts. Our group-contrastive learning facilitates the learning of high-level quality representations that can generalize well to diverse content and distortions.

The features from both pathways can be combined to learn a simple regressor trained with few samples from any IQA dataset. Additionally, predictions can be made in a zero-shot setting using the learned features, which can then be combined to provide a single objective score. We show through extensive experiments that our framework shows superior performance in both the data-efficient as well as zero-shot settings. We summarize the main contributions of our framework as follows:

- A quality-aware contrastive loss that weighs positive and negative training pairs using a “soft” perceptual similarity measure between a pair of samples to enable representation learning invariant to distortion types.
- An unsupervised task-specific adaptation of a vision-language model to capture semantic quality information. We achieve this by separating higher and lower-quality groups of images based on quality-relevant antonym text prompts.
- Superior performance of our method over other NR-IQA methods trained using few samples (data-efficient) on several IQA datasets to highlight the generalizability of our features. Additionally, we show superior cross-database prediction performance.
- A zero-shot quality prediction method using the learned features and its superior performance compared to other zero-shot (or completely blind) methods.

2. Related Work

2.1. Supervised NR-IQA

Many popular supervised NR-IQA methods such as BRISQUE [17], DIIVINE [19], BLIINDS [24], CORNIA

[35] predict quality using hand-crafted natural scene statistics based features. Such methods have succeeded when images contain synthetic distortions but often suffer when the distortions are more complex or authentic. To mitigate this, several deep learning-based methods have emerged that are either trained in an end-to-end fashion [3, 11, 41] or use a pre-trained feature encoder that can be fine-tuned for IQA [41]. Further, transformer-based models have shown promise on authentic and synthetically distorted images [6, 27, 29, 37]. Methods such as MetaIQA [44] employ meta-learning to learn from synthetic data and adapt to real-world images efficiently. A recent method, LIQE [42] adapts the CLIP model for IQA via scene and distortion classification along with supervised fine-tuning on several IQA datasets. However, the model requires multiple annotations per image during training making the model infeasible when adapting to newer and more complex datasets in the data-efficient regime.

2.2. Self-Supervised Quality Feature Learning

Although supervised NR-IQA methods have shown reasonable performance in quality prediction, they still possess the limitation of requiring large amounts of human annotations for training. One of the earliest approaches in this domain was through the design of quality-aware codebooks [35]. Later, different ranking-based methods were used for quality-aware pre-training [14]. Contrastive learning-based training such as CONTRIQUE [16], Re-IQA [25] and QPT [4] learn quality representations by contrasting multiple levels of synthetic distortions. While Re-IQA [25] also uses high and low-level features, our method significantly differs from Re-IQA in how the low-level and high-level features are designed. Further, all the above methods neither consider the generalizability to unseen distortions nor do they consider the data-efficient evaluation setting.

2.3. Zero-Shot or Completely Blind (CB) IQA

Another class of IQA methods are zero-shot or completely blind and do not require any human opinions for their design. For example, NIQE [18] neither requires training on a dataset of annotated images nor knowledge about possible degradations. IL-NIQE [38] improves over NIQE by integrating other quality-aware features based on Gabor filter responses, gradients, and color statistics. However, both methods tend to fail on authentic and other complex distortions. A recent method [2] learns deep features using contrastive learning to predict quality without any supervision. However, the performance shown on in-the-wild IQA datasets still provides scope for further improvement. Leveraging the contextual information from CLIP [23], CLIP-IQA [28] shows that a zero-shot application of the CLIP model can yield promising quality predictions. However, zero-shot methods tend to have poorer perfor-

mance and motivate the use of limited labels on target IQA applications to improve performance.

2.4. Data-Efficient IQA

IQA in the low-data setting remains relatively unexplored. Data-efficient image quality assessment (DEIQT) [22] shows that IQA models can be efficiently fine-tuned with very few annotated samples from a target dataset, enabling generalization through data efficiency. Further, with a sufficient number of training samples, data-efficient training can achieve performances of full dataset supervision on multiple IQA datasets. However, DEIQT still requires end-to-end fine-tuning of a transformer model, leading to increased training times.

3. Method

We first describe our approach to learning generalizable low-level and high-level quality representations. The overall framework is illustrated in Fig. 1. We discuss how quality is predicted in the data-efficient and zero-shot settings.

3.1. Low-Level Representation Model

Contrastive learning for image quality [4, 16, 25] discriminates images based on varied types and levels of distortions to capture low-level information. While this provides very good pre-trained feature encoders without learning from human labels, images with different distortion types are often treated as a negative pair with respect to an anchor image, and the features of such images are pulled apart. This leads to two main issues. An image with a different distortion type may have a similar perceptual quality as the anchor. Secondly, since image representations with different distortion types are separated, this hurts the model’s generalizability to represent unseen distortions. The goal of our work is to address both these limitations through a quality-aware contrastive learning loss.

We introduce a novel quality-aware contrastive loss, where positive and negative pairs (pairs of images considered similar and dissimilar in quality, respectively) are selected based on their perceptual similarity. This allows a soft weighting such that a similarity weight close to one treats the pair of images as positive and pulls their corresponding representations closer. Similarly, image features are pulled apart when the perceptual similarity is near zero. This differs from the way prior methods use contrastive learning. In particular, our framework allows the selection of pairs regardless of their distortions, allowing for generalization.

Image Augmentation: In order to train the feature encoder using contrastive learning, we generate multiple synthetically distorted versions of a camera-captured image and sample fragments from each image. Four synthetic distortions are generated: blur, compression, noise, and color sat-

uration, at two levels each. Fragment sampling has proven effective in retaining the global quality information of an image [32]. To obtain fragments, we divide an image into grids, and random mini-patches are extracted from each of the grid locations. The mini-patches are then stitched together to yield a single fragmented image that is used to train the model. An augmentation is generated by randomly sampling another set of mini-patches from the same image to obtain another fragmented image. Note that this augmentation is quality preserving and can be used as a hard-positive pair in contrastive loss.

Quality-Aware Contrastive Loss: We contrast multiple distorted versions of the same scene to learn quality representations and mitigate content bias. Consider a batch of N_b images $\{\mathbf{x}_i\}_{i=1}^{N_b}$, where each image has D distorted versions. Let \mathbf{x}_i^j and \mathbf{x}_i^k denote two distorted versions of an image \mathbf{x}_i where $j, k \in \{1, 2, \dots, D\}$. Let \mathbf{z}_i^j and \mathbf{z}_i^k be the respective unit-norm feature representations obtained as $\mathbf{z}_i^j = f_\theta(T(\mathbf{x}_i^j))$ and $\mathbf{z}_i^k = f_\theta(T(\mathbf{x}_i^k))$, where $T(\cdot)$ is the fragment sampling operation and $f_\theta(\cdot)$ is the feature encoder. Let $s(\cdot, \cdot) : \mathbb{R}^{M \times N} \times \mathbb{R}^{M \times N} \rightarrow [0, 1]$ denote a perceptual similarity measure between two images with the same content. Further, let $p_\tau(\mathbf{z}_1, \mathbf{z}_2) = \exp(\mathbf{z}_1 \cdot \mathbf{z}_2 / \tau)$. We overcome the limitation of existing contrastive learning methods which require hard positives and negatives through the above soft similarity measure to label positives and negatives. The similarity measures the closeness of distorted versions in terms of intrinsic quality attributes and provides a confidence weight in contrastive loss.

Our quality-aware contrastive loss is given by $\mathcal{L}_{QACL} = \sum_{i=1}^{N_b} \sum_{j=1}^D \mathcal{L}_i^j$, where \mathcal{L}_i^j is given by

$$\mathcal{L}_i^j = -\log \frac{p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^{j+}) + \sum_{k \neq j} s(\mathbf{x}_i^j, \mathbf{x}_i^k) p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^k)}{p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^{j+}) + \sum_{k \neq j} p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^k)}, \quad (1)$$

where \mathbf{z}_i^{j+} is the representation of an augmentation of the image \mathbf{x}_i^j , and τ_1 is a temperature hyperparameter. Note that since $p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^k) = s(\mathbf{x}_i^j, \mathbf{x}_i^k) p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^k) + (1 - s(\mathbf{x}_i^j, \mathbf{x}_i^k)) p_{\tau_1}(\mathbf{z}_i^j, \mathbf{z}_i^k)$, \mathbf{x}_i^k is treated as similar to \mathbf{x}_i^j with weight $s(\mathbf{x}_i^j, \mathbf{x}_i^k)$ and dissimilar with weight $(1 - s(\mathbf{x}_i^j, \mathbf{x}_i^k))$. The similarity function makes the learning distortion type agnostic since it measures relative degradation without knowledge of the distortion type, making the learned features generalizable to different (and unseen) distortions. The InfoNCE [20] loss can be seen as a special case of \mathcal{L}_{QACL} when $s(\cdot, \cdot) = 0$.

It is desirable that the perceptual similarity measure used satisfies a few properties: (1) It captures intrinsic quality-specific attributes, such as structure, sharpness, or contrast, (2) It is capable of handling various distortion types used during training and correlates well with human judgments on these distortions, (3) It captures local and global qual-

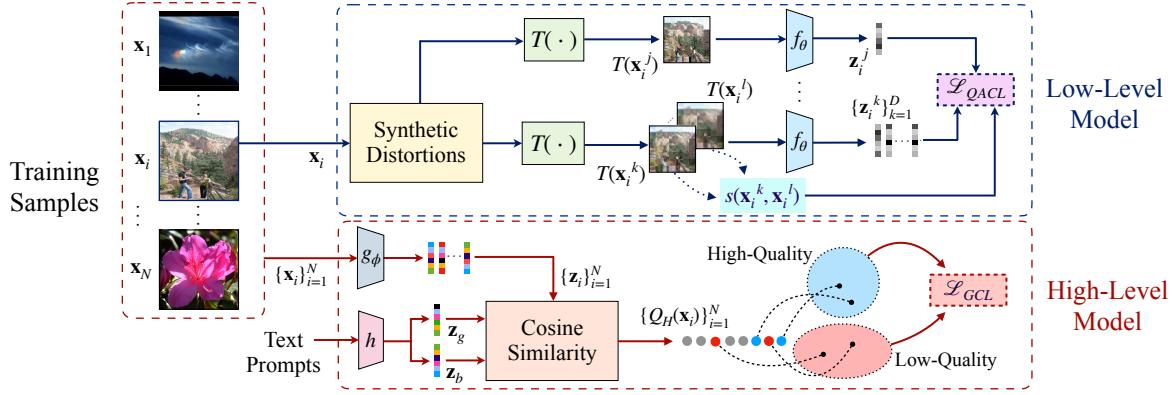


Figure 1. Illustration of the GRepQ framework. The low-level model, f_θ , is trained using multiple distorted versions of an image \mathbf{x}_i subjected to the fragment sampling operation $T(\cdot)$. \mathbf{x}_i^j denotes an anchor image. The perceptual similarity measure $s(\cdot, \cdot)$ is used to weigh the feature similarities of pairs of distorted images \mathbf{x}_i^k and \mathbf{x}_i^l in Eq. (1). The high-level model, g_ϕ , is trained using Eq. (3) after selecting groups of features based on their cosine similarities with embeddings \mathbf{z}_g and \mathbf{z}_b of antonym text prompts that relate to higher and lower quality respectively. Embeddings are obtained from the text-encoder $h(\cdot)$.

ity information relevant to the human visual system, and (4) It predicts similarity with fairly low complexity to enable faster training times. We explore different similarity measures such as FSIM [39], SSIM [30], GMSD [33], MSSSIM [31] and LPIPS [40] in our work. Degraded reference IQA [1] has also shown that such similarity measures can be used to distinguish between distorted versions of a degraded reference image. Finally, we note that such similarity measures can be used to compare different distorted versions of an anchor image since all of them have the same content. Thus, we do not include variations of different images in this loss.

3.2. High-Level Representation Model

To understand the scene context for IQA, we adapt the CLIP model to the IQA task. Image quality can be obtained using CLIP by measuring the cosine similarity between the image feature and the text embeddings of a pair of antonym prompts such as ‘‘a good photo.’’, ‘‘a bad photo.’’ [28]. Although CLIP has a reasonable zero-shot quality prediction performance in terms of correlation with human opinion, the representations are not specifically crafted for the task of IQA, leading to a performance gap. We bridge this gap by fine-tuning the image encoder of CLIP through an unsupervised loss, as described below.

Contrastive Learning Over Groups: Analyses of vision-language models show that text representations are richer than image representations [13, 21]. Thus, we fix the text encoder in the CLIP model and only update the image encoder for the IQA task. The text representations corresponding to the antonym prompts remain the same during training and testing. We propose a loss for updating the im-

age encoder that aims at separating images in a batch into groups based on how close the image representations are to two text-prompt embeddings. We then seek to align the representations of images within each group and separate the representations across groups. Such a loss simultaneously ensures that the intra-group feature entropy (entropy of representations within each group) is minimized and the inter-group entropy (entropy of features between groups) is maximized [10, 43].

Consider a batch consisting of N images $\{\mathbf{x}_i\}_{i=1}^N$ with visual representations $\{\mathbf{z}_i\}_{i=1}^N$. The representations are obtained as $\mathbf{z}_i = g_\phi(\mathbf{x}_i)$, where $g_\phi(\cdot)$ is the CLIP image encoder. Let \mathbf{z}_g and \mathbf{z}_b correspond to the prompt representations of ‘‘a good photo.’’ and ‘‘bad photo.’’ respectively. We construct two groups of images, \mathcal{S}_g and \mathcal{S}_b , that correspond to higher and lower quality respectively based on the quality estimated as

$$Q_H(\mathbf{x}_i) = \frac{1}{1 + \exp(k_2(\mathbf{z}_i \cdot \mathbf{z}_b - \mathbf{z}_i \cdot \mathbf{z}_g))}, \quad (2)$$

where k_2 is a scaling parameter. Let the features $\{\mathbf{z}_i\}_{i=1}^N$ sorted in increasing order of Q_H be $\{\mathbf{z}_{(1)}, \mathbf{z}_{(2)}, \dots, \mathbf{z}_{(N)}\}$. We obtain the groups as $\mathcal{S}_b = \{\mathbf{z}_{(i)}\}_{i=1}^M$ and $\mathcal{S}_g = \{\mathbf{z}_{(i)}\}_{i=N-M+1}^N$, where $M = \text{round}(N/k)$, k is a hyper-parameter that decides the separability of lower and higher quality groups within a batch of images, and M denotes the group size. Let $\mathcal{S}_b(i) = \mathcal{S}_b \setminus \{\mathbf{z}_i\}$ and $\mathcal{S}_g(i) = \mathcal{S}_g \setminus \{\mathbf{z}_i\}$. Our group contrastive loss used for fine-tuning is expressed

as

$$\begin{aligned} \mathcal{L}_{GCL} = & - \sum_{\mathbf{z}_i \in \mathcal{S}_g} \log \frac{\sum_{\mathbf{z}_j \in \mathcal{S}_g(i)} p_{\tau_2}(\mathbf{z}_i, \mathbf{z}_j)}{\sum_{\mathbf{z}_j \in \mathcal{S}_g(i) \cup \mathcal{S}_b} p_{\tau_2}(\mathbf{z}_i, \mathbf{z}_j)} \\ & - \sum_{\mathbf{z}_i \in \mathcal{S}_b} \log \frac{\sum_{\mathbf{z}_j \in \mathcal{S}_b(i)} p_{\tau_2}(\mathbf{z}_i, \mathbf{z}_j)}{\sum_{\mathbf{z}_j \in \mathcal{S}_b(i) \cup \mathcal{S}_g} p_{\tau_2}(\mathbf{z}_i, \mathbf{z}_j)}. \end{aligned} \quad (3)$$

While creating groups, a quality separation gap and closeness of quality scores of images in each group are necessary for effective contrastive learning. The parameter k controls this separability and is a hyperparameter that needs to be appropriately chosen.

3.3. Mapping Representations to Objective Quality

Data-Efficient Quality Prediction: Once the high and low-level features are learned, they are concatenated and regressed with mean opinion scores on the evaluation datasets using a few samples from each dataset. We use a linear SVR $f_d(\cdot) : \mathbb{R}^P \rightarrow \mathbb{R}$ on features of target datasets, where P is the feature dimension. The data-efficient quality of any new image \mathbf{x} can simply be computed using its corresponding feature representation $\mathbf{z}_x \in \mathbb{R}^P$ as

$$\text{GRepQ}_D(\mathbf{x}) = f_d(\mathbf{z}_x). \quad (4)$$

Our approach offers the advantage of requiring no end-to-end training using the limited labels on a new target database.

Zero-Shot Quality Prediction: We use different approaches for the low-level and high-level representations to predict quality without using any supervision. For the low-level features, we compute a distance between the features of the input image and that of a corpus of pristine images similar to NIQE as

$$d(\mathbf{x}) = \sqrt{(\mu_p - \mu_d)^T \left(\frac{\Sigma_p + \Sigma_d}{2} \right)^{-1} (\mu_p - \mu_d)}, \quad (5)$$

where μ_p and Σ_p are the mean and covariance of the representations from the low-level encoder corresponding to patches of pristine images. μ_d and Σ_d are the mean and covariance of the representations of the patches from an image \mathbf{x} . Here, non-overlapping patches of size $R \times R$ are extracted from the image to estimate the relevant statistics of the features. The low-level quality is then predicted as

$$Q_L(\mathbf{x}) = \frac{1}{1 + \exp(k_1 d(\mathbf{x}))}, \quad (6)$$

where k_1 is a scaling parameter. The quality from the high-level representations can be predicted using Eq. (2). The overall image quality is then measured as

$$\text{GRepQ}_Z(\mathbf{x}) = Q_H(\mathbf{x}) + Q_L(\mathbf{x}), \quad (7)$$

and is illustrated in Fig. 2.

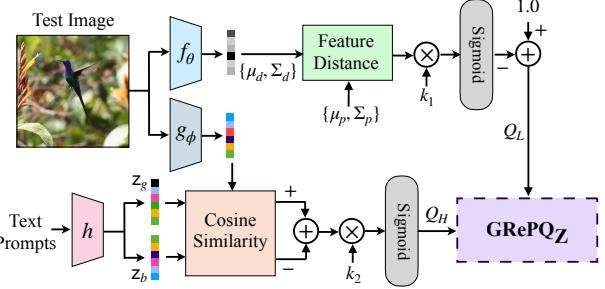


Figure 2. GRepQ_Z: Computing zero-shot image quality. The text-prompts used for evaluation are "a good/bad photo.".

4. Experiments

4.1. Training and Implementation Details

Training Dataset for Representation Learning: We train the low and high-level feature encoders on the FLIVE dataset [36] using a subset of 5000 real-world images encompassing a variety of authentic distortions with different resolutions and aspect ratios. The diverse content and distortions make it conducive to learning representations that can be generalized to diverse images. No human annotations were used during this training.

Low-Level Encoder: We use a ResNet18 (performance with a Resnet50 was found to be similar) without pre-trained weights for the low-level feature encoder. The contrastive loss in Eq. (1) is trained by projecting the features from the penultimate layer of ResNet18 onto \mathbb{R}^{128} . Images are fragmented into 7×7 grids, and random minipatches from each grid location are stitched together to form 224×224 sized patches. The temperature τ_1 is fixed at 0.5. A batch consists of 8 images with 8 distorted versions each. The model is trained for 15 epochs using the AdamW [15] optimizer with a weight decay of 0.05 and an initial learning rate of 10^{-4} . A cosine learning rate scheduler is used. To guide the quality-aware contrastive training, we employ FSIM as the perceptual similarity measure.

High-Level Encoder: We fine-tune CLIP's image encoder while keeping the text encoder fixed. The image encoder consists of a Resnet50 backbone with an additional attention-pooling layer. To enable contrastive learning over groups specified in Eq. (3), a projection head is used to contrast features in \mathbb{R}^{128} . The images are center-cropped to a size of 224×224 , and a batch size of $N = 128$ is used. Based on the coarse predictions obtained using Eq. (2), we use a separability hyperparameter $k = 8$ to divide the batch of images into groups of size $M = 16$. Once the groups are formed, the image encoder is trained using Eq. (3) with a temperature $\tau_2 = 0.1$. The model is trained for 15 epochs using an Adam optimizer with an initial learning rate of 5×10^{-6} . The scaling parameter k_2 is set to 10.

Method	Method	CLIVE			KonIQ			CSIQ			LIVE			PIPAL		
		50	100	200	50	100	200	50	100	200	50	100	200	50	100	200
TReS [6]	End-to-end	0.670	0.751	0.799	0.713	0.719	0.791	0.791	0.811	0.878	0.901	0.927	0.957	0.186	0.349	0.501
HyperIQA [27]		0.648	0.725	0.790	0.615	0.710	0.776	0.790	0.824	0.909	0.892	0.912	0.929	0.102	0.302	0.379
DEIQT [22]		0.667	0.718	0.812	0.638	0.682	0.754	0.821	0.891	0.941	0.920	0.942	0.955	0.396	0.410	0.436
MANIQA [34]		0.642	0.769	0.797	0.652	0.755	0.810	0.794	0.847	0.874	0.909	0.928	0.957	0.136	0.361	0.470
LIQE [42]		0.691	0.769	0.810	0.759	0.801	0.832	0.838	0.891	0.924	0.904	0.934	0.948	-	-	-
Resnet50 [8]	Simple Feature Regression	0.576	0.611	0.636	0.635	0.670	0.707	0.793	0.890	0.935	0.871	0.906	0.922	0.150	0.220	0.302
CLIP [23]		0.664	0.721	0.733	0.736	0.770	0.782	0.841	0.892	0.941	0.896	0.923	0.941	0.254	0.303	0.368
CONTRIQUE [16]		0.695	0.729	0.761	0.733	0.794	0.821	0.840	0.926	0.940	0.891	0.922	0.943	0.379	0.437	0.488
Re-IQA [25]		0.591	0.621	0.701	0.685	0.723	0.754	0.893	0.907	0.923	0.884	0.894	0.929	0.280	0.350	0.431
GRepQ _D (LL)		0.531	0.565	0.613	0.620	0.647	0.679	0.794	0.805	0.832	0.866	0.880	0.886	0.395	0.410	0.431
GRepQ _D (HL)		0.740	0.770	0.796	0.794	0.813	0.843	0.869	0.905	0.932	0.904	0.927	0.944	0.410	0.415	0.427
GRepQ _D (HL + LL)		0.760	0.791	0.822	0.812	0.836	0.855	0.878	0.914	0.941	0.926	0.937	0.953	0.489	0.518	0.548

Table 1. SRCC performance comparison of GRepQ_D with other NR-IQA methods trained using few labels on various IQA databases. The methods are segregated into end-to-end trained (top five) and feature-learning-based (next four) methods. LL and HL correspond to low and high-level models respectively. The best-performing methods are bolded.

Zero-Shot Quality Prediction using Low-Level Encoder:

For the zero-shot quality prediction using Eq. (5), we select 125 pristine image patches as used in literature [2] (chosen based on sharpness and colorfulness). Patches of size 96×96 are extracted from the pristine images and the test image. The scaling parameter k_1 is set to 0.01.

All the implementations were done in PyTorch using two 11GB Nvidia GeForce RTX 2080 Ti GPUs.

4.2. Experimental Setup

We present the details of the two main evaluation settings: data-efficient setting and the zero-shot setting. In the data-efficient setting, we train our data-efficient framework GRepQ_D, using a few samples from each evaluation dataset. We randomly split each evaluation dataset into 80% and 20% and use the 20% subset for testing. We select a random subset of 50, 100, or 200 samples from the 80% for training a linear support vector regressor (SVR) on the features. We use Spearman’s rank order correlation coefficient (SRCC) between the objective and subjective scores to evaluate the models’ performance. We report the median performance obtained across 10 splits of each evaluation dataset. The results with respect to Pearson’s linear correlation coefficient (PLCC) are given in the supplementary. In the zero-shot setting, no training on any evaluation dataset is required, and we test on the entire evaluation dataset.

Evaluation Datasets: We choose a variety of datasets spanning different types of distortions to demonstrate the effectiveness of our framework for the three experimental settings. Since the training images are sampled from the FLIVE dataset, we do not evaluate them on FLIVE. We evaluate two popular in-the-wild datasets: CLIVE [5], KONIQ [9], and three synthetic or processed image datasets: LIVE-IQA [26], CSIQ [12] and PIPAL [7]. **CLIVE** contains 1,162 images captured from multiple mobile devices. **KonIQ-10K** contains 10073 in-the-wild images. **LIVE-IQA** [26] contains 29 scenes along with 779

distorted images containing JPEG compression, blur, noise, and fast-fading distortions. **CSIQ** [12] consists of 30 original images with 866 distorted images with blur, contrast, and JPEG compression distortions. **PIPAL** is a large IQA database consisting of 23,200 images with 40 different distortions per image, including GAN-generated artifacts, making this dataset very challenging to evaluate.

4.3. Data-Efficient Setting

We compare GRepQ_D with other state-of-the-art (SoTA) end-to-end NR-IQA methods: TReS [6], HyperIQA [27], and MANIQA [34], the data-efficient method DEIQT [22], and feature based methods: Resnet50 [8], CLIP [23], CONTRIQUE [16] and Re-IQA [25]. We note that LIQE [42] is not trainable on PIPAL and thus its entry is left blank. For the methods requiring feature regression, the SVR parameters are optimized to yield the best performances. To ensure fair comparisons, the median performance of all methods over ten train-test splits are reported.

Tab. 1 presents comparisons on the data-efficient training of GRepQ_D against other NR-IQA methods. The results indicate that GRepQ_D outperforms other methods on all datasets in almost all three data regimes (50, 100, and 200 samples). We notice that GRepQ_D outperforms even end-to-end trained models despite using a simple SVR. The superior performance over Re-IQA, which may also be considered as an ensemble of two sets of features, demonstrates the superiority of both our low and high level features. While it may appear that the high-level model performs better than the low-level model in most of the scenarios, we provide examples in Sec. 4.6, where the low-level model could also be more accurate. Thus, there is a need for both the high and low-level representations. As an extreme case, we also present results in the fully-supervised setting in the supplement.

Method	CLIVE	KonIQ	CSIQ	LIVE	PIPAL
NIQE [18]	0.463	0.530	0.613	0.836	0.153
IL-NIQE [38]	0.440	0.507	0.814	0.847	0.282
CL-MI [2]	0.507	0.645	0.588	0.663	0.303
CLIP-IQA [28]	0.612	0.700	0.690	0.652	0.261
GRepQ _Z	0.740	0.768	0.693	0.741	0.436

Table 2. Performance comparison of GRepQ_Z (zero-shot) with other zero-shot methods on various IQA databases.

Training	FLIVE		KonIQ		CSIQ	
	CLIVE	KonIQ	CLIVE	KonIQ	CSIQ	LIVE
HyperIQA	0.758	0.735	0.785	0.772	0.744	0.926
TReS	0.713	0.740	0.786	0.733	0.761	-
CONTRIQUE	0.710	0.781	0.731	0.676	0.823	0.925
DEIQT	0.733	0.781	0.794	0.744	0.781	0.932
GRepQ _C	0.774	0.815	0.774	0.792	0.770	0.893

Table 3. Cross-dataset performance of GRepQ_C along with other NR-IQA methods. Results for methods apart from CONTRIQUE are from [22].

4.4. Zero-Shot Setting

Since zero-shot methods are trained without human supervision, we compare GRepQ_Z with unsupervised or completely blind NR-IQA methods such as NIQE [18], IL-NIQE [38], contrastive learning with mutual information (CL-MI) [2], and CLIP-IQA [28]. We utilize entire evaluation databases for testing all the methods.

Tab. 2 shows that GRepQ_Z consistently outperforms other methods on three out of five datasets by considerable margins. GRepQ_Z achieves SoTA performance even on the challenging PIPAL dataset, containing diverse distortions, particularly images restored by various restoration (including GAN-based) methods for super-resolution and denoising. A 44% improvement in SRCC is shown over the second-best-performing algorithm (CL-MI). The lower performance of GRepQ_Z on LIVE and CSIQ is attributed to content bias of both the low-level and high-level models. Although the low-level model is trained in a content conditional manner, the features perhaps do suffer from some residual content bias. Since LIVE and CSIQ contain very few unique scene content, the residual content bias leads to reduced performance of our zero-shot model. Despite these challenges, GRepQ_Z still achieves competitive performance, showing its generalization capability in the zero-shot setting.

4.5. Cross-Database Experiments

We also show the effectiveness of our features through cross-database experiments. Here, a single linear SVR (ridge regressor) is trained on an entire dataset and tested on other intra-domain datasets in the authentic and synthetic image settings. The results in Tab. 3 indicate that GRepQ_C (defined as the cross-dataset prediction eval-

Similarity Measure	50	100	200
None	0.381	0.413	0.452
SSIM	0.533	0.558	0.590
MS-SSIM	0.527	0.561	0.575
GMSD	0.544	0.570	0.583
LPIPS	0.578	0.605	0.629
FSIM	0.620	0.647	0.679

Table 4. SRCC performance analysis on the KonIQ dataset of the impact of different perceptual similarity measures on the low-level model under the data-efficient setting.

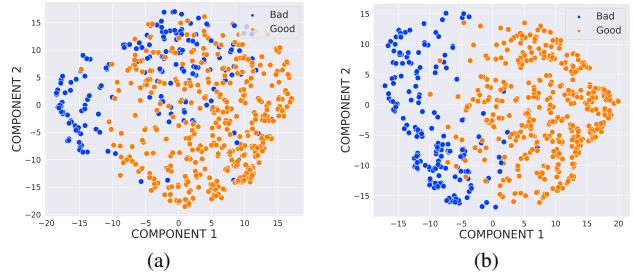


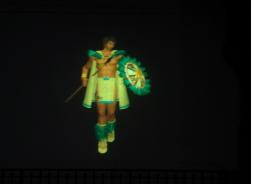
Figure 3. A visualization of feature representations of images from a combination of CLIVE and the test set of KonIQ using t-SNE for features from (a) the zero-shot CLIP image encoder and (b) our fine-tuned high-level encoder. Blue and orange points correspond to bad and good-quality images respectively.

ated using Eq. (4)) achieves competitive (also best) performances in most of the evaluation settings.

4.6. A Deeper Understanding of GRepQ Features

Choice of Perceptual Similarity Measures in the Low-Level Feature Encoder: We compare different popular perceptual similarity measures such as SSIM [30], MS-SSIM [31], FSIM [39], LPIPS [40] and GMSD [33] used in the low-level feature encoder in Tab. 4. The low-level encoders are trained using these measures under similar training settings. We also train an encoder without any similarity measure (denoted by None) to show a need for quality-aware contrastive learning. In this case, all the other distorted versions of an image are treated as negatives, while the augmented version is the only positive. In Tab. 4, we show the low-level encoder’s data-efficient performances on KonIQ. We see that FSIM outperforms all other measures. We note that the superior performance of FSIM in this context is consistent with its superior performance as an FR-IQA metric across multiple datasets.

Analyzing High-Level Feature Representations: We analyze the impact of our group-contrastive learning in improving the high-level quality representations. For this analysis, we identify extremely good and extremely bad quality images based on mean opinion score (MOS) greater

			
(a)	(b)	(c)	(d)
28.97	24.49	75.63	65.15
40.56	25.43	62.04	61.29
27.66	39.40	72.33	49.85

Human Opinion Score
GRepQ_D (LL) Prediction
GRepQ_D (HL) Prediction

Figure 4. Demonstrating the complementarity of high and low-level model predictions. Images with different MOS from the KonIQ-10K database are listed. Low and high-level model predictions are mentioned below their respective MOS. Predictions that agree with human opinions are marked with green while erroneous predictions are marked with red.

than 75 or less than 25 respectively on the combined CLIVE and KonIQ datasets. We show the feature representations of the CLIP model in Fig. 3a and those of our model in Fig. 3b. We see that our learned representations are better separable between the higher and lower-quality images. This leads to the superior performance of our high-level model when compared to CLIP-IQA.

Complementarity of High and Low-Level Features: We present a qualitative and quantitative analysis of the complementarity of representations from both encoders. We show examples of when the two models outperform each other in Fig. 4. For instance, Fig. 4a shows that the low-level model makes an erroneous prediction since only the object in focus is blurred, but the background is relatively clean. Fig. 4d shows that the image does not contain enough contextual information for the high-level model to make an accurate prediction. We also perform an error-based feature complementarity analysis in Fig. 5. In particular, we compute the absolute error between the MOS predicted by the high and low-level models and the true MOS and show them in four quadrants. We see several examples where one of the models performs much better than the other. This shows that the models have complementary behavior in many examples.

Limitations: In the low-data setting, the low-level model does not perform as well as the high-level model on in-the-wild datasets. Since the low-level model is more suited to capture varied distortion levels rather than content, synthetic datasets benefit more from this model. Secondly, the high-level model uses fixed prompts and can be further improved through prompt engineering or tuning.

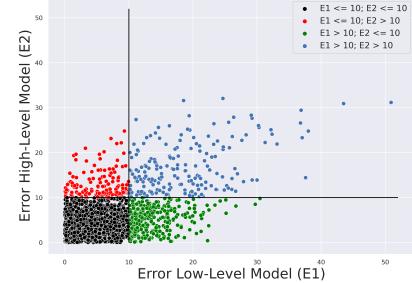


Figure 5. Analysis of absolute error in quality predictions from high and low-level models on the KonIQ-10K database.

5. Concluding Remarks

We design generalizable low-level and high-level quality representations that enable IQA in a data-efficient setting. Specifically, we learn low-level features using a novel quality-aware contrastive learning strategy that is distortion-agnostic. Secondly, we present a group-contrastive learning framework that learns to elicit semantic-based high-level quality information from images. We show that both sets of representations lead to accurate prediction of quality scores in both the data-efficient and zero-shot settings on diverse datasets. This demonstrates the generalizability of our learned features. Future advances in self-supervised learning and quality-specific prompt engineering could be used to further enhance the generalizability of models for data-efficient NR IQA.

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

References

- [1] Shahrukh Athar and Zhou Wang. Degraded reference image quality assessment. *IEEE Transactions on Image Processing*, 2023. 4
- [2] 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. 2, 6, 7
- [3] 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. 2
- [4] Lei Chen, Le Wu, Zhenzhen Hu, and Meng Wang. Quality-aware unpaired image-to-image translation. *IEEE Transactions on Multimedia*, 21(10):2664–2674, 2019. 1, 2, 3
- [5] 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, 2015. 6
- [6] 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. 2, 6
- [7] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision (ECCV) 2020*, pages 633–651. Springer International Publishing, 2020. 6
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [9] 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. 6
- [10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 4
- [11] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016. 2
- [12] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010. 6
- [13] Manling Li, Ruochen Xu, Shuhang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 4
- [14] 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. 2
- [15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5
- [16] 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. 1, 2, 3, 6
- [17] 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
- [18] 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, 7
- [19] 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
- [20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
- [21] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 4
- [22] Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 1, 3, 6, 7
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 6
- [24] 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
- [25] 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 (CVPR)*, pages 5846–5855, June 2023. 1, 2, 3, 6
- [26] 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. 6
- [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

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 2, 6
- [28] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In AAAI, 2023. 2, 4, 7
- [29] Jing Wang, Haotian Fan, Xiaoxia Hou, Yitian Xu, Tao Li, Xuechao Lu, and Lean Fu. Mstraq: No reference image quality assessment based on swin transformer with multi-stage fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1269–1278, 2022. 2
- [30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4, 7
- [31] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 4, 7
- [32] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *Computer Vision–ECCV 2022: 17th European Conference, Proceedings*, pages 538–554. Springer, 2022. 3
- [33] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2013. 4, 7
- [34] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 6
- [35] 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
- [36] 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. 5
- [37] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021. 2
- [38] 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, 7
- [39] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assess-
ment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 4, 7
- [40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4, 7
- [41] 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. 2
- [42] 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, 6
- [43] Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, and Jiashi Feng. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. *Advances in Neural Information Processing Systems*, 34:29848–29860, 2021. 4
- [44] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 2