This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ARNIQA: Learning Distortion Manifold for Image Quality Assessment

Lorenzo Agnolucci Leonardo Galteri Marco Bertini Alberto Del Bimbo

University of Florence - Media Integration and Communication Center (MICC)
Florence, Italy

[name.surname]@unifi.it

Abstract

No-Reference Image Quality Assessment (NR-IQA) aims to develop methods to measure image quality in alignment with human perception without the need for a high-quality reference image. In this work, we propose a self-supervised approach named ARNIQA (leArning distoRtion maNifold for Image Quality Assessment) for modeling the image distortion manifold to obtain quality representations in an intrinsic manner. First, we introduce an image degradation model that randomly composes ordered sequences of consecutively applied distortions. In this way, we can synthetically degrade images with a large variety of degradation patterns. Second, we propose to train our model by maximizing the similarity between the representations of patches of different images distorted equally, despite varying content. Therefore, images degraded in the same manner correspond to neighboring positions within the distortion manifold. Finally, we map the image representations to the quality scores with a simple linear regressor, thus without fine-tuning the encoder weights. The experiments show that our approach achieves state-of-the-art performance on several datasets. In addition, ARNIQA demonstrates improved data efficiency, generalization capabilities, and robustness compared to competing methods. The code and the model are publicly available at https://github.com/miccunifi/ARNIQA.

1. Introduction

Image Quality Assessment (IQA) refers to the computer vision task of automatically evaluating the quality of images with a high correlation with human judgments. Specifically, No-Reference IQA (NR-IQA) focuses on devising methods that can be used when a high-quality reference image is unavailable. NR-IQA finds diverse applications in industries and research domains, including image restoration [14, 33], captioning [4], and multimedia streaming [1].

Although supervised learning techniques have shown notable advances in NR-IQA [7, 29, 30, 42], their effectiveness is based on labeled data. Acquiring such annotations is challenging and resource-intensive, given the requirement for a substantial number of ratings to obtain dependable mean opinion scores. For example, the KADID dataset [15], which comprises 10125 images synthetically degraded with several distortion types, required approximately 300K annotations. This inherent dependence on labeled data hampers the scalability and broad applicability of supervised approaches.

More recently, several works based on self-supervised learning [2, 3, 9] have been presented [19, 26, 44]. These methods involve the pre-training of an encoder on unlabeled data with a contrastive loss. Then, the image representa-

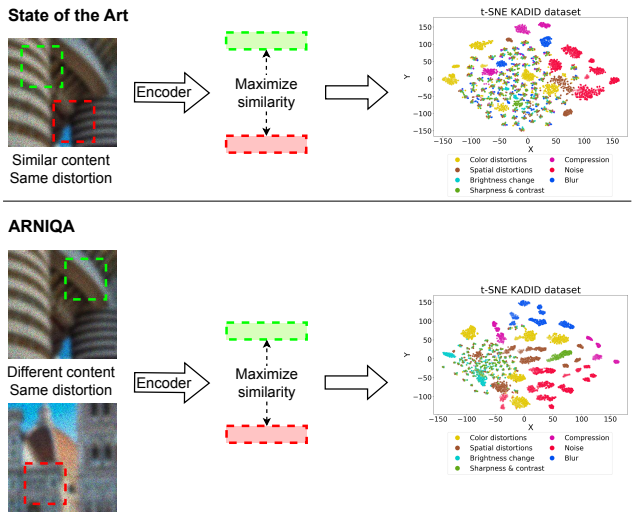


Figure 1. Comparison between our approach and the State of the Art for NR-IQA. While the SotA maximizes the similarity between the representations of crops from the same image, we propose to consider crops from different images degraded equally to learn the image distortion manifold. The t-SNE visualization of the embeddings of the KADID dataset [15] shows that, compared to Re-IQA [26], ARNIQA yields more discernable clusters for different distortions. In the plots, a higher alpha value corresponds to a stronger degradation intensity.

tions are mapped to the quality scores with a fine-tuning of the encoder weights [44] or by just using a linear regression [19, 26]. For example, Re-IQA [26] generates image representations by concatenating low-level and high-level features obtained through a quality-aware and content-aware encoder, respectively. Existing methods involve maximizing the similarity between the representations of two crops of the same distorted image. Therefore, since crops share similar visual information, the model is exposed to content-dependent degradation patterns, which inevitably leads to content-dependent embeddings. The upper part of Fig. 1 shows the t-SNE visualization [31] of the embeddings of the KADID [15] dataset generated by Re-IQA. We notice that the representations related to some types of distortions, *e.g.* blur, are scattered across the space, without being confined into separable clusters. This result stems from the training strategy, as images distorted equally may correspond to different representations due to their diverse content.

In contrast, we propose a self-supervised approach, named ARNIQA (leArning distoRtion maNifold for Image Quality Assessment)[1], to model the image distortion manifold so that images that exhibit similar degradation patterns correspond to resembling embeddings, despite varying content. We refer to *image distortion manifold* as the continuous space of all the possible degradations to which an image can be subjected. Different regions along this manifold represent various types and degrees of degradation. For example, images showing distinct blur and noise patterns lie in different areas of the manifold. Similarly, images subjected to varying compression rates using the same algorithm correspond to diverse regions within the space. Such a distortion manifold represents image quality in an intrinsic manner. In fact, images that show similar degrees and patterns of degradation are prone to be perceived as having similar quality. At the same time, images exhibiting comparable levels and types of distortion correspond to similar positions in the manifold. Therefore, to map the representation in the manifold to a quality score, it is sufficient to train a simple linear regressor, without the need of fine-tuning the encoder weights. Moreover, by focusing on the inherent distortions within images rather than being dependent on their content, our approach significantly reduces the complexity of the learning process [29]. Given two different images that are degraded in the same way, our training strategy consists of extracting a crop from each of them and maximizing the similarity between their representations. At the same time, we maximize the distance from the embeddings of other images degraded in a different manner. In this way, our model learns to recognize image degradation despite varying content. To improve contrastive learning performance, we present a strategy to ensure the presence of hard negative examples within each batch by also considering half-scale images. To train our model, we propose to synthetically distort pristine images with a wide variety of degradations. To this end, we introduce an image degradation model that produces random compositions of consecutively applied distortions. Our degradation model is capable of generating about 100 times more possible distortion compositions than existing approaches. In the lower section of Fig. 1 we report the t-SNE visualization of the embeddings of the KADID dataset obtained by ARNIQA. We notice that compared to Re-IQA [26], our approach produces more easily distinguishable clusters for different types of degradation, thanks to our training strategy.

Extensive experiments show that ARNIQA achieves state-of-the-art performance on datasets with both synthetic and authentic (*i.e.* real-world) distortions. Furthermore, since our learning process is less complex, the proposed approach proves to be more data efficient than competing methods, requiring only up to 0.5% of the training images compared to the competitors. In addition, cross-dataset evaluation and the gMAD competition [18] demonstrate that ARNIQA has better generalization capabilities and is more robust than the baselines.

We summarize our contributions as follows:

1. We propose ARNIQA, a self-supervised approach for learning the image distortion manifold. By maximizing the similarity between the embeddings of different images distorted equally, we make the encoder generate similar representations for images exhibiting the same degradation patterns regardless of their content;
2. We introduce an image degradation model that randomly assembles ordered sequences of distortions, with $1.9 \cdot 10^9$ distinct possible compositions, for synthetically degrading images;
3. ARNIQA achieves state-of-the-art performance on NR-IQA datasets with both synthetic and authentic distortions while showing enhanced data efficiency, generalization capabilities, and robustness.

2. Related Work

2.1. No-Reference Image Quality Assessment

Due to its importance in both industry and computer vision tasks, No-Reference Image Quality Assessment (NR-IQA) has been an active area of research for several years [6, 12, 19, 20, 24, 26, 30, 32, 34].

Traditional methods [20–22, 35, 41], such as BRISQUE [20] and NIQE [21], rely on the extraction of handcrafted features from the images. Subsequently, they employ a regression model to predict quality scores. Codebook-based approaches, such as CORNIA [36] and HOSA [34], build a visual codebook from local patches to obtain quality-aware features. In recent years, methods using supervised

[1] Pronounced as the English word "*arnica*" (/ˈɑːr.nɪk.ə/).

learning achieved a significant boost in performance in NR-IQA [7, 12, 29, 30, 37, 39, 42]. For example, HyperIQA [30] presents a self-adaptive hypernetwork that distinguishes content understanding from quality predictions. The most similar to our work is Su *et al*. [29], which learns the image distortion manifold in a supervised manner on IQA datasets. Given that it requires distortion-specific information for training, it cannot be used for NR-IQA on datasets with authentic degradations. On the contrary, we model the distortion manifold using unlabeled data with self-supervised learning. Due to their dependence on ground-truth quality scores for training, supervised methods suffer from the scarcity of labeled data for IQA, which are expensive and time-consuming to collect.

Recently, self-supervised learning has emerged as a promising technique for NR-IQA [19, 26, 44]. Self-supervised methods train an encoder on unlabeled data with a contrastive loss and then use its image representations to obtain the final quality scores, either by fine-tuning the model weights [44] or using a linear regressor [19,44]. QPT [44] proposes a quality-aware contrastive loss based on the assumption that the quality of patches is similar for the same distorted image but differs as the image or the degradations vary. CONTRIQUE [19] models the representation learning problem as a classification task, where each class is composed of images degraded equally. Re-IQA [26] trains a quality-aware and a content-aware encoder to generate low-level and high-level representations, respectively. Existing methods are based on maximizing the similarity between the representations of crops of the same distorted image. In contrast, we maximize the similarity between the embeddings of patches that belong to different images that were degraded equally, regardless of varying content, to model the image distortion manifold. After training, we freeze the encoder weights and map the image representation to the final quality scores with a simple linear regressor.

2.2. Image Degradation Models

Image degradation models aim to synthetically distort images so that the degradation patterns closely resemble those found in real-world scenarios. They play an important role in both blind image restoration [28, 33, 38, 40, 43] and IQA [7,26,42,44]. Degradation models differ mainly in how many distinct types of distortion they consider and how they compose them. Specifically, the number of times and the order in which they apply the degradations. RealESRGAN [33] proposes a second-order degradation model, *i.e*. that performs the distortion process twice but with different parameters. The images are degraded sequentially with one distortion from each of 4 predefined groups, always following the same order. Re-IQA [26] considers 25 distortion types but applies only one of them to each image, thus not studying combined degradation patterns. QPT [44] presents a second-order degradation model with skip and shuffle operations. It takes into account 3 distortion groups comprising a total of 9 degradation types. In contrast, we introduce an image degradation model that randomly composes ordered sequences of consecutively applied distortions. Given that we consider 24 distortion types divided into 7 groups, we obtain 100 times more possible compositions than existing methods. We rely on our degradation model to synthetically degrade the training images.

3. Proposed Approach

Our approach relies on the SimCLR [2] framework to train a model composed of a pre-trained ResNet-50 [2] encoder and a 2-layer MLP projector that reduces the dimension of the features. We employ unlabeled pristine images distorted with the proposed degradation model for self-supervised learning. After training, we discard the projector and consider the encoder output features as the image representations. Finally, we freeze the encoder and train a linear regressor on top of it to obtain the quality score of an image from its representation.

3.1. Image Degradation Model

To effectively learn the image distortion manifold, during training our model must be exposed to a very wide range of diverse degradation patterns. Additionally, it is imperative to possess information about the nature and intensity of degradations within each image for self-supervised learning with a contrastive loss. To address these requirements, we propose to train our model using synthetically degraded images. To this end, we need to make two considerations. First, a broad spectrum of distortion types, spanning varying degrees of intensity, must be taken into account to create a rich collection of degradation patterns. Second, we also need to consider the case of multiple distortions applied at once to investigate how the degradations appear when combined together. Therefore, we introduce an image degradation model that randomly composes ordered sequences of consecutively applied distortions to generate images that exhibit a large variety of degradation patterns. Figure 2 shows an overview of the proposed degradation model.

We consider 24 distinct degradation types D divided into the 7 distortion groups $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_7\}$ defined by the KADID [15] dataset. Each distortion has $L = 5$ levels of intensity. See the supplementary material for more details on the specific degradation types. The distortion groups we consider are: 1) Brightness change; 2) Blur; 3) Spatial distortions; 4) Noise; 5) Color distortions; 6) Compression; 7) Sharpness & contrast. Each of them is defined as $\mathcal{G}_i = \{\dots, D^{ij}, \dots\}$, where $i \in \{1, \dots, 7\}$ is the index of the distortion group within $\mathcal{G}$ and $j \in \{1, \dots, |G_i|\}$ indicates the index of the degradation type within $\mathcal{G}_i$, with $|\cdot|$ that represents the cardinality.

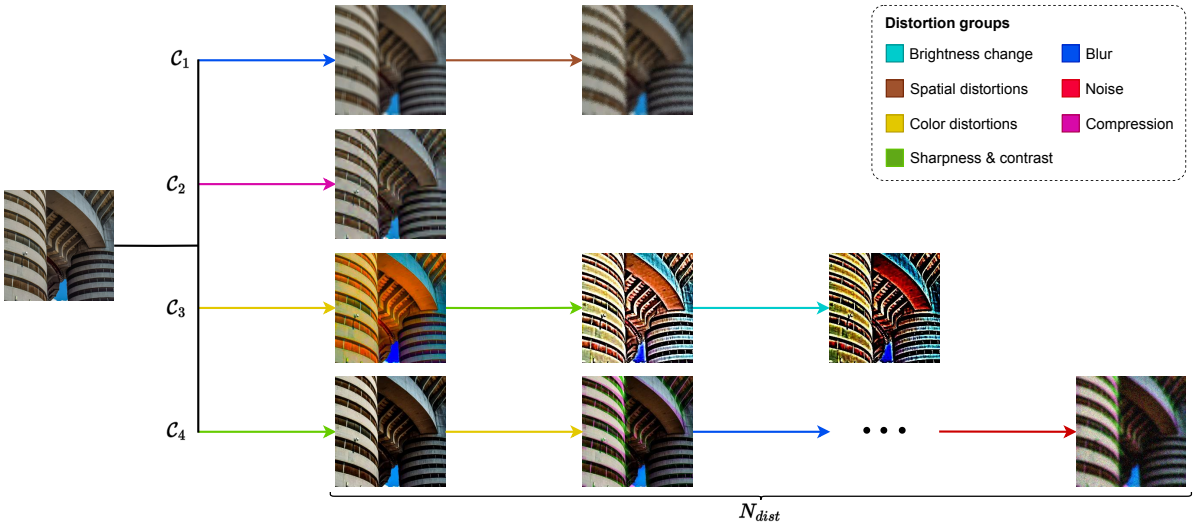


Figure 2. Overview of the proposed image degradation model. We randomly assemble distortion compositions $\mathcal{C}$, *i.e*. ordered sequences of distortions applied consecutively, to synthetically generate images with a wide variety of degradation patterns. Each distortion composition contains a maximum of N_{dist} degradations sampled from 7 distinct distortion groups.

Let $\underline{I}$ be a pristine image. Our aim is to obtain a randomly selected distortion composition $\mathcal{C}$, *i.e*. an ordered sequence of distortions that generates the degraded image I from $\underline{I}$. We define N_{dist} as the maximum number of different distortions that can be applied to $\underline{I}$. First, we randomly select a number $n_{dist} = \{1, \dots, N_{dist}\}$ of distortions. Then, we sample n_{dist} distortion types with a uniform distribution from n_{dist} different degradation groups. In other words, as in [40, 43], for each distortion composition, there can be a maximum of one degradation for each group. Finally, we shuffle the order of the selected distortions and sample a level of intensity for each of them with a given probability distribution. In the end, we obtain a distortion composition $\mathcal{C} = \{\dots, D_k^{ij_l}, \dots\}$ where $k \in \{1, \dots, n_{dist}\}$ is the distortion index within $\mathcal{C}$, i and j are defined above and $l \in \{1, \dots, L\}$ is the intensity level. Since we want our model to also have access to pristine images during training, we define a hyperparameter p_{prist} and apply the degradation composition to an image with probability $1 - p_{prist}$. Compared to the 9 types of distortion considered by QPT [44], we take significantly more degradation patterns into account. Moreover, contrary to Re-IQA [26], we consecutively apply multiple distortions to the same image, thus studying the effect of their combination.

Applying multiple distortions with a high level of intensity to the same image usually results in a complete disruption of its content. Although our aim encompasses learning areas of the distortion manifold corresponding to very severe degradations, our primary focus resides in regions that are more likely to be related to real-world scenarios. These regions correspond to degradation compositions that alter the content of the image, but not so severely as to make it unrecognizable. In fact, when evaluating the performance of an image restoration model or assessing the quality of a picture uploaded to social platforms, it is unlikely that the images under consideration would be degraded to the point of rendering their content indistinguishable. Therefore, we propose to sample the intensity level of each distortion with a Gaussian distribution with mean 0 and standard deviation σ. In this way, lower intensity levels are more likely to be sampled, leading to less severe degradation compositions. Thus, we model regions of the distortion manifold corresponding to degradations most probably corresponding to real-world scenarios in a more fine-grained manner.

Ultimately, our image degradation model is capable of yielding a large variety of distinct distortion compositions. Specifically, the number of possible ways in which the degradations can be assembled is given by:

$$\sum_{m=1}^{N_{dist}} m! L^m \left[\sum_{i=1}^{7} |G_i| \sum_{j=2}^{7} |G_j| \dots \sum_{k=m}^{7} |G_k| \right] \quad (1)$$

As an example, with $N_{dist} = 4$, we obtain $1.9 \cdot 10^9$ possible compositions, which are about 100 times more than the $2 \cdot 10^7$ available with the model proposed in QPT [44].

3.2. Training Strategy

Existing self-supervised NR-IQA methods, such as Re-IQA [26], extract two crops from a single distorted image. Then, they maximize the similarity between their representations. Since the crops share similar visual information, the models learn content-dependent distortion features. In contrast, we maximize the similarity of the representations of crops from two different images with completely diverse content but distorted in the same manner. This way, the en-

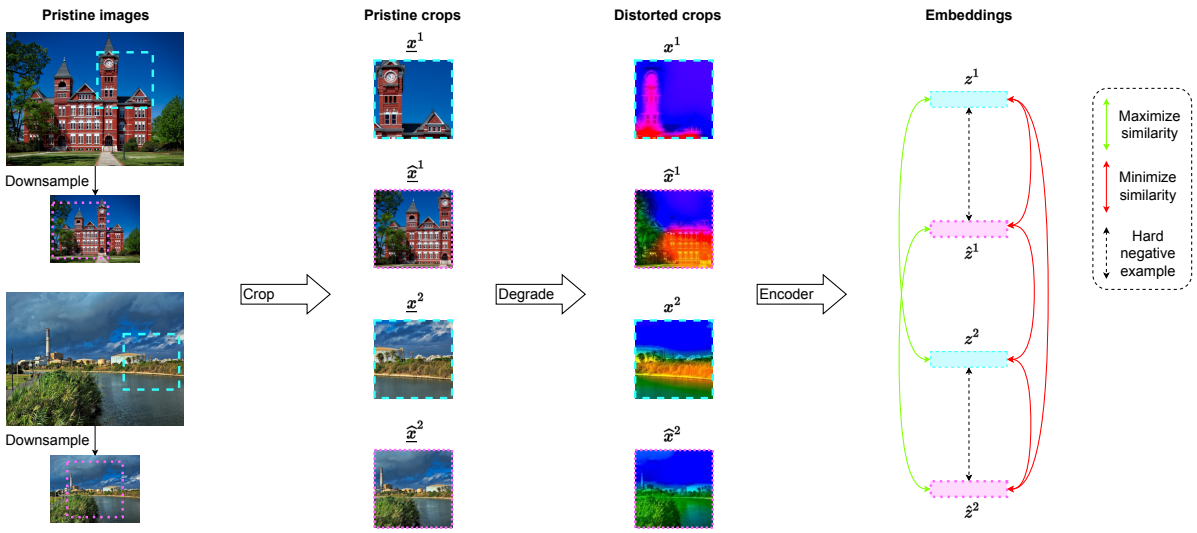


Figure 3. Overview of the proposed training strategy. Given two pristine images, we extract two crops and degrade them equally. Then, we maximize the similarity of their embeddings. At the same time, we minimize the similarity with respect to the embeddings of degraded crops from the half-scale versions of the original images. These embeddings constitute hard negative examples for the representations of the full-scale images since they share similar content and differ only for a downsample distortion. Notice how the original and half-scale degraded crops differ despite being degraded in the same way due to the downsampling operation.

coder learns to model the distortion manifold, thus yielding resembling embeddings for images that exhibit similar degradation patterns, despite varying content. Figure 3 shows an overview of our training strategy.

Our approach is based on SimCLR [2]. SimCLR is a framework for self-supervised learning based on a contrastive loss. Given a training example (*i.e.* an image), SimCLR constructs a positive pair for the contrastive loss by generating two views of the original image with random augmentation techniques. The training process aims to maximize the similarity between the representations of the two views of each training example while maximizing the distance between the embeddings of all the other augmented images in the batch. Thus, the number of examples in each batch is doubled. Intuitively, given that our goal is to learn the image distortion manifold, we can interpret a specific distortion composition as a training example. Therefore, by using it to degrade two different images, we are generating the two views considered by SimCLR. Formally, let $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_B\}$ be a batch of distortion compositions obtained with the proposed degradation model, where B is the batch size. Similarly, let $\mathcal{B}^1 = \{\underline{x}_1^1, \dots, \underline{x}_B^1\}$ and $\mathcal{B}^2 = \{\underline{x}_1^2, \dots, \underline{x}_B^2\}$ be two batches of pristine images randomly selected from the training dataset. For each pair $(\underline{x}_i^1, \underline{x}_i^2)$ where $i \in \{1, \dots, B\}$, we extract a random crop from each image and employ $\mathcal{C}_i$ to obtain the degraded version (x_i^1, x_i^2). Each pair constitutes the two views of the SimCLR framework, and their embeddings represent a positive pair in the contrastive loss.

However, since the proposed degradation model has a very large number of possible compositions, the given batch of distortion compositions $\mathcal{C}$ could lead to considerably different image pairs. In that case, it would be trivial for the model to discriminate between the different examples, making the learning process less effective. To avoid this issue, we propose a strategy to ensure the presence of hard negative examples in each batch, which is known to enhance contrastive learning [11, 25]. Given an image pair $(\underline{x}_i^1, \underline{x}_i^2)$ defined as above, we downsample the images to half size before cropping, resulting in $(\underline{\widehat{x}}_i^1, \underline{\widehat{x}}_i^2)$. After applying $\mathcal{C}_i$, we obtain the distorted image pair $(\widehat{x}_i^1, \widehat{x}_i^2)$. Given that the downsampling operation fundamentally reduces the pixel count, it inherently results in information loss and thus can be viewed as a degradation. Therefore, this process can be likened to prepending a downsampling degradation to each distortion composition $\mathcal{C}_i$. Finally, we apply this technique to all the image pairs and add the new B pairs to the batch, thereby doubling the batch size and the number of negative examples, which improves the performance of contrastive learning [2, 8]. Moreover, since we use all the images both at full-scale and half-scale, the size of the training dataset is also doubled. Thanks to our strategy, the model always has to discriminate between the two examples x_i^1 and $\widehat{x}_i^1$, which mutually serve as hard negatives for each other. Indeed, they share similar content as they are crops taken from the same image at two different scales, and their degradation differs only for a downsample distortion. Therefore, by minimizing the similarity between the representations of x_i^1 and $\widehat{x}_i^1$, our model learns to discriminate between images with slightly different degradation, even if they share similar content. The same considerations apply for x_i^2 and $\widehat{x}_i^2$. CONTRIQUE [19] considers images at half-scale as well

but regards them as positive examples belonging to the same distortion class. On the contrary, we treat the half-scale resolution crops as challenging negative examples, since they actually differ for a single distortion.

Formally, let $f(\cdot)$ be the ResNet-50 encoder and $g(\cdot)$ the 2-layer MLP projector that we use for dimensionality reduction. Then, given an image $x \in \mathbb{R}^{3\times H\times W}$, with H and W representing respectively its height and width, we compute its representation z with:

$$h = f(x) \in \mathbb{R}^C, \qquad z = g(h) = g(f(x)) \in \mathbb{R}^D \tag{2}$$

where C and D are the number of channels of the encoder and the projector, respectively. First, we generate the embeddings of all the views, both at full- and half-scale. Then, following SimCLR, we employ the NT-Xent contrastive loss [2] for training. To this end, we define the loss terms:

$$\begin{aligned} \ell_i^{1,2} &= -\log \frac{\gamma_{i,i}^{1,2}}{\sum_{k=1}^{B}\left[\gamma_{i,k}^{1,2} + \gamma_{i,\widehat{k}}^{1,2} + \gamma_{i,\widehat{k}}^{1,1}\right] + \sum_{k\neq i}^{B}\gamma_{i,k}^{1,1}} \\ \widehat{\ell}_i^{2,1} &= -\log \frac{\gamma_{\widehat{i},\widehat{i}}^{2,1}}{\sum_{k=1}^{B}\left[\gamma_{\widehat{i},\widehat{k}}^{2,1} + \gamma_{\widehat{i},k}^{2,1} + \gamma_{\widehat{i},k}^{2,2}\right] + \sum_{k\neq i}^{B}\gamma_{\widehat{i},\widehat{k}}^{2,2}} \end{aligned} \tag{3}$$

where $\gamma_{i,\widehat{k}}^{1,2} = e^{(cos(z_i^1, \widehat{z}_k^2)/\tau)}$ and $cos(\cdot)$ and τ represent the cosine similarity and a temperature hyperparameter, respectively. Hence, the overall training loss is given by:

$$\mathcal{L} = \frac{1}{4B}\sum_{i=1}^{B}\left[\ell_i^{1,2} + \ell_i^{2,1} + \widehat{\ell}_i^{1,2} + \widehat{\ell}_i^{2,1}\right] \tag{4}$$

Intuitively, the loss maximizes the similarity between the representation of each view and the corresponding one, while maximizing the distance to all the other views, both at full-scale and half-scale. Therefore, we consider 2 (views) $\times$ 2 (scales) $\times$ B (batch size)$=4B$ elements in total.

After training, our model has learned a distortion manifold and hence generates similar embeddings for images degraded in the same way, regardless of their content.

4. Experimental Results

4.1. Implementation Details

We train our model for 10 epochs using a stochastic gradient descent optimizer with momentum 0.9 and weight decay $1e{-}4$. Starting from a learning rate of $1e{-}3$, we employ a cosine annealing with warm restarts scheduler [16]. Differently from [19, 26], we start from a pre-trained ResNet-50 encoder and fine-tune its weights during training. The encoder and the projector have a number of channels C and D of 2048 and 128, respectively. During training, we use a patch size of 224, a temperature τ of 0.1, and a batch size of 16. Regarding the image degradation model, we set the maximum number of distortions N_{dist} to 4, the probability of using pristine images p_{prist} to 0.05, and the standard deviation of the Gaussian distribution σ to 2.5.

4.2. Datasets

We employ the 140K pristine images from the KADIS dataset [15] for training, discarding the 700K degraded ones it provides. In fact, we use our image degradation model to obtain the degraded versions of the pristine images.

We test ARNIQA on datasets with both synthetic and authentic distortions. These consist of collections of degraded images labeled with subjective opinions of picture quality in the form of Mean Opinion Score (MOS). We consider four synthetically degraded datasets: LIVE [27], CSIQ [13], TID2013 [23], and KADID [15]. LIVE comprises 779 images degraded with 5 types of distortion at 5 levels of intensity, with 29 reference images as the base. CSIQ, on the other hand, stems from 30 reference images, each undergoing 6 types of distortions at 5 levels of intensity, yielding 866 images. TID2013 and KADID contain 3000 and 10125 images synthetically degraded with 24 and 25 types of distortion at 5 different levels of intensity, starting from 25 and 81 reference images, respectively. Regarding the datasets with authentic distortions, we consider FLIVE [37] and SPAQ [5]. FLIVE is the largest existing dataset for NR-IQA, comprising nearly 40K real-world images. SPAQ contains 11K high-resolution images captured by several mobile devices. Similar to [5, 19], for evaluation, we resize the SPAQ images so that the shorter side is 512.

4.3. Evaluation Protocol

To evaluate the performance, we employ Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) to measure prediction monotonicity and accuracy, respectively.

Following [19, 26], we randomly divide the datasets into 70%, 10%, and 20% splits corresponding to training, validation, and test sets, respectively. Splits are selected based on reference images to ensure no overlap of contents. We employ the ground-truth MOS scores of the training split to train a Ridge regressor [10] with an L2 loss. Note that we do not perform any fine-tuning of the encoder weights during the evaluation. Similarly to [19, 26], we use the validation split to identify the regularization coefficient of the regressor via a grid search over values in the range $\left[10^{-3}, 10^{3}\right]$. During testing, we compute the image features at full-scale and half-scale and concatenate them to obtain the final representation, as in [19]. Similarly to [44], we take the four corners and the center crops at both scales and average the corresponding predicted quality scores to obtain the final one. To remove any bias in the selection of the training set, we repeat the training/test procedure 10 times and report the median results. Given the large size of the dataset, for

Method	Type	Synthetic Distortions								Authentic Distortions				Average	
		LIVE		CSIQ		TID2013		KADID		FLIVE		SPAQ			
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [20]	Handcrafted	0.939	0.935	0.746	0.829	0.604	0.694	0.528	0.567	0.288	0.373	0.809	0.817	0.652	0.703
NIQE [21]		0.907	0.901	0.627	0.712	0.315	0.393	0.374	0.428	0.211	0.288	0.700	0.709	0.522	0.572
CORNIA [36]	Codebook	0.947	0.950	0.678	0.776	0.678	0.768	0.516	0.558	–	–	0.709	0.725	–	–
HOSA [34]		0.946	0.950	0.741	0.823	0.735	0.815	0.618	0.653	–	–	0.846	0.852	–	–
DB-CNN [42]	Supervised learning	0.968	0.971	0.946	0.959	0.816	0.865	0.851	0.856	0.554	0.652	0.911	0.915	0.841	0.870
HyperIQA [30]		0.962	0.966	0.923	0.942	0.840	0.858	0.852	0.845	0.535	0.623	0.916	0.919	0.838	0.859
TReS [7]		0.969	0.968	0.922	0.942	0.863	0.883	0.859	0.858	0.554	0.625	–	–	–	–
Su *et al*. [29]		**0.973**	**0.974**	0.935	0.952	0.815	0.859	0.866	0.874	–	–	–	–	–	–
CONTRIQUE [19]	SSL + LR	0.960	0.961	0.942	0.955	0.843	0.857	**0.934**	**0.937**	0.580	0.641	0.914	0.919	0.862	0.878
Re-IQA [26]		0.970	0.971	0.947	0.960	0.804	0.861	0.872	0.885	**0.645**	**0.733**	**0.918**	**0.925**	0.859	0.889
ARNIQA	SSL + LR	0.966	0.970	**0.962**	**0.973**	**0.880**	**0.901**	0.908	0.912	0.595	0.671	0.905	0.910	**0.869**	**0.890**

Table 1. Comparison between the proposed approach and competing methods on datasets with synthetic and authentic distortions. Best and second-best scores are highlighted in bold and underlined, respectively. – denotes results not reported in the original paper. SSL and LR stands for self-supervised learning and linear regression, respectively.

Training	Testing	Method				
		HyperIQA	Su *et al*.	CONTRIQUE†	Re-IQA†	**ARNIQA**
LIVE	CSIQ	0.744	0.777	0.803	0.795	**0.904**
LIVE	TID2013	0.541	0.561	0.640	0.588	**0.697**
LIVE	KADID	0.492	0.506	0.699	0.557	**0.764**
CSIQ	LIVE	0.926	**0.930**	0.912	0.919	0.921
CSIQ	TID2013	0.541	0.550	0.570	0.575	**0.721**
CSIQ	KADID	0.509	0.515	0.696	0.521	**0.735**
TID2013	LIVE	0.876	0.892	**0.904**	0.900	0.869
TID2013	CSIQ	0.709	0.754	0.811	0.850	**0.866**
TID2013	KADID	0.581	0.554	0.640	0.636	**0.726**
KADID	LIVE	**0.908**	0.896	0.900	0.892	0.898
KADID	CSIQ	0.809	0.828	0.773	0.855	**0.882**
KADID	TID2013	0.706	0.687	0.612	**0.777**	0.760

Table 2. Cross-dataset evaluation results for the SRCC metric. † denotes results evaluated by us with the official pre-trained models. Best scores are highlighted in bold.

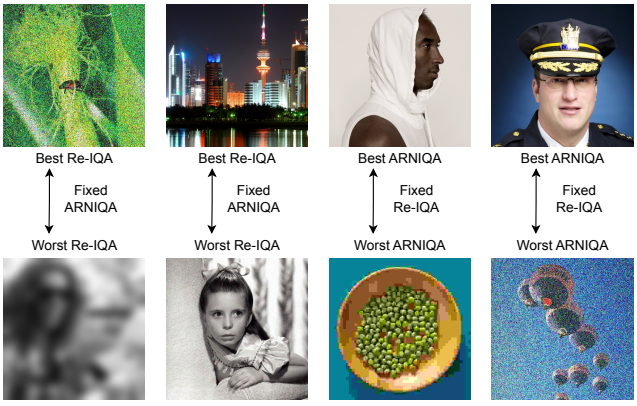
FLIVE we use only the official splits [37].

4.4. Results

In Tab. 1 we compare the performance of the proposed approach with other state-of-the-art methods. ARNIQA achieves competitive performance for both synthetic and authentic distortions and obtains the best results on average. In particular, we notice how our method outperforms Su *et al*. [29], which also aims to learn the distortion manifold but in a supervised manner and directly on IQA datasets. Furthermore, contrary to our approach, Su *et al*. cannot be evaluated on datasets with authentic degradations, as it requires distortion-specific information for training. Compared to other self-supervised approaches, namely CONTRIQUE [19] and Re-IQA [26], ARNIQA achieves comparable or better performance. However, our method is significantly more data-efficient than the competitors. Indeed, we employ 140K (training dataset) $\times$ 2 (scales) $\times$ 10 (epochs) = 2.8M images for training. In contrast, doing similar computations, we get that CONTRIQUE uses 65M images, while Re-IQA requires 512M and 38M images for the content and quality encoders, respectively. See the supplementary material for more details. Therefore, ARNIQA achieves state-of-the-art performance while requiring only up to 0.5% of the data compared to competing methods. The reason is that focusing solely on the degradation patterns within images reduces the complexity of the learning process compared to depending on image content as well, as also observed by [29].

We evaluate the generalization capabilities of our model by measuring cross-dataset performance. We train the regressor on the whole training dataset and then use it to obtain the quality predictions on the testing dataset. We report the results for the SRCC metric in Tab. 2. ARNIQA significantly outperforms all the competing methods. In particular, the proposed approach achieves the largest improvements compared to the baselines when training on a dataset comprising few distortion types (*e.g*. CSIQ) and testing on one with a large variety of different degradations (*e.g*. TID2013). We hypothesize that the reason behind this outcome is that our method makes the encoder model the regions of the distortion manifold that correspond to distinct types of degradations in a consistent way. In other words, the mapping from the distortion manifold to the quality scores is consistent across different types of degradation. Therefore, a regressor trained by mapping only some specific regions of the manifold – *i.e*. considering only a few different distortions – to quality scores behaves well even when used on unseen degradation types. We will study this phenomenon more thoroughly in future work.

To evaluate the robustness of the model, we conduct the group maximum differentiation (gMAD) competition [18] between ARNIQA and Re-IQA on the Waterloo Exploration Database [17], a dataset with synthetically degraded images without MOS annotations. We fix one model to act as the defender and we group its quality predictions into several levels. Then, the other model functions as the at-



(a) (b) (c) (d)

Figure 4. gMAD competition results between ARNIQA and Re-IQA [26]. (a) and (b): Fixed ARNIQA at a low- and high-quality level, respectively. (c) and (d): Fixed Re-IQA at a low- and high-quality level, respectively.

tacker by identifying the image pairs within each level that differ the most in terms of quality. Therefore, for a model to be robust, the selected image pairs should exhibit similar quality when functioning as the defender and show a noticeable quality difference when acting as the attacker. We show the results in Fig. 4. When we fix ARNIQA (Figs. 4a and 4b), Re-IQA is unable to identify image pairs showing an obvious quality difference. On the contrary, when ARNIQA acts as the attacker (Figs. 4c and 4d), it manages to spot the failures of Re-IQA by finding image pairs that clearly have significantly different quality. Thus, the proposed approach proves to be more robust than Re-IQA.

4.5. Ablation Studies

Image Degradation Model We conduct ablation studies on our image degradation model: 1) *RealESRGAN*: we replicate the degradation model of RealESRGAN [33]; 2) *w/o gaussian*: we sample the intensity level of each degradation with a uniform distribution instead of a Gaussian one; 3) $N_{dist} = 1$ and 4) $N_{dist} = 7$: we reduce and increase the maximum number of distorsions N_{dist}, respectively.

The upper part of Tab. 3 shows the results for the SRCC metric. We observe that the RealESRGAN degradation model obtains poor performance on all the datasets. Expectedly, considering only 4 distortion groups and applying them always in the same order limits the variety of degradation patterns the model is exposed to during training, hampering the learning process. We notice that sampling the levels of intensity of the degradations with a uniform distribution degrades the performance compared to using a Gaussian one. This is because it leads to more coarse modeling of the regions of the manifold corresponding to degradations more likely to be related to those found in real-world scenarios. Finally, the variants of the degradation model with $N_{dist} = 1$ and $N_{dist} = 7$ generate images that respectively contain no combined degradation patterns and too strong distortions, thus hampering the training process.

Method	LIVE	CSIQ	TID2013	KADID	Average
RealESRGAN	0.926	0.896	0.616	0.727	0.791
w/o gaussian	0.965	0.953	0.866	**0.920**	0.926
$N_{dist} = 1$	0.966	0.957	0.857	0.916	0.924
$N_{dist} = 7$	**0.970**	0.957	0.868	0.902	0.924
same image	0.940	0.863	0.721	0.775	0.825
w/o HN	0.966	0.960	0.851	0.908	0.921
ARNIQA	0.966	**0.962**	**0.880**	0.908	**0.929**

Table 3. Ablation studies results for the SRCC metric. Best scores are highlighted in bold.

Training Strategy We perform ablation studies on our training strategy: 1) *same image*: we extract the crops from the same image, instead of from two different ones; 2) *w/o HN*: we do not employ our strategy to obtain hard negative examples and, for a fair comparison, we double the batch size to have the same number of negative examples.

We report the results for the SRCC metric in the lower section of Tab. 3. We observe that extracting two crops from the same degraded image leads to poor performance. Even if it proved to be a viable technique to achieve state-of-the-art results in NR-IQA, it requires more convoluted approaches compared to ours, such as considering multiple loss terms [44] or two different encoders [26]. Furthermore, we notice that our strategy to guarantee the presence of hard negatives in every batch improves the results compared to using the same number of randomly sampled negative examples, as expected for contrastive learning [11, 25].

5. Conclusion

In this work, we present a self-supervised approach, named ARNIQA, to learn the image distortion manifold for NR-IQA. First, we introduce an image degradation model that randomly assembles ordered sequences of distortions, with about 100 times more possible compositions than competing methods. Second, we propose a training strategy that maximizes the similarity between the embeddings of crops belonging to distinct images degraded equally, regardless of their content. This way, we model the distortion manifold so that a simple linear regressor can effectively map image representations to quality scores. The experiments show that ARNIQA achieves state-of-the-art performance on datasets with both synthetic and authentic distortions. Also, our method exhibits enhanced generalization capabilities, data efficiency, and robustness compared to the baselines. In future work, we will study how our learned distortion manifold can be used for blind image restoration.

Acknowledgments This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Perceptual quality improvement in videoconferencing using keyframes-based gan. *IEEE Transactions on Multimedia*, 2023. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3, 5, 6

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[4] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3646–3656, 2020. 1

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

[6] Leonardo Galteri, Lorenzo Seidenari, Pietro Bongini, Marco Bertini, and Alberto Del Bimbo. Lanbique: Language-based blind image quality evaluation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2s):1–19, 2022. 2

[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 1220–1230, 2022. 1, 3, 7

[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. 5

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[10] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 6

[11] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 5, 8

[12] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 2, 3

[13] 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

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1

[15] 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. 1, 2, 3, 6

[16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[17] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016. 7

[18] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2016. 2, 7

[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. 1, 2, 3, 5, 6, 7

[20] 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, 7

[21] 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

[22] 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

[23] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Color image database TID2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*, pages 106–111. IEEE, 2013. 6

[24] Vishnu Prabhakaran and Gokul Swamy. Image quality assessment using semi-supervised representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 538–547, 2023. 2

[25] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020. 5, 8

[26] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 1, 2, 3, 4, 6, 7, 8

[27] 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

[28] Pranjay Shyam, Kyung-Soo Kim, and Kuk-Jin Yoon. Giqe: Generic image quality enhancement via nth order iterative degradation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2077–2087, 2022. 3

[29] Shaolin Su, Qingsen Yan, Yu Zhu, Jinqiu Sun, and Yanning Zhang. From distortion manifold to perceptual quality: a data efficient blind image quality assessment approach. *Pattern Recognition*, 133:109047, 2023. 1, 2, 3, 7

[30] 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. 1, 2, 3, 7

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[32] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*, 2023. 2

[33] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 3, 8

[34] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016. 2, 7

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

[36] 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, 7

[37] 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. 3, 6, 7

[38] Zongsheng Yue, Qian Zhao, Jianwen Xie, Lei Zhang, Deyu Meng, and Kwan-Yee K Wong. Blind image super-resolution with elaborate degradation modeling on noise and kernel. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2128–2138, 2022. 3

[39] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017. 3

[40] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 3, 4

[41] 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

[42] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 1, 3, 7

[43] Wenlong Zhang, Guangyuan Shi, Yihao Liu, Chao Dong, and Xiao-Ming Wu. A closer look at blind super-resolution: Degradation models, baselines, and performance upper bounds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 527–536, 2022. 3, 4

[44] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023. 1, 2, 3, 4, 6, 8