Transformer-Based No-Reference Image Quality Assessment via Supervised Contrastive Learning

Jinsong Shi, Pan Gao*, Jie Qin*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
{srache, pan.gao}@nuaa.edu.cn, qinjiebuaa@gmail.com

Abstract

Image Quality Assessment (IQA) has long been a research hotspot in the field of image processing, especially No-Reference Image Quality Assessment (NR-IQA). Due to the powerful feature extraction ability, existing Convolution Neural Network (CNN) and Transformers based NR-IQA methods have achieved considerable progress. However, they still exhibit limited capability when facing unknown authentic distortion datasets. To further improve NR-IQA performance, in this paper, a novel supervised contrastive learning (SCL) and Transformer-based NR-IQA model SaTQA is proposed. We first train a model on a large-scale synthetic dataset by SCL (no image subjective score is required) to extract degradation features of images with various distortion types and levels. To further extract distortion information from images, we propose a backbone network incorporating the Multi-Stream Block (MSB) by combining the CNN inductive bias and Transformer long-term dependence modeling capability. Finally, we propose the Patch Attention Block (PAB) to obtain the final distorted image quality score by fusing the degradation features learned from contrastive learning with the perceptual distortion information extracted by the backbone network. Experimental results on six standard IQA datasets show that SaTQA outperforms the state-of-the-art methods for both synthetic and authentic datasets. Code is available at https://github.com/I2-Multimedia-Lab/SaTQA.

Introduction

Image Quality Assessment (IQA) refers to the quantitative analysis of the content of an image, thus quantifying the degree of visual distortion of a distorted image. The relevant evaluation methods are generally divided into two types: subjective quality assessment (Mantiuk, Tomaszewska, and Mantiuk 2012) and objective quality assessment (Yang et al. 2009). Among them, the objective quality assessment is divided into three categories according to the extent of information the algorithm needs from reference image (Wang and Bovik 2006): full reference quality assessment (FR-IQA) (Wang et al. 2004), reduced-reference quality assessment (RR-IQA) (Rehman and Wang 2012) and no reference quality assessment (NR-IQA) (Mittal, Moorthy, and Bovik
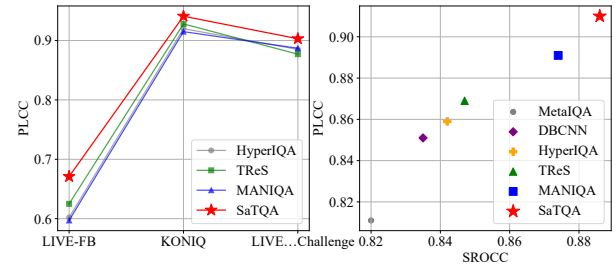

Figure 1: Quantitative comparison of NR-IQA methods. The left figure presents our PLCC results on three authentic datasets compared to the latest NR-IQA method. The right figure presents the average performance of our method compared to other methods on all datasets. Higher coefficient matches perceptual score better.

2012). The FR-IQA and RR-IQA methods calculate the image quality by comparing the difference between the reference image and the distorted image, while the NR-IQA method evaluates the image quality based on the distorted image itself without any reference information, making it the most challenging task and more promising in practical applications.

With recent advance of deep learning techniques in computer vision, more and more data-driven approaches have been developed for NR-IQA (Mittal, Moorthy, and Bovik 2012; Zhang, Zhang, and Bovik 2015; Xue, Zhang, and Mou 2013; Golestaneh, Dadsetan, and Kitani 2022; Yang et al. 2022). However, most existing deep learning based NR-IQA models are trained using the whole training dataset without distinguishing different distortion types and levels, which may render them exhibiting limited capability when facing a large-scale authentic dataset, such as LIVE-FB (Ying et al. 2020). As shown in Fig 1, even the recent most powerful Transformer-based NR-IQA model MANIQA (Yang et al. 2022), the PLCC result on LIVE-FB is down to less than 0.6. This is because, for human eyes, even a same level of distortion but induced by two different types of noises may result in very different visual perception. Thus, without distinguishing different noise types and levels during training, the model may not perform well in the face of a large-scale dataset containing excessive distortions. Some CNN-

---

*Corresponding authors

based methods, such as CONTRIQUE (Madhusudana et al. 2022), attempted to solve this problem by specifying the distortion type and level or using other type of content as the label for distortion classification during training. However, this strategy is not applicable to large-scale authentic dataset, where the distortion type and level are impossible to explicitly specify. To tackle this problem, we propose a supervised contrastive learning based quality assessment method. By conducting self-supervised pre-training on the large scale synthetic dataset KADIS (Lin, Hosu, and Saupe 2020), degradation features corresponding to different types and levels of distorted images are obtained as prior knowledge for final quality score regression. In the architecture design of IQA model, traditional methods usually employ CNN for perceptual distortion feature extraction (Saad, Bovik, and Charrier 2012; Mittal, Moorthy, and Bovik 2012; Zhang, Zhang, and Bovik 2015; Xue, Zhang, and Mou 2013). However, the perception of image quality by humans is usually performed by viewing the entire image, i.e., by integrating the content distortion globally over the whole image. Due to limited receptive field of CNN, the performance of CNN based IQA models may be restricted. On the contrary to CNN, Transformer can integrate global distortion information of images due to its long range dependency modeling ability, which is thus naturally suitable for NR-IQA. Inspired by this, a plethora of Transformer based NR-IQA approaches have been proposed (You and Korhonen 2021; Ke et al. 2021; Golestaneh, Dadsetan, and Kitani 2022; Yang et al. 2022). Nevertheless, pure Transformer has weak ability to extract local details such as image edges and textures, which may degrade the final prediction performance. In addition, most previous NR-IQA methods, such as MANIQA (Yang et al. 2022), used the pretrained Transformer as the backbone that is originally designed for image classification, which may not be well suitable for IQA that is closely related to human eye perception. As can be seen from the Grad-CAM (Selvaraju et al. 2017) in Fig. 2, pure ViT (Dosovitskiy et al. 2020) model focuses more on the semantic region of the image, while missing the distorted areas of the image. To this end, we propose a Multi-Stream Block (MSB), which leverages the CNN and the Transformer features, aiming to combine the strengths of CNN's edge extraction while maintaining the global modeling capability of the Transformer. We further propose the Patch Attention Block (PAB) to fuse the learned perceptual feature with the degradation features from contrastive learning, and the fused features will be used for the prediction of objective image quality scores.

The main contributions of this paper are as follows:

- We use supervised contrastive learning to perform self-supervised pre-training on the large-scale dataset KADIS, and the learned degradation features corresponding to various distortion types and levels of images are used to guide the training of the quality score generation network.

- We propose the MSB module that combines the features of CNN and Transformer, which introduces the inductive bias in CNN while ensuring the long-term dependency

Figure 2: Grad-CAM (Selvaraju et al. 2017) of ViT and SaTQA on distorted image.

dency modeling capability of Transformer for enhancing the ability of the backbone network to extract perceptual distortion information.

- Extensive experiments on six IQA datasets containing synthetic and authentic show that our proposed model outperforms current mainstream NR-IQA methods by a large margin.

Related Work

No-Reference Image Quality Assessment

The previous NR-IQA methods were mainly oriented to quality assessment task for specific distortion types (Li et al. 2013, 2015; Liu, Klomp, and Heynderickx 2009; Wang et al. 2019). These methods use statistical information on images to perform quality assessment of images with known distortion types. However, they are less effective when quality assessment is performed on images with unknown distortion types and therefore restrictive. Current NR-IQA methods focus on general-purpose, and can be further divided into two types according to the feature extraction method: natural scene statistics (NSS)-based metrics (Gao et al. 2013; Mittal, Moorthy, and Bovik 2012; Moorthy and Bovik 2010; Saad, Bovik, and Charrier 2012; Fang et al. 2014) and learning-based metrics (Ye and Doermann 2012; Ye et al. 2012; Zhang et al. 2015, 2018; Lin and Wang 2018; Su et al. 2020; Zhu et al. 2020). The NSS-based methods consider that natural images have regular statistical features (e.g., brightness, gradient, etc.), and images with different distortion types and levels will have different effects on this regularity. Based on this regulation, (Mittal, Moorthy, and Bovik 2012) used the NSS features in the spatial domain to construct the NR-IQA model. (Fang et al. 2014) used the distribution function obtained by fitting the distorted image features to construct the evaluation model based on the moment characteristics of the image and the information entropy. (Moorthy and Bovik 2010) used the discrete wavelet transform (DWT) to extract NSS features, while (Saad, Bovik, and Charrier 2012) used the discrete cosine transform (DCT) to evaluate the image quality. However, the NR-IQA methods based on NSS require manual extraction of local or global features of the image, and the feature representation extracted by this hand-crafted approach is often not effective. Although these NR-IQA methods show some performance gain on some synthetic distorted datasets, the results are not as good as on authentic distorted datasets. With the success of deep
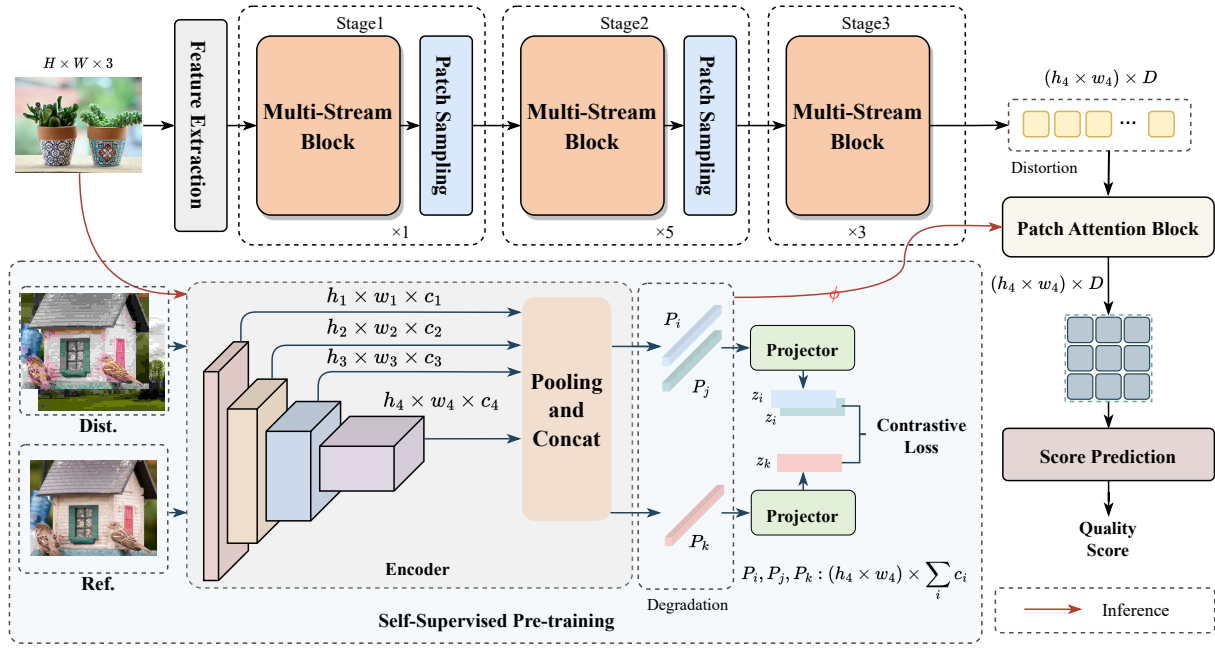
Figure 3: The overall framework of our approach for no-reference image quality assessment. P_i, P_j, and P_k refer to the degradation features, z_i, z_j, and z_k correspond to the vectors after projected. ϕ denotes stopping gradient propagation.

learning on various vision tasks, CNN-based NR-IQA methods have also achieved some success on distorted datasets. DBCNN (Zhang et al. 2018) uses a CNN structure based on bilinear pooling for NR-IQA model building and achieves good performance gain on synthetic datasets. Hallucinated-IQA (Lin and Wang 2018) proposes an NR-IQA model based on generative adversarial networks, which obtains the final quality score by pairing the generated hallucinated reference image with the distorted image. HyperIQA (Su et al. 2020) predicts image quality by fusing distortion information from images of different scales. MetaIQA (Zhu et al. 2020) trains the model to acquire prior knowledge of distorted images via meta-learning and solves the generalization performance problem of NR-IQA task to some extent.

Transformer for IQA

In the past few years, CNNs had become the popular backbone of computer vision tasks and achieved success in the field of IQA. Unfortunately, CNNs are highly flawed in capturing non-local information and have a strong localization bias. In addition, the shared convolutional kernel weights and translational invariance properties of CNNs also lead to the inability of CNNs to handle complex feature combinations. Inspired by the use of Transformer for long-range dependency modeling in the NLP domain (Vaswani et al. 2017), ViT (Dosovitskiy et al. 2020), a Transformer model for computer vision tasks, has achieved initial success. In the field of IQA, IQT (You and Korhonen 2021) applies a hybrid architecture to extract feature information of reference and distorted images by CNN and use it as input to Transformer. MUSIQ (Ke et al. 2021) proposes a multi-scale image quality Transformer, which solves the problem

of different image resolutions by encoding three scales of an image. TReS (Golestaneh, Dadsetan, and Kitani 2022) proposes to use relative ranking and self-consistency loss to train the model and thus enhance the evaluation performance. MANIQA (Yang et al. 2022) proposes a multi-dimension attention mechanism for multi-dimensional inter-action on channel and spatial domains, and achieves promising evaluation performance on synthetic datasets. However, all these Transformer based models are trained without distinguishing the latent degradation from other ones, which may not generalize well to a large-scale authentic dataset.

Contrastive Learning

Unlike supervised learning, self-supervised learning aims to acquire features using unlabeled data. Current contrastive learning methods mainly focus on instance recognition tasks, i.e., the image itself and its corresponding enhanced samples are considered as the same class (Dosovitskiy et al. 2014; Chen et al. 2020; He et al. 2020). In addition, some self-supervised methods learn data features by constructing auxiliary tasks, which are often large and do not require manual annotation. Usually, these auxiliary tasks include rotation prediction (Gidaris, Singh, and Komodakis 2018), image restoration (Pathak et al. 2016), etc. (Liu, Van De Weijer, and Bagdanov 2019) used image ranking as an auxiliary task and the proposed NR-IQA model achieved good performance on synthetic datasets. CONTRIQUE (Madhusudana et al. 2022) used distortion type and level as an auxiliary task by comparing user-generated content (UGC) with the images in synthetic datasets, finally, the learned features were used for image quality prediction. In this paper, we proposed a supervised contrastive learning (SCL) approach inspired

by this auxiliary task construction method. But, taking one step further, we remove the use of UGC images and use the reference images from the synthetic dataset alone for comparasion.

Proposed Method

In this section, we detail our proposed model, which is an NR-IQA method based on supervised contrastive learning and Transformer named **SaTQA**, and Fig. 3 shows the architecture of our proposed model.

Overall Pipeline

Given a distorted image $I \in \mathbb{R}^{H \times W \times 3}$, where H and W denote height and width, respectively, and the goal of NR-IQA is to predict its objective quality score. First, we use ViT (Dosovitskiy et al. 2020) to extract the original features of the image, where each layer of extracted features defines $F_i \in \mathbb{R}^{(h \times w) \times C_i}$, where $i \in \{1, 2, ..., 12\}$. In this paper, we extract four of these layers and concatenate them together by channel dimension, denoted as $\widehat{F} \in \mathbb{R}^{\sum_i (h \times w) \times C_i}$, where $i \in \{1, 2, 5, 6\}$, C_i denotes the dimension of the ViT layer, h and w denotes the height and width of the corresponding output, respectively. Then we downscale $\widehat{F}$ along channel dimension to get $\widehat{F}' \in \mathbb{R}^{h \times w \times D}$. Next we use a three-stage MSB module for image distortion extraction, where the first two stages are down-sampled after MSB processing. The distortion information is extracted by the backbone network and the output is $R \in \mathbb{R}^{(h_4 \times w_4) \times D}$, where D denotes the channel dimension of the features, h_4 and w_4 denotes the height and width of the last stage output. At the same time, the distorted image is passed through ResNet-50 (He et al. 2016) trained with SCL and the output is P, $P \in \mathbb{R}^{(h_4 \times w_4) \times \sum_i c_i}$. We make P and R have the same shape, and thus $D = \sum_i c_i$. Next, the PAB module fuses the features of P and R to obtain the feature S for the image quality, and the final S is regressed to the objective quality score corresponding to the image.

Supervised Contrastive Learning

Self-supervised learning improves model performance by augmenting the model's ability to extract data features through auxiliary tasks. In the NR-IQA domain, we need the model to distinguish the distortion types and levels of different images, and the target of this auxiliary task is thus similar to the classification of image. However, different from traditional supervised classification, we train the model using SCL on a large-scale synthetic dataset to obtain a more robust and generalized degradation features than classification, so that the degradation from the same distortion type (level) are pulled closer together than degradation from different distortion type (level). We then generalize the pre-trained model to other dataset with latent degradation, e.g., authentic distortion dataset. As shown in Fig. 4, at the beginning, the distribution of various distorted images on the feature space is random and unordered. Our goal is that images with the same distortion and level can be clustered together after using supervised contrastive learning.
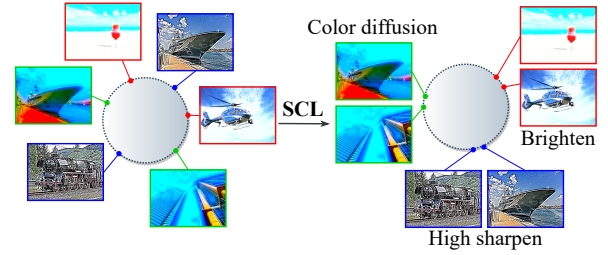

Figure 4: Left: Distribution of images with the same level of different distortion types in the feature space. Right: Distribution of images after training with SCL, where the images with color diffusion, high sharpen, brighten are clustered together, respectively.

In our proposed supervised contrastive learning for distorted image clustering, firstly, I_u^v is defined to denote the distorted image, $u \in \{1, ..., U\}$ to denote the different distortion types corresponding to I, and $v \in \{1, ..., V\}$ to denote the distortion level under each u, where U and V denote the total number of distortion types and levels. Thus, together with the class of the reference image, there are a total of $U \times V + 1$ classes. That is, this auxiliary task is transformed into distinguish different types and levels of distorted images problem. Inspired by (Chen et al. 2020), we use encoder network $f(.)$ together with projector $g(.)$ structure to learn the degradation features of the images. In this paper, the encoder uses ResNet-50, and the projector uses a multilayer perceptron (MLP). For a given image x, $r = f(x)$, $z = g(r)(z \in \mathbb{R}^D)$, where z denotes the D-dimensional feature of the output, r denotes the feature representation of the encoder output. As with (Chen et al. 2020), the encoder output is L_2 normalized before being passed to the projector. The contrast loss function uses a supervised normalized temperature-scaled cross entropy (NT-Xent) (Khosla et al. 2020), which can be formulated as:

$$\mathcal{L}_{fea} = \frac{1}{|P(i)|} \sum_{j \in P(i)} - \log \frac{\exp(\phi(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{I}_{k \neq i} \exp(\phi(z_i, z_k)/\tau)} \tag{1}$$

where N denotes the number of all distortions and references images in a mini-batch, $\mathbb{I}$ denotes the indicator function, τ denotes the temperature parameter, $P(i)$ denotes the set belonging to the same class i, and $|P(i)|$ is its cardinality. $\phi(m, n)$ denotes $\phi(m, n) = m^T n / \|m\|_2 \|n\|_2$.

Multi-Stream Block

As shown in Fig. 5 (a), we propose a Multi-Stream Block (MSB) that combines CNN to extract low-level edge features and Transformer to extract high-level content features of images. First, we reshape the feature $\widehat{F}'$ after ViT extraction into feature map $\widetilde{F} \in \mathbb{R}^{D \times h \times w}$; then we divide it into three parts, $\chi_1 \in \mathbb{R}^{D_1 \times h \times w}$, $\chi_2 \in \mathbb{R}^{D_2 \times h \times w}$, and $\chi_3 \in \mathbb{R}^{D_3 \times h \times w}$, along the channel dimension, where χ_1 and χ_2 are feed to CNN and χ_3 is feed to Transformer. χ_1 will go through deformable convolution (DeformConv) and linear layers to get χ_1'. χ_2 will first go through depthwise
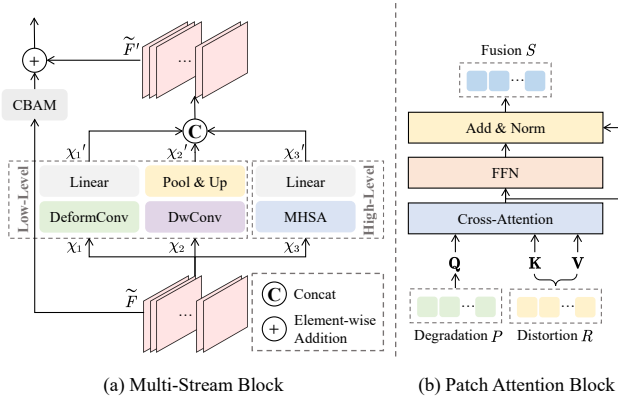
(a) Multi-Stream Block   (b) Patch Attention Block

Figure 5: Overview of MSB and PAB.

Datasets	Dist. Images	Dist. types	Dataset types
LIVE	799	5	Synthetic
CSIQ	866	6	Synthetic
TID2013	3000	24	Synthetic
KADID-10K	10125	25	Synthetic
KADIS	700k	25	Synthetic
LIVEC	1162	-	Authentic
KONIQ	10073	-	Authentic
LIVE-FB	39810	-	Authentic

Table 1: Summary of IQA datasets.

convolution (DwConv) and max-pooling layers, and then up-sample to get χ_2'. χ_3 will go through Multi-Head Self-Attention (MHSA) and linear layers to get χ_3'. These can be formulated as:

$$\chi_1' = \text{FC}(\text{Deform}(\text{Conv}(\chi_1))) \tag{2}$$

$$\chi_2' = \text{UpSample}(\text{MaxPool}(\text{DwConv}(\chi_2))) \tag{3}$$

$$\chi_3' = \text{FC}(\text{MHSA}(\chi_3)) \tag{4}$$

Finally, χ_1', χ_2' and χ_3' are concatenated together and reshaped, which is denoted as $\widetilde{F}' \in \mathbb{R}^{D \times h \times w}$. In addition, we connect $\widetilde{F}$ to $\widetilde{F}'$ with residual learning through the CBAM (Woo et al. 2018) (*i.e.*, channel attention and spatial attention) module. In this paper, the percentage of χ_1, χ_2 and χ_3 in the channel dimension will change with different stages. At the beginning, all three share the same percentage in the channel, and then χ_3 will gradually increase its percentage in the total channels. Note that, in CNN branches, deformable convolution can augment the spatial sampling locations, and depthwise convolution can reduce the number of computations and supplement the channel attention that is not modeled in Transformer branch.

Patch Attention Block

After self-supervised pre-training on the large-scale IQA dataset KADIS, the degradation P can be used to distinguish the distortion level and type of the image to be evaluated. As shown in Fig. 5 (b), We map P to Query ($\mathbf{Q}$) and R to Key ($\mathbf{K}$) and Value ($\mathbf{V}$), and then calculate them by Eq. (5) to obtain the final image quality features S. With this kind of computation, the model can achieve good performance in the evaluation and converge faster even on small-scale IQA datasets due to P-assisted training.

$$S = \text{Softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^{\text{T}}}{\sqrt{\text{D}}}\right) \times \mathbf{V} \tag{5}$$

Quality Prediction

For score prediction, we first reshape the tokens from patch attention block into feature maps, and then the quality score

was regressed with the quality vector obtained via global average pooling (GAP). In each batch of images, we train our quality prediction model by minimizing the regression loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \|p_i - y_i\|_1 \tag{6}$$

where N denotes the batch size, p_i denotes the image quality score predicted by the model for the ith image, and y_i denotes the corresponding objective quality score.

Experiments

Datasets

For self-supervised pre-training, we use the KADIS (Lin, Hosu, and Saupe 2020) dataset. The KADIS dataset contains 140,000 original reference images, 700,000 distorted images, and has no subjective quality score. It has 25 distortion types with five distortion levels for each distortion type. In addition, we evaluated the performance of our proposed model on six standard synthetic and authentic IQA datasets. The synthetic datasets include CSIQ (Larson and Chandler 2010), TID2013 (Ponomarenko et al. 2015), and KADID-10K (Lin, Hosu, and Saupe 2019), and the authentic datasets include LIVE Challenge (Ghadiyaram and Bovik 2015), KONIQ (Hosu et al. 2020), and LIVE-FB (Ying et al. 2020). Tab. 1 shows a summary of the datasets used in the experiments. The commonly used observer ratings for the images are expressed by Mean opinion score (MOS) and Differential Mean opinion score (DMOS), where larger MOS values indicate better image quality and larger DMOS values indicate poorer image quality.

Evaluation Metric

We use Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) to measure the performance of the NR-IQA method. The values of SROCC and PLCC are in the range [-1, 1], and the higher absolute values indicate higher correlation and vice versa.

Implementation Details

In this paper, all of our experiments are performed using PyTorch on a single NVIDIA GeForce RTX3090. During image preprocessing, we cropped 8 random 224×224 patches from each image and randomly flipped the cropped patches

	CSIQ		TID2013		KADID-10K		LIVE Challenge		KONIQ		LIVE-FB	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
DIIVINE	0.776	0.804	0.567	0.643	0.435	0.413	0.591	0.588	0.558	0.546	0.187	0.092
BRISQUE	0.748	0.812	0.571	0.626	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303
ILNIQE	0.865	0.822	0.648	0.521	0.558	0.534	0.508	0.508	0.537	0.523	0.332	0.294
BIECON	0.823	0.815	0.762	0.717	0.648	0.623	0.613	0.613	0.654	0.651	0.428	0.407
MEON	0.864	0.852	0.824	0.808	0.691	0.604	0.710	0.697	0.628	0.611	0.394	0.365
WaDIQaM	0.844	0.852	0.855	0.835	0.752	0.739	0.671	0.682	0.807	0.804	0.467	0.455
DBCNN	0.959	0.946	0.865	0.816	0.856	0.851	0.869	0.869	0.884	0.875	0.551	0.545
TIQA	0.838	0.825	0.858	0.846	0.855	0.850	0.861	0.845	0.903	0.892	0.581	0.541
MetaIQA	0.908	0.899	0.868	0.856	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0.540
P2P-BM	0.902	0.899	0.856	0.862	0.849	0.840	0.842	0.844	0.885	0.872	0.598	0.526
HyperIQA	0.942	0.923	0.858	0.840	0.842	0.844	0.885	<u>0.872</u>	0.920	0.907	0.602	0.544
TReS	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	<u>0.928</u>	<u>0.915</u>	<u>0.625</u>	<u>0.554</u>
MANIQA	<u>0.968</u>	<u>0.961</u>	<u>0.943</u>	<u>0.937</u>	<u>0.946</u>	<u>0.944</u>	<u>0.887</u>	0.871	0.915	0.880	0.597	0.543
SaTQA (Ours)	**0.972**	**0.965**	**0.948**	**0.938**	**0.949**	**0.946**	**0.903**	**0.877**	**0.941**	**0.930**	**0.676**	**0.582**

Table 2: Comparison of SaTQA *v.s.* state-of-the-art NR-IQA algorithms on synthetically and authentically distorted datasets. The best performances are in bold, and the second best are underlined. Some are borrowed from (Golestaneh, Dadsetan, and Kitani 2022).

Dataset	LIVE					CSIQ					
Type	JP2K	JPEG	WN	GB	FF	WN	JPEG	JP2K	FN	GB	CC
BRISQUE	0.929	0.965	0.982	0.964	0.828	0.723	0.806	0.840	0.378	0.820	0.804
ILNIQE	0.894	0.941	0.981	0.915	0.833	0.850	0.899	0.906	0.874	0.858	0.501
HOSA	0.935	0.954	0.975	0.954	0.954	0.604	0.733	0.818	0.500	0.841	0.716
BIECON	0.952	**0.974**	0.980	0.956	0.923	0.902	0.942	0.954	0.884	0.946	0.523
WaDIQaM	0.942	0.953	0.982	0.938	0.923	0.974	0.853	0.947	0.882	0.976	0.923
PQR	**0.953**	0.965	0.981	0.944	0.921	0.915	0.934	0.955	0.926	0.921	0.837
HyperIQA	0.949	0.961	0.982	0.926	0.934	0.927	0.934	0.960	0.931	0.915	0.874
MANIQA	0.870	0.895	0.984	0.959	0.896	0.966	0.971	0.973	0.977	0.956	0.946
Ours	0.947	0.965	**0.989**	**0.988**	**0.955**	**0.985**	**0.984**	**0.991**	**0.986**	**0.980**	**0.970**

Table 3: SROCC comparisons of individual distortion types on the LIVE and CSIQ datasets.

horizontally. The trained patches inherit the quality score of the original images. The Encoder used for contrastive learning is a modified ResNet-50 (He et al. 2016), and the feature extraction part of the backbone network uses ViT-B/8 (Dosovitskiy et al. 2020), where the patch size is 8. SaTQA contains three stages, the first two stages consist of Multi-Steam Block (MSB) and Patch Sampling, and the last stage contains only MSB, where D is set to 768 and h_4 and w_4 are 7. In the three stages, the number of channels per branch within MSB D_1, D_2 and D_3 are [256, 256, 256], [192,192, 384], and [48,48,672]. The same training strategy as used in the existing NR-IQA algorithm. For the synthetic dataset, we divide the dataset based on the reference image. In the training process, we use AdamW optimizer with learning rate of 2×10^{-5}, weight decay of 1×10^{-2}, and learning strategy of cosine annealing, where T_{max} is set to 50 and eta_{min} is 0. Experiments are trained for 150 epochs. The loss function used is L1Loss, and the batch size is 4. For testing, we generate the final objective quality score by averaging the 8 cropped patch scores of the distorted images. We report the average of the results by running the experiment 10 times with different seeds.

Performance Comparison

Individual dataset evaluations. As shown in Tab. 2, the SROCC and PLCC of SaTQA on all datasets are outperform of the state-of-the-art methods (Saad, Bovik, and Charrier 2012; Mittal, Moorthy, and Bovik 2012; Zhang, Zhang, and Bovik 2015; Kim and Lee 2016; Ma et al. 2017; Bosse et al. 2017; Zhang et al. 2018; You and Korhonen 2021; Zhu et al. 2020; Ying et al. 2020; Su et al. 2020; Golestaneh, Dadsetan, and Kitani 2022; Yang et al. 2022; Xu et al. 2016; Zeng, Zhang, and Bovik 2017). As the current correlation coefficients of the NR-IQA method on most syhthetic datasets already converge to 1 (especially the performance on CSIQ approaches saturation), and it is thus extremely difficult to further make significant improvement on those datasets. Nevertheless, the performance of SaTQA on these synthetic datasets is still improved, which demonstrates the effectiveness of our proposed model. Moreover, SaTQA performs well on the authentic dataset, not only on the small-scale dataset LIVE Challenge, but also on the current large-scale LIVE-FB dataset.

Individual distortion evaluations. Since there are various types of distortion in images, especially in the authentic dataset, an image may contain multiple types of distortion

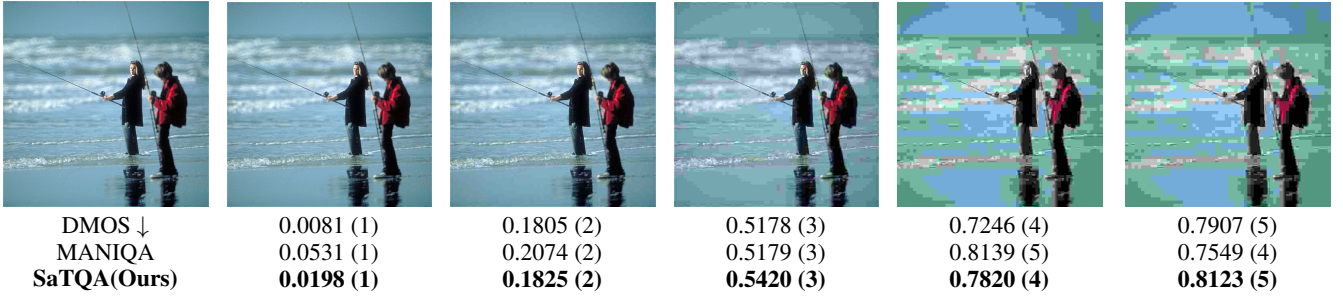|					
DMOS ↓	0.0081 (1)	0.1805 (2)	0.5178 (3)	0.7246 (4)	0.7907 (5)
MANIQA	0.0531 (1)	0.2074 (2)	0.5179 (3)	0.8139 (5)	0.7549 (4)
SaTQA(Ours)	**0.0198 (1)**	**0.1825 (2)**	**0.5420 (3)**	**0.7820 (4)**	**0.8123 (5)**

Figure 6: Example images of JPEG distortion types from the CSIQ test dataset. DMOS denotes the human rating, and smaller DMOS values indicate better image quality. The number in the parenthesis denotes the rank among considered distorted images.

Trained on	KonIQ	LIVEC	LIVE	
Test on	LIVEC	KONIQ	CSIQ	TID2013
WaDIQaM	0.682	0.711	0.704	0.462
P2P-BM	0.770	0.740	0.712	0.488
HyperIQA	0.785	0.772	0.744	0.551
TReS	0.786	0.733	0.761	0.562
MANIQA	0.773	0.759	0.813	0.581
Proposed	**0.791**	**0.788**	**0.831**	**0.617**

Table 4: SROCC evaluations on cross datasets, where bold entries indicate the best performers.

(e.g., blur, noise, etc.). To validate the general performance of SaTQA on the distortion types, we conducted experiments on synthetic datasets LIVE and CSIQ.As can be seen from Tab. 3, our model has a strong generalization performance on CSIQ and LIVE dataset.

Cross-dataset evaluations. To verify the generalization performance of the model on different datasets, we conducted cross-dataset experiments. In this validation, the training process is performed on a specific dataset and the test is executed on another different dataset without any parameter tuning during the test. As shown in Tab. 4, our model has strong generalization ability.

Ablation Study

In Tab. 5, we provide ablation studies to validate the effect of each component of our proposed model. We use the ViT pre-trained on ImageNet as our baseline (#1), with the patch size of 8, and the original CLS token for quality regression.

Supervised Contrastive Learning. To efficiently utilize a large amount of unlabeled data and enhance the prediction performance of the model on the authentic IQA dataset, we propose the SCL module. The results in #2 indicate that the use of SCL can significantly improve the performance of the model on the LIVE Challenge (+0.044 PLCC, +0.037 SROCC). For CSIQ dataset (+0.041 PLCC, +0.04 SROCC).

Multi-Stream Block. To leverage the feature extraction capabilities of CNN and Transformer for IQA tasks, we propose the MSB. In #3, the model has 0.01 improvement in SROCC and PLCC on both datasets.

Patch Attention Block. PAB is propesed to utilize the

SCL	MSB	PAB	LIVEC		CSIQ	
			PLCC	SROCC	PLCC	SROCC
			0.827	0.803	0.922	0.916
✓			0.871	0.840	0.963	0.956
✓	✓		0.880	0.851	0.970	0.963
✓	✓	✓	**0.903**	**0.877**	**0.972**	**0.965**

Table 5: Ablation study on modules. #1 represents the baseline performance without using any proposed components.

degradation learned by SCL as a query to compute cross-attention with the perceptual distortion extracted by MSB. From #4, by adding PAB, the model has a 0.02 improvement in SROCC and PLCC on the LIVE Challenge dataset and a 0.002 increase on the CSIQ dataset.

Visual Result Analysis

We provide the DMOS prediction results of MANIQA and SaTQA for JPEG distortion images on CSIQ test set as an example. As shown in Fig. 6, it can be seen that SaTQA performs better than MANIQA both in image distortion level ranking and in overall DMOS value prediction, where MANIQA produces errors in the prediction of the fourth and fifth distortion levels.

Conclusion

In this paper, we propose SaTQA, a specialized model for NR-IQA by combining supervised contrastive learning and Transformer. Initially, SCL is utilized to address incapability of existing models when facing authentic datasets for quality prediction. To enhance the feature extraction capability for distorted images and overcome the limitations of global modeling in CNN, we introduce the MSB module. Additionally, we design the PAB module to effectively fuse image degradation features and distortion information. Experimental results on six standard IQA datasets demonstrate the remarkable performance achieved by our model.

Acknowledgements

References

Bosse, S.; Maniry, D.; Müller, K.-R.; Wiegand, T.; and Samek, W. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1): 206–219.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27.

Fang, Y.; Ma, K.; Wang, Z.; Lin, W.; Fang, Z.; and Zhai, G. 2014. No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters*, 22(7): 838–842.

Gao, X.; Gao, F.; Tao, D.; and Li, X. 2013. Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning. *IEEE Transactions on neural networks and learning systems*, 24(12).

Ghadiyaram, D.; and Bovik, A. C. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1): 372–387.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. 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*, 1220–1230.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Kim, J.; and Lee, S. 2016. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1): 206–220.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1): 011006.

Li, L.; Lin, W.; Wang, X.; Yang, G.; Bahrami, K.; and Kot, A. C. 2015. No-reference image blur assessment based on discrete orthogonal moments. *IEEE transactions on cybernetics*, 46(1): 39–50.

Li, L.; Zhu, H.; Yang, G.; and Qian, J. 2013. Referenceless measure of blocking artifacts by Tchebichef kernel analysis. *IEEE Signal Processing Letters*, 21(1): 122–125.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. IEEE.

Lin, H.; Hosu, V.; and Saupe, D. 2020. Weak supervision for deep IQA feature learning. *arXiv preprint arXiv:2001.08113*.

Lin, K.-Y.; and Wang, G. 2018. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 732–741.

Liu, H.; Klomp, N.; and Heynderickx, I. 2009. A no-reference metric for perceived ringing artifacts in images. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(4): 529–539.

Liu, X.; Van De Weijer, J.; and Bagdanov, A. D. 2019. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1862–1878.

Ma, K.; Liu, W.; Zhang, K.; Duanmu, Z.; Wang, Z.; and Zuo, W. 2017. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3): 1202–1213.

Madhusudana, P. C.; Birkbeck, N.; Wang, Y.; Adsumilli, B.; and Bovik, A. C. 2022. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31: 4149–4161.

Mantiuk, R. K.; Tomaszewska, A.; and Mantiuk, R. 2012. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, 2478–2491. Wiley Online Library.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Moorthy, A. K.; and Bovik, A. C. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters*, 17(5): 513–516.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30: 57–77.

Rehman, A.; and Wang, Z. 2012. Reduced-reference image quality assessment by structural similarity estimation. *IEEE transactions on image processing*, 21(8): 3378–3389.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing*, 21(8): 3339–3352.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. 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*, 3667–3676.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, G.; Wang, Z.; Gu, K.; Li, L.; Xia, Z.; and Wu, L. 2019. Blind quality metric of DIBR-synthesized images in the discrete wavelet transform domain. *IEEE Transactions on Image Processing*, 29: 1802–1814.

Wang, Z.; and Bovik, A. C. 2006. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1): 1–156.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xu, J.; Ye, P.; Li, Q.; Du, H.; Liu, Y.; and Doermann, D. 2016. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9): 4444–4457.

Xue, W.; Zhang, L.; and Mou, X. 2013. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 995–1002.

Yang, J.; Hou, C.; Zhou, Y.; Zhang, Z.; and Guo, J. 2009. Objective quality assessment method of stereo images. In *2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, 1–4. IEEE.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.

Ye, P.; and Doermann, D. 2012. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7): 3129–3138.

Ye, P.; Kumar, J.; Kang, L.; and Doermann, D. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, 1098–1105. IEEE.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. 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*, 3575–3585.

You, J.; and Korhonen, J. 2021. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1389–1393. IEEE.

Zeng, H.; Zhang, L.; and Bovik, A. C. 2017. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8): 2579–2591.

Zhang, P.; Zhou, W.; Wu, L.; and Li, H. 2015. SOM: Semantic obviousness metric for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2394–2402.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2018. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.

Zhu, H.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2020. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14143–14152.