Kernel-aware Dual-level Statistical Distance for Image Quality Assessment with Relaxed Reference

Xinpeng Li, Ting Jiang, Qingbo Wu, *Member, IEEE,* Hongliang Li, *Senior Member, IEEE,* Bing Zeng, *Fellow, IEEE,* Shuaicheng Liu, *Member, IEEE*

***Abstract*—This paper presents an unified image quality assessment method with relaxed reference including the pixel-wise aligned, non-aligned, or completely irrelevant pristine images. To achieve this appealing flexibility, we propose a Kernel-aware Dual-level Statistical Distance (KDSD). Firstly, a parallel Fourier and Vanilla convolution module are applied to both the distorted image and its reference to extract the complementary dual-level features, which capture the global and local information from the frequency and spatial domains respectively. Then, to avoid the potential position misalignment between the distorted image and its reference, we measure their feature difference with the maximum mean discrepancy (MMD) based statistical distance, which is applied to two distributions without considering the spatial position of each element. Benefits from the kernel embedding, the MMD could represent the high-dimensional features' statistical distance via regular Hilbert space operations without complex density estimation. Finally, the MMDs calculated from the global- and local-level features are fed to the regression head to obtain the predicted subjective score. Extensive experiments show that the proposed method outperforms the state-of-the-arts by a relatively large margin with the pixel-wise aligned, non-aligned or completely irrelevant reference. Remarkably, our non-aligned method performs even better than many other aligned methods.**

***Index Terms*—Image quality assessment, maximum mean discrepancy, Fourier convolution.**

I. Introduction

Image Quality Assessment (IQA) [1]–[6] is a task that evaluates the image quality based on the human preferences. A good image quality assessment can benefit various image-related tasks, such as super-resolution [7], image denoise [8], and HDR [9].

IQA tasks can be broadly divided into reference and no-reference approaches, where the former one requires a reference image during the evaluation while the latter one has no reference image. Studies have shown that humans are good at comparing two images rather than making judgments by looking at only a single image [10]. Therefore, reference approaches are more accurate than no-references ones. However, in reality, it is difficult, if not impossible, to obtain a reference image whose content is exactly the same as that of the distorted image. Therefore, non-aligned reference IQA (NAR-IQA) approaches have been proposed [2], [11], allowing the relaxed references where reference and distorted images can have different contents.

Manuscript submitted Oct. 24, 2022.

Authors are with School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China, 611731.Corresponding Author: Shuaicheng Liu (liushuaicheng@uestc.edu.cn)

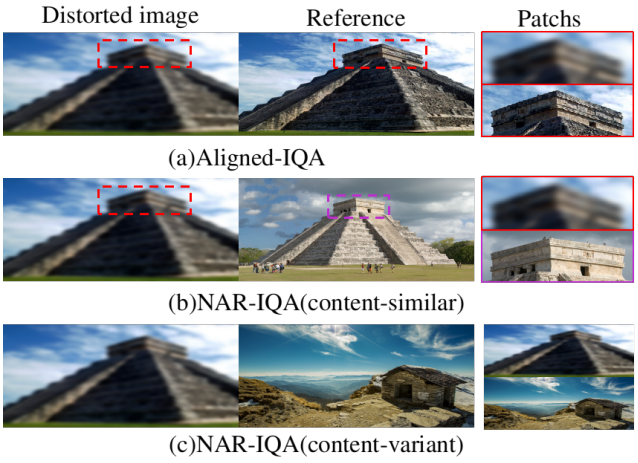

Fig. 1. (a) Aligned-IQA: distorted and reference images are aligned pixel-wisely. (b) NAR-IQA (content-similar): reference and distorted images are not aligned, but similar contents exist. (c) NAR-IQA (content-variant): the contents in two images are completely different.

As shown in Fig. 1, NAR-IQA can be further divided into NAR-IQA (content-similar) and NAR-IQA (content-variant) based on the similarity of the image contents, which we refer to as NAR-IQA-S and NAR-IQA-V, respectively. Specifically, NAR-IQA-S needs a reference image with similar content to the distorted image. Such reference images can be easily obtained if the contents of distorted images are common or famous places, as shown in Fig. 1 (b). However, if no similar contents can be acquired, we can resort to the NAR-IQA-V, where a randomly chosen reference image is adopted, as shown in Fig. 1 (c). Notably, for NAR-IQA-V, difference cannot be directly measured between reference and distorted images due to their variant contents. The reference image can be used to learn what an undistorted image looks like as a reference for learning the degree of distortion of the distorted image.

The reference-based IQA methods consist of three steps: First, extracting features of distorted and reference images. Second, measuring the distance between the extracted features. Third, predicting the score based on the results obtained in the second step. All the steps are important with more credits given to the first two.

With respect to the feature extraction, deep IQA methods often extract features with convolutional neural networks [2], [12]. However, classical convolutional neural networks can only provide local receptive fields with local features, which is not friendly when the distortion area is relatively large.

The situation becomes even worse, when the reference and distorted images have different contents. For example, In Fig. 1 (a), the local feature from small receptive field may be enough when reference and distorted images are aligned pixel-wisely. However, when content varies, local features are insufficient.

Therefore, for non-aligned reference approaches, larger receptive fields are preferred. Recently, transformer-based method [13] have been proposed and show that large receptive field is beneficial to the IQA task. However, none of the existing non-aligned reference IQA methods consider this problem, and they all use traditional CNN for feature extractions. Instead of introducing transformer or other alternatives for global features, in this work, we adopt Fourier convolutions [14]. We proposes a pipeline that uses classical convolution and Fourier convolution for the dual-level feature extraction at multiple scales simultaneously. Fourier convolution is an application of Fourier transform in neural networks [15]. According to the spectral convolution theorem, change a value in frequency domain would have a global impact on the original signal. We can obtain the global receptive field through the Fourier convolution. In this work, we show that it is a good choice for non-aligned reference IQA. In addition, Fourier convolution can extract features of different frequencies, which is beneficial for the assessment of distortions such as blur and noise.

Having the features been extracted, the next step is to calculate their distance. If reference and distorted images are aligned, this step is straightforward. Most of the previous methods choose to calculate the feature distance rudely, such as absolute differences [16], [17]. However, similar methods are still used in non-aligned approaches [2], [11], which is unreasonable. Because, the gap between the distorted and reference images is not only the degree of distortion, but also the varying image contents. In this work, we measure their statistical distance instead of the position-related difference.

Recently, [18] utilizes the local sliced Wasserstein distance to measure the statistical distance between distorted and reference images under the situation of full aligned IQA, showing the superiority of statistical distance over non-statistical ones. However, this method is not applicable to NAR-IQA. Because Wasserstein distance is used to measure the minimum cost of converting one distribution to another. However, in NAR-IQA this transformation is absent. Various methods can be applied to count the statistical distances between two distributions, such as entropy, mutual information, and Kullback–Leibler divergence. However, these methods either require density estimation or sophisticated space-partitioning/bias-correction strategies, which is inapplicable to high-dimensional data. In this work, we adopt maximum mean discrepancy (MMD) and show that it is a more reasonable choice. The MMD is the moments supremum of arbitrary order of two distributions. MMD can be solved by the kernel embedding, which can avoid density estimation and parametric assumptions. Furthermore, the kernel-aware method has good generalization and property of finite sample convergence. By combining MMD with our feature extractor, we propose a Kernel-aware Dual-level Statistical Distance (KDSD) which can measure the statistical distance between dual-level features in a infinite-dimensional reproducing kernel Hilbert space.

We conduct extensive experiments on 4 representative datasets, LIVE [19], CSIQ [20], TID2013 [21] and KADID-10k [22]. Our method achieves state-of-the-art performance in both aligned and non-aligned IQA tasks. In sum, main contributions of this work includes:

- We propose a new feature extraction method with classical and Fourier convolutions which can combines the local and global information effectively.
- We introduce MMD into the IQA task and show its effectiveness for the evaluation of statistical distance between distorted and reference images.
- Extensive experiments on 4 benchmarks show that our method outperforms state-of-the-art methods. In particular, our method using non-aligned references performs even better than other methods using aligned references.

II. Related Work

A. Traditional Image Quality Assessment

Image quality assessment aims to evaluate image quality in accordance with human preferences. IQA methods can be broadly divided into reference and no-reference approaches. For reference-based methods, they can be further divided into aligned and non-aligned categories.

Full-Reference IQA (FR-IQA) belongs to the aligned category, which often evaluates the quality in terms of PSNR and SSIM [23]. However, these objective metrics may not fully reflect the human options. As a result, FR-IQA methods try to involve more subjective ingredients. Traditional FR-IQA methods evaluate the similarity between the distorted and reference images by calculating the distance between aligned pixels, or hand-crafted features extracted form them [24]. Deep methods use convolutional neural networks to extract image features. For example, [18] proposed a semi-supervised approach to make predictions with the help of unlabled data. [13] introduced the transformer to enhance the capability of feature extraction. All these methods, however, require strict alignment between distorted and reference images, which is not applicable in many scenarios where reference labels are unavailable.

NR-IQA has wider application prospects because it does not require any other assistances at all. It is the first time that neural networks are used in the NR-IQA task in [25]. Since then, many NN-based methods have been proposed. [26] did not directly regress the MOS, but first regress the quality map obtained by the existing FR-IQA methods to reduce the difficulty of the regression. [27], [28] introduced prior knowledge from other tasks to the NR-IQA task. A network pre-trained on ImageNet for classification tasks was used in [27]. [28] uses mate learning to make the pre-trained network better cope with different types of distortion.

B. Non-aligned Reference Image Quality Assessment

Non-Aligned Reference IQA (NAR-IQA) just require relaxed references which could not be strictly aligned with distorted images. The content of reference images can be either similar or even completely different from distorted images. On

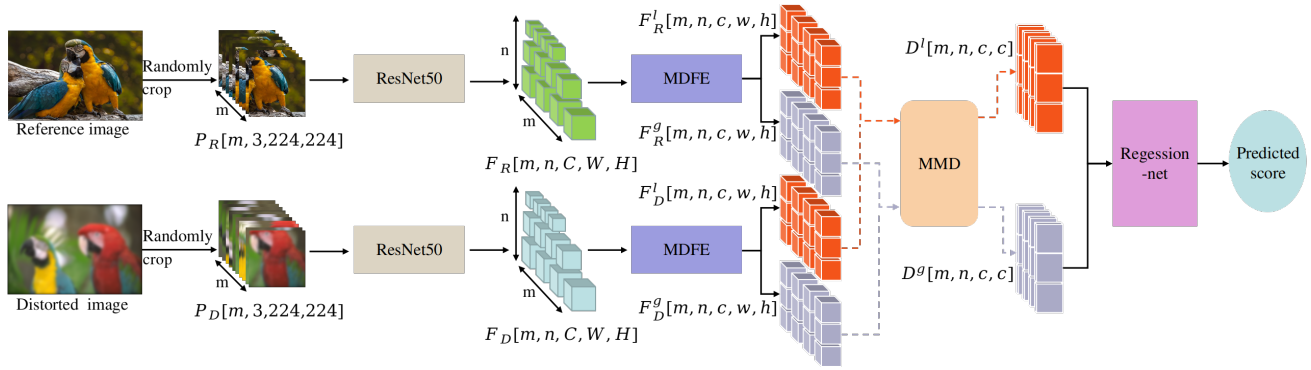

Fig. 2. The pipeline of our method. First we extract the multi-scale features of distorted and reference images separately through a pre-trained ResNet50. Further, we use a multi-scale dual-level feature extractor (MDFE) to perform feature extraction in the local and global levels. Then we calculate the distance matrices of the extracted features through MMD. Finaly predicted score produced by dealing these distance matrices through a regression network. The two MDFEs share weights.

one hand, a reference image is available to support the quality assessment, on the other hand, the strict alignment requirement is relaxed to facilitate a wide application range. [11] first proposed the concept of NAR-IQA, aiming at the similar contents, NAR-IQA-S. [1] applied quality-discriminable image pairs to rank image scores. [2] proposed NAR-IQA-V for the first time, where the contents of the reference and distorted images can be different. In this work, we target at two important aspects of NAR-IQA, the importance of the global understanding and the statistical distance calculation.

C. Fourier Convolution

Our method is also related to Fourier convolution, which maps features to the frequency domain by Fourier transform, followed by convolutions. Fourier convolution can realize the global receptive field without increasing any parameters, which has shown its success in various tasks, including but not limited to, image recognition, action recognition, and image inpainting [15], [29]. In this work, we introduce the Fourier convolution to the task of IQA to obtain global receptive fields.

III. Proposed Method

A. Overall Architecture

Figure 2 shows our pipeline, which takes a reference image and a distorted image as input and predicted score of the distorted image as output. First, we send reference patches $P_R = \{p_{R_i}\}(i = 1, \ldots, m)$ and distorted patches $P_D = \{p_{D_i}\}(i = 1, \ldots, m)$ extracted from reference and distorted images into a pre-trained network for the feature extraction. ResNet50 [30] pre-trained on ImageNet [31] is adopted. Second, the features with n different scales are obtained from different blocks of the ResNet50: $F_R = \{f_{R_i}^j\}(i = 1, \ldots, m; j = 1, \ldots, n)$, $F_D = \{f_{D_i}^j\}(i = 1, \ldots, m; j = 1, \ldots, n)$, and $[C, W, H] = \{[c_1, w_1, h_1], \ldots, [c_n, w_n, h_n]\}$ in Fig. 2 means the set of channel, width and height of the features. Then F_R and F_D will be sent to our proposed multi-scale dual-level feature extractor (MDFE).

In MDFE, we extract the features of each scale at local and global level, and then rescale them to the same scale through adaptive average pooling (AdaAvgPooling), yielding F_R^l, F_R^g, F_D^l, and F_D^g, where l represents the local and g represents the global. These features are all in the same real space $\mathbb{R}^{mn \times c \times w \times h}$, where c represents the number of feature channels, and w and h represent the width and height of the features. Third, we calculate the distance matrices of F_R^l and F_D^l to obtain the $D^l \in \mathbb{R}^{mn \times c \times c}$ in the local level by MMD, and calculate the distance matrices of F_R^g and F_D^g to obtain the $D^g \in \mathbb{R}^{mn \times c \times c}$ in the global level through MMD. It is worth noting that we do not directly calculate the MMD of the features, but use the intermediate result as the distance matrices. Finally, the distance matrices are sent to a regression network to obtain the final score. In the following, we will describe the details.

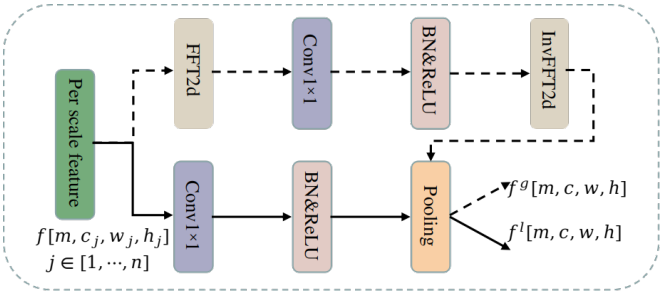

Fig. 3. The structure of MDFE. The features with different scale perform the same operations as shown in this figure. Fourier convolution is used in the above branch to extract global features, and another branch extracts local features through classical convolution.

B. Multi-scale Dual-level Feature Extractor

As shown in Fig. 3, our proposed MDFE extracts features through two branches at each scale. One of the branches is batch normalized after a 1×1 convolution and then activated by ReLU, which performs feature extraction at the local level. The other branch applies Fourier transform to perform the spectral transformation. We adopt the Fourier Unit module of Fast Fourier Convolution [15], which consists of two steps, first we apply Real FFT2d to the features:

$$RealFFT2d : \mathbb{R}^{c \times w \times h} \rightarrow \mathbb{C}^{c \times \frac{w}{2} \times h} \quad (1)$$

Second, real and imaginary parts are concatenated:

$$ComplexToReal : \mathbb{C}^{c\times \frac{w}{2}\times h} \rightarrow \mathbb{R}^{2c\times \frac{w}{2}\times h} \tag{2}$$

followed by a 1×1 convolution. Similarly, we also batch normalize the features and activate them by ReLU on this branch. Finally, we convert the generated features back by inverse Fourier transform, which also consists of two steps, first we split real and imaginary parts up:

$$RealToComplex : \mathbb{R}^{2c\times \frac{w}{2}\times h} \rightarrow \mathbb{C}^{c\times \frac{w}{2}\times h} \tag{3}$$

Second, the inverse Fourier transform is applied:

$$InverseRealFFT2d : \mathbb{C}^{c\times \frac{w}{2}\times h} \rightarrow \mathbb{R}^{c\times w\times h} \tag{4}$$

According to the spectral convolution theorem, the global information of the entire feature map are contained in this branch. The results obtained by MDFE with different sizes will be converted to the same size through AdaAvgPooling.

C. Calculation of Distance Matrices

Unlike previous methods, which directly calculate difference between features, we introduce MMD to calculate their statistical distances in terms of their distributions. According to the MMD, if and only if all their moments of any order are the same, the two distributions are identical. Usually, we cannot solve MMD directly, but we can solve the square of MMD through the kernel embedding. Specifically, given two distributions: $X = \{x_1, x_2 \ \ldots, x_p\}$ and $Y = \{y_1, y_2, \ldots, y_q\}$, the square of MMD is calculated as:

$$\begin{aligned} MMD^2[X, Y; \mathcal{F}] =& \frac{1}{p^2}\sum_{i=1,j=1}^{p} k(x_i, x_j) + \frac{1}{q^2}\sum_{i=1,j=1}^{q} k(y_i, y_j) \\ &- \frac{2}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} k(x_i, y_j), \end{aligned} \tag{5}$$

where k and $\mathcal{F}$ represent the reproducing kernel and the unit ball for a reproducing kernel Hilbert space, respectively. Note that we have $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$, where $\Phi(x) = k(x, \cdot)$. When the number of elements in the two distributions is the same, Eq. 5 can be simplified as:

$$\begin{aligned} MMD^2[X, Y; \mathcal{F}] = \frac{1}{p^2}\sum_{i=1,j=1}^{p} &(k(x_i, x_j) + k(y_i, y_j) \\ &- 2k(x_i, y_j)) \end{aligned} \tag{6}$$

Eq. 6 can be viewed as the average of all elements of a matrix M, where the i-th row and j-th column of the M can be expressed as:

$$M_{i,j} = k(x_i, x_j) + k(y_i, y_j) - 2k(x_i, y_j) \tag{7}$$

We use a function φ to represent the process of obtaining the matrix M, distances matrices can be generated by:

$$\begin{aligned} D_i^l &= \varphi(F_{D_i}^l, F_{R_i}^l) \\ D_i^g &= \varphi(F_{D_i}^g, F_{R_i}^g), \end{aligned} \tag{8}$$

where $i \in (1, 2, \ldots, mn)$ represents the i-th feature. We treat a feature $F \in \mathbb{R}^{c\times w\times h}$ as a distribution of the length c, where each element has length $w \times h$. That is, the feature on each channel is used as an element in the distribution.

D. Regression

Finally, the generated distance matrices will be used to calculate the final score through three MLPs f, g and ψ:

$$\widetilde{y} = \psi(f(D^l), g(D^g)), \tag{9}$$

where f and g map D^l and D^g into two vectors, which are concatenated and mapped into a score through ψ. We use L_1 loss to supervise $\widetilde{y}$ and the ground truth y to train the model:

$$loss = |\widetilde{y} - y| \tag{10}$$

IV. Experiments

A. Experiment Settings

1) Datasets: Four IQA datasets including LIVE [19], CSIQ [20], TID2013 [21] and KADID-10k [22] are used in this paper. The specifics of them are shown in Table II. Among them, LIVE and CSIQ are two small-scale datasets, TID2013 is a medium-scale dataset and KADID-10k is a large-scale dataset. In these datasets, TID2013 and KADIA-10k use the mean opinion score (MOS) as the labels, while CSIQ and LIVE use DMOS, where DMOS is inversely proportional to MOS. We follow previous works [13], [18] and randomly divide every dataset into training, validation, and test sets according to the ratio of 60%, 20%, and 20%.

For NAR-IQA-S, we refer to the practice in [11] and rescale and rotate the original corresponding reference images then use these rescaled and rotated images as references. The ranges of rescale factor and rotation are $[0.95, 1.05]$ and $[-5°, 5°]$, respectively. In NAR-IQA-V, we refer to [2] and randomly select images for training and testing from 800 images in the training set and 100 images in the test set of DIV2K-HR [42], respectively.

2) Evaluation Criterias: We use the Spearman's rank order correlation coefficient (SRCC), and Pearson's linear correlation coefficient (PLCC) as evaluation criterias, which are widely used in the previous IQA methods [13], [18]. Among them, SRCC pays attention to whether the ranking of predicted scores is correct, while PLCC pays attention to whether the predicted score is close to the ground truth.

B. Implementation Details

The number of input patches m is set to 10, the size of each patch is 224×224, and the scales of multi-scale features n are set to 4. In all experiments, we utilize random horizontal and vertical flips as data augmentation during training. The initial learning rate is set to 2×10^{-5}, and we use the ADAM optimizer for training. All experiments are done on NVIDIA-2080ti GPUs.

C. Comparisons with the State-of-the-art Methods

We compare our method with other methods under three experimental settings: FR-IQA, NAR-IQA-S, and NAR-IQA-V. In FR-IQA, we compare some well-known and widely used methods such as PSNR, SSIM [16], and VIF [32], and some representative and recent methods, including PieAPP [17],

TABLE I
COMPARISON OF OUR METHOD AND OTHER METHODS IN THREE IQA TASKS, THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE UNDERLINED. '-' MEANS NOT APPLICABLE.

IQA Type	Method	LIVE		CSIQ		TID2013		KADID-10k	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR-IQA	PSNR	0.873	0.865	0.810	0.819	0.687	0.677	0.676	0.675
	SSIM [16]	0.948	0.937	0.865	0.852	0.727	0.777	0.724	0.717
	VIF [32]	0.964	0.960	0.911	0.913	0.677	0.771	0.679	0.687
	PieAPP [17]	0.918	0.909	0.890	0.873	0.670	0.749	0.836	0.836
	LPIPS [33]	0.932	0.934	0.903	0.927	0.670	0.749	0.843	0.839
	IQT [13]	0.970	-	0.943	-	0.899	-	0.948	0.950
	DISTS [4]	0.955	0.955	0.946	0.946	0.830	0.855	0.887	0.886
	ISPL-FR [18]	0.970	0.978	0.965	0.968	0.924	0.912	0.944	0.943
	CVRKD [2]	0.960	0.965	0.958	0.965	0.928	0.935	0.957	0.959
	Ours	**0.979**	**0.982**	**0.977**	**0.983**	**0.956**	**0.967**	**0.972**	**0.974**
NR-IQA	BRISQUE [34]	0.939	0.935	0.746	0.829	0.604	0.694	-	-
	FRIQUEE [35]	0.940	0.944	0.835	0.874	0.680	0.753	-	-
	BMPRI [36]	0.931	0.933	0.909	0.934	0.928	0.947	-	-
	BIECON [37]	0.961	0.962	0.815	0.823	0.717	0.762	-	-
	WaDIQaM-NR [38]	0.954	0.963	-	-	0.761	0.787	-	-
	DIQaM-NR [38]	0.960	**0.972**	-	-	0.835	0.855	-	-
	IW-CNN [39]	0.963	0.964	0.812	0.791	0.800	0.802	-	
	DB-CNN [40]	**0.968**	0.971	**0.946**	**0.959**	0.816	**0.865**	0.501	0.569
	HyperIQA [41]	0.962	0.966	0.923	0.942	**0.840**	0.858	**0.852**	**0.845**
NAR-IQA-S	IQT [13]	0.946	0.951	0.909	0.919	0.918	0.930	0.931	0.935
	CVRKD [2]	0.951	0.954	0.952	0.954	0.926	0.932	0.954	0.957
	Ours	**0.976**	**0.974**	**0.963**	**0.977**	**0.953**	**0.961**	**0.970**	**0.973**
NAR-IQA-V	IQT [13]	0.899	0.895	0.894	0.917	0.729	0.771	0.825	0.838
	CVRKD [2]	0.955	0.960	0.827	0.874	0.815	0.853	0.911	0.919
	Ours	**0.957**	**0.963**	**0.931**	**0.940**	**0.822**	**0.872**	**0.921**	**0.927**

TABLE II
THE INTRODUCTION TO THE IQA DATASETS. #REF., #DIS., AND #DIS.TYPE REPRESENT THE NUMBER OF REFERENCE IMAGES, DISTORTED IMAGES, AND DISTORTION TYPES, RESPECTIVELY.

Dataset	#Ref.	#Dis	#Dis.Type	Score Range
LIVE	29	779	5	[0,100]
CSIQ	30	866	6	[0,1]
TID2013	25	3000	24	[0,9]
KADID-10k	81	10125	25	[1,5]

LPIPS [33], IQT [13], DISTS [4], ISPL [18] and CVRKD [2]. As shown in Table I, our method outperforms other methods on all four datasets, especially on two larger datasets, TID2013 and KADID-10k.

There are only two influential works DCNN [11] and CVRKD [2] for NAR-IQA so far. Among them, DCNN proposed concept of NAR-IQA-S for the first time. However we do not compare with it due to the fact that it is not open-sourced, and the implementation is hard due to insufficient details. Moreover, it did not report on the dataset as we adopted. Therefore, we compare with CVRKD and some recently published methods. In NAR-IQA-S, our method not only outperforms other methods, but also performs even better than the results of other methods in FR-IQA. From Table I, we can see that the results obtained by all methods have a decline compared with the results of FR-IQA, but the decline is not too much. This shows that IQA can also be done well when the contents of the reference and distorted images are similar.

NAR-IQA-V is a recently proposed form of IQA in CVRKD [2]. CVRKD mainly uses knowledge distillation to transfer the knowledge learned in FR-IQA to NAR-IQA. For fairness, we compare other methods without knowledge distillation in Table I, and we will compare the performance of our method and other methods when both use knowledge distillation in the ablation study. Our method outperforms other methods, because they still uses the feature extraction method and distance measurement method used in the FR-IQA task, which is inappropriate in NAR-IQA. From Table I, we can see that compared with FR-IQA and NAR-IQA-S, the accuracy of NAR-IQA-V has dropped a lot. We found that NAR-IQA-V is almost the same level as NR-IQA on small-scale and medium-scale datasets, but NAR-IQA-V performs much better than NR-IQA on the large-scale dataset KADID-10k. The data volume of CSIQ is small and its distortion types are relatively few, so many methods are easy to overfit. However, our approach comprehensively considers the local and global statistical distances to better distinguish some subtleties. Therefore, our method significantly outperforms other methods on the NAR-IQA-V task on this dataset. As mentioned before, different kinds of IQAs have different advantages and disadvantages, which can be applied to different needs.

D. Ablation Study

1) Effectiveness of Fourier Convolution in MDFE: We separately test the effectivenes of different branches in MDFE for three types of IQA tasks. As shown in Table III: in FR-IQA and NAR-IQA-S, either using classical convolution alone or using Fourier convolution alone can make accurate predictions. When they are both used at the same time, the local information and the global information are complemented, which can further improve the accuracy.

For NAR-IQA-V: we find that using Fourier convolution alone performs better than just using classical convolution. This is because distorted and reference images are not aligned,

TABLE III
SRCC OF USING DIFFERENT BRANCH FORMS IN MDFE. THE L REPRESENTS THE BRANCH THAT USES CLASSICAL CONVOLUTION TO EXTRACT LOCAL FEATURES, AND G REPRESENTS THE BRANCH THAT USES FOURIER CONVOLUTION TO EXTRACT GLOBAL FEATURES.

IQA Type	L	G	Live	CSIQ	TID	KADID
FR-IQA	✓	✗	0.968	0.966	0.914	0.962
	✗	✓	0.964	0.966	0.937	0.966
	✓	✓	**0.979**	**0.977**	**0.956**	**0.972**
NAR-IQA-S	✓	✗	0.958	0.940	0.942	0.962
	✗	✓	0.957	0.947	0.922	0.958
	✓	✓	**0.976**	**0.963**	**0.953**	**0.970**
NAR-IQA-V	✓	✗	0.945	0.890	0.783	0.905
	✗	✓	0.948	0.893	0.802	**0.935**
	✓	✓	**0.957**	**0.931**	**0.822**	0.921

TABLE IV
COMPARISON OF RESULTS USING DIFFERENT WAYS TO MEASURE THE DISTANCE BETWEEN DISTORTED AND REFERENCE IMAGES. THE *sub.* INDICATES THAT THE ACQUIRED FEATURES ARE DIRECTLY SUBTRACTED, WHILE KLD AND SWD MEAN KULLBACK-LEIBLER DIVERGENCE AND SLICED WASSERSTEIN DISTANCE.

IQA Type	Method	CSIQ		KADID-10k	
		SRCC	PLCC	SRCC	PLCC
FR-IQA	*Sub.*	0.955	0.960	0.949	0.952
	L_1	0.973	0.978	0.965	0.967
	L_2	0.971	0.974	0.964	0.965
	cos	0.973	0.976	0.964	0.966
	KLD	0.970	0.975	0.968	0.958
	SWD	0.967	0.978	0.966	0.969
	MMD	**0.977**	**0.983**	**0.972**	**0.974**
NAR-IQA-S	*Sub.*	0.963	0.966	0.949	0.946
	L_1	0.954	0.959	0.964	0.966
	L_2	0.955	0.967	0.954	0.960
	cos	0.945	0.961	0.923	0.929
	KLD	0.957	0.970	0.954	0.958
	SWD	0.945	0.954	0.955	0.957
	MMD	**0.963**	**0.977**	**0.970**	**0.973**
NAR-IQA-V	*Sub.*	0.897	0.921	0.889	0.899
	L_1	0.877	0.922	0.894	0.903
	L_2	0.918	0.932	0.918	0.922
	cos	0.830	0.866	0.914	0.921
	KLD	0.854	0.895	0.911	0.923
	SWD	0.901	0.921	0.868	0.876
	MMD	**0.931**	**0.940**	**0.921**	**0.927**

and only extracting features at the local level cannot extract enough information to determine the distortion degree of distorted images. It is always beneficial to combine local and global information on small-scale and medium-scale datasets, *i.e.*, LIVE, CSIQ and TID2013. However, for the large-scale dataset KADID-10k, only using Fourier convolution in MDFE is more effective than using both convolution forms at the same time. This is because the features from distorted and reference images are often completely irrelevant in the NAR-IQA-V task, which sometimes leads to some interference with the prediction. This kind of interference is especially serious when the amount of data is relatively large.

2) Effectiveness of MMD: In Table IV, we show the effectiveness of using different ways to measure the distance between distorted and reference images on the results. We selected several common methods such as L_1, L_2, and Kullback-Leibler divergence to compare. The results show that MMD can achieve better results especially in NAR-IQA. This is because methods such as L_1 and L_2 measure the distance too roughly and Kullback-Leibler divergence is limited by the rationality of density estimation. For sliced Wasserstein distance, because it sometimes cannot find similar features in the NAR-IQA task, it cannot solve the distance of distribution transformation well, so its performance is not particularly good.

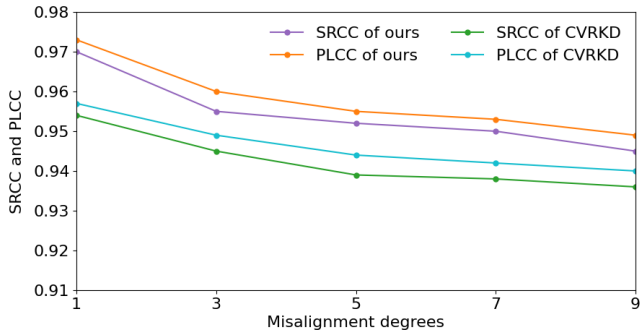

Fig. 4. SRCC and PLCC of KADID-10k using references with different degrees of misalignment in NAR-IQA-S.

TABLE V
SRCC AND PLCC OF DIFFERENT TYPES OF DISTORTIONS IN THE CSIQ. IN THE TABLE, AWGN MEANS ADDITIVE WHITE GAUSSIAN NOISE, FNOISE MEANS ADDITIVE PINK GAUSSIAN NOISE, AND JPEG AND JPEG2000 ARE TWO KINDS OF COMPRESSIONS.

Dis. Type	FR-IQA		NAR- IQA-S		NAR- IQA-V	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
awgn	0.975	0.979	0.946	0.961	0.852	0.849
blur	0.968	0.982	0.954	0.978	0.937	0.936
contrast	0.956	0.958	0.955	0.957	0.813	0.855
fnoise	0.982	0.989	0.956	0.963	0.924	0.921
jpeg	0.955	0.981	0.913	0.979	0.921	0.966
jpeg2000	0.953	0.984	0.968	0.987	0.970	0.980

3) Comparison of Results at Different Misalignment Degrees in NAR-IQA-S: In order to analyze the impact of misalignment degree, we set a variety of the misalignment degrees, which are denoted as r, and the corresponding rescaling range and rotation range are: $[1 - \frac{5\times r}{180}, 1 + \frac{5\times r}{180}]$ and $[-5^\circ \times r, 5^\circ \times r]$. As shown in Fig. 4, when the degree of misalignment increases, the accuracy of both our method and CVRKD decreases, but both SRCC and PLCC of ours are higher than those of CVRKD for each degree. When $r = 9$, we find that the accuracy is close to the result of NAR-IQA-V, which means that it is difficult to find the corresponding relationship between reference and distorted images when the degree of misalignment is too large. So if we want to get more accurate results, we still need reference images that are similar enough to the distorted images.

4) Evaluation with Different Non-aligned References: To more intuitively show the impact of different NAR images on the IQA task, we use all the images in the test-set of DIV2K-HR [42] to be combined with the distorted images in KADID-10k for testing. Then for the 7 distortion types in KADID-10k: blur, color distortion, compression, noise, brightness, and sharpness & contrast distortion, we show some examples in Fig. 5. In Fig. 5, the leftmost column are the distorted images to be tested and their corresponding mean opinion scores (MOSs), and the right is the results obtained using different NAR references, where i-th represents the score that is i-th

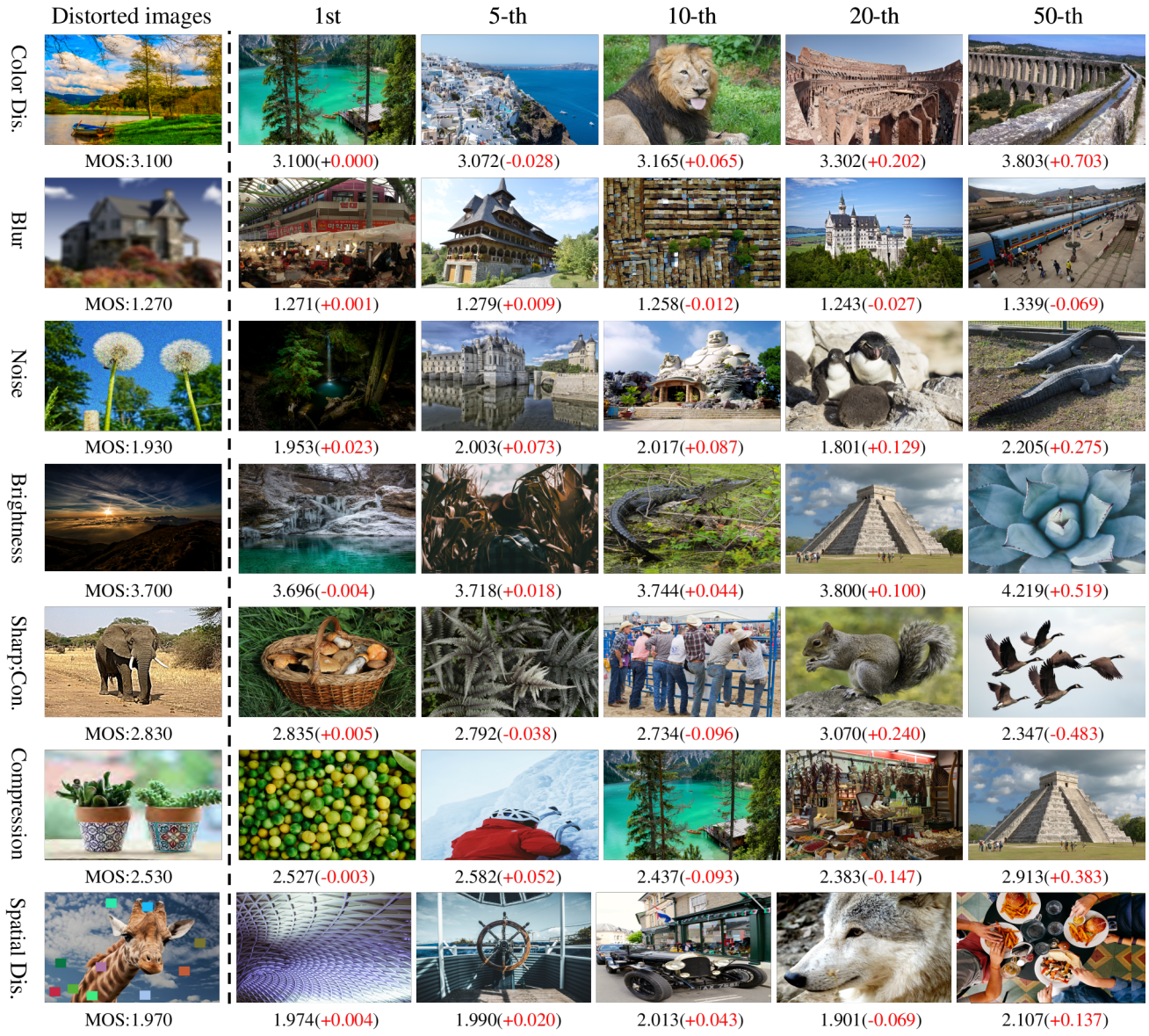

Fig. 5. Comparison of prediction results using different NAR. In the figure, Dis. represents distortion, and Con. represents contrast.

closest to the ground truth in the scores obtained by using all the images in the test set of DIV2k-HR to test respectively. The results show that the choice of different reference images has a great impact on the results, and it seems that there is no regularity in which image is the most suitable as the reference image.

5) Comparison of Results with Different Types of Distortions: As shown in Table V, we counted the respective accuracy of different distortion types under the three IQA forms. In FR-IQA, our method can give accurate predictions for various distortions. For NAR-IQA-S compared to FR-IQA, the accuracy of the blur, contrast, and jpeg2000 distortions is almost unchanged. However, for jpeg, awgn and fnoise, the accuracy drops significantly, which shows that these distortions have stricter alignment requirements. For NAR-IQA-V, the prediction accuracy of almost all distortions has dropped a lot, especially for images with contrast distortion. This is because some small changes in contrast are difficult to measure whether it is unreasonable, only it can be found by comparing them with their corresponding references. At the same time, we find that the accuracy of images with jpeg2000 distortion in the three IQA forms has almost no change, indicating that the evaluation of the degree of this distortion does not have strict requirements for whether the contents of the reference images and distorted images are similar.

6) Comparison of NAR-IQA-V Results with Knowledge Distillation: For a more comprehensive comparison, we further uses knowledge distillation in the NAR-IQA-V task. In the process of knowledge distillation, the knowledge in a teacher network N_T trained in the FR-IQA task is transferred to a student network N_S, so that N_S can use the NAR for image quality assessment.

Like CVRKD, we do not jointly train N_T and N_S, but first use distorted patches and fully aligned reference patches

TABLE VI
COMPARISON OF OUR METHOD AND CVRKD FOR THE NARIQA-V TASK WHEN BOTH USE KNOWLEDGE DISTILLATION.

Loss	LIVE		CSIQ		TID2013		KADID-10k	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CVRKD	0.958	0.962	0.885	0.914	0.831	0.862	0.945	0.950
ours	**0.968**	**0.971**	**0.931**	**0.943**	**0.835**	**0.865**	**0.950**	**0.955**

TABLE VII
COMPARISON OF RESULTS UNDER DIFFERENT SETTINGS IN NAR-IQA-V. THE TOP-i MEANS THAT ALL IMAGES IN THE TEST SET OF DIV2K-HR ARE USED FOR TESTING RESPECTIVELY AND THEN THE AVERAGE OF i PREDICTION RESULTS WHICH ARE CLOSEST TO THE GROUND TRUTH IS USED AS THE FINAL SCORE.

Setting	Method	CSIQ		KADID-10k	
		SRCC	PLCC	SRCC	PLCC
top1	CVRKD	0.955	0.970	0.993	0.993
	ours	**0.993**	**0.997**	**0.998**	**0.999**
top5	CVRKD	0.944	0.961	0.989	0.990
	ours	**0.989**	**0.994**	**0.998**	**0.998**
top10	CVRKD	0.934	0.952	0.986	0.987
	ours	**0.986**	**0.992**	**0.996**	**0.997**

$\{P_D, P_R\}$ to train N_T according to Eq. 10.

Then we fix the parameters of N_T and use N_T to guide N_S learning. Given a set of distorted patches and NAR patches $\{P_D, P_{NAR}\}$ as input, N_S can obtain a set of internal features $\{f_f^S, f_g^S\}$, where f_f^S and f_g^S are obtained by f and g in Eq. 9, respectively. At the same time, the fixed N_T uses $\{P_D, P_R\}$ as input to also obtain internal features$\{f_f^T, f_g^T\}$. Finally, L_2 loss is used to transfer knowledge from N_T to N_S:

$$L_{KD_d} = \|f_f^S - f_f^T\|_2 + \|f_g^S - f_g^T\|_2 \tag{11}$$

Except L_{KD_d}, the label loss L_{KD_l} between the prediction of N_s and ground truth y is also used to optimize N_S:

$$L_{KD_l} = |N_s(P_D, P_{NAR}) - y| \tag{12}$$

The final loss L_{KD} of N_S is the combine of the L_{KD_d} and L_{KD_l}:

$$L_{KD} = L_{KD_d} + L_{KD_l} \tag{13}$$

We compare our method with CVRKD when both use knowledge distillation. The experimental results are shown in the Table VI, our method outperforms CVRKD on all 4 datasets, and the leading degree is greater than that in Table I, which indicates that our method can make knowledge transfer better by obtaining a larger receptive field and using statistical distance to measure the distance.

7) Comparison of Results under Different Settings in NAR-IQA-V: In NAR-IQA-V, we randomly select a reference image in the test-set of DIV2K-HR [42] for a distorted image. In this way we can get the performance of our model on the NAR-IQA-V task in general. Here we test the upper limit of our method on the NAR-IQA-V task. Specifically, given a distorted image, we use each image in the test-set of DIV2K-HR as the reference and then predict a set of scores. Then the average of i scores closest to the ground truth among them is used as the final prediction, where $i \in [1, 5, 10]$. As shown in Table. VII, our method achieves very high accuracy in all three settings, whether using the small-scale CSIQ [20] or the large-scale KADID-10k [22]. Compared with CVRKD [2], our method performs better in all three settings, especially on the small-scale dataset CSIQ. This shows that our method is not only more accurate than other methods in general, but also the upper limit of our method is higher. However, in reality, in the NAR-IQA-V task, we do not know which reference images are better to choose as reference images, so the data in this experiment only presents the level that the NAR-IQA-V task can achieve under ideal conditions. Further exploration is needed to achieve such accurate predictions.

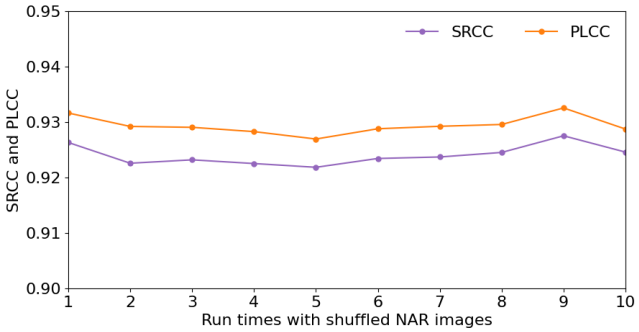

Fig. 6. Stability evaluation of the proposed method. The figure shows the results of 10 tests using randomly selected non-aligned references under NAR-IQA-V task.

TABLE VIII
COMPARISON WITH USING DIFFERENT LOSS ON KADID.

Loss	FR-IQA		NAR-IQA-S		NAR-IQA-V	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
L_2	0.968	0.970	0.952	0.954	0.907	0.915
L_1	**0.972**	**0.974**	**0.970**	**0.973**	**0.950**	**0.955**

8) Stability about Using Different Non-aligned References: To test the robustness of the proposed method, we randomly select references from the test set of DIV2K-HR and conduct 10 tests on KADID-10k. Fig. 6 shows the results, we can find that the fluctuation of results is very small, which shows that our method has good robustness in the case of choosing different non-aligned references.

9) Comparison of Results Using Different Loss Functions: As shown in Table VIII, we achieve better results using $L1$ loss when training with different losses. This is because the types of distortion between different distorted images are sundry, and the degrees of distortion are also complex, while L_1 loss is not easily affected by outliers and is more robust.

10) Evaluation with Different Number of Input Patches: As shown in Fig. 7, we conduct experiments on KADID-10k using different number of input patches. The accuracy is relatively low when using few patches, and the prediction accuracy increases with the increase of the input patches number under the three IQA tasks. For FR-IQA and NAR-IQA-S, when the number of patches increases to 7, the accuracy will unchanged when increasing the number of patches. For NAR-IQA-V, when the number of patches reaches 13, then increasing the number of patches will have no obvious benefit. Since the more patches are inputted, the more time-consuming the model is, we use 10 patches as input in all three IQA tasks, considering both efficiency and accuracy.

11) Stability about Random Selection of Input Patches: This paper uses randomly cropping to get patches as input. To

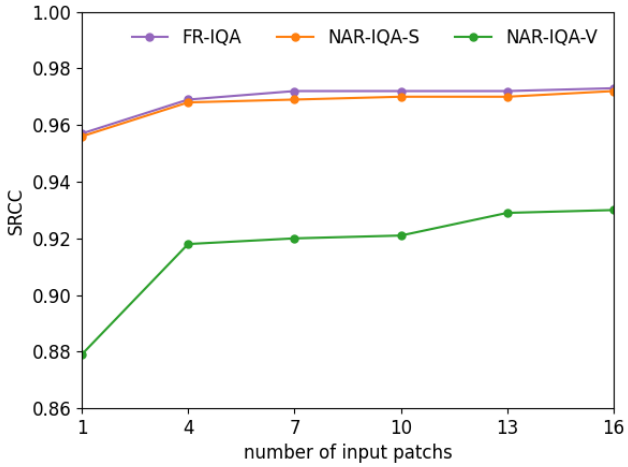

Fig. 7. SRCC of KADID-10k while the number of input patches is different.

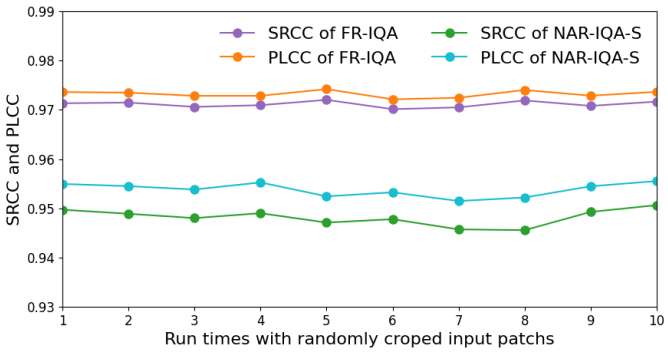

Fig. 8. Stability evaluation of the proposed method. The figure shows the results of 10 tests using randomly cropped inputs under the FR-IQA and NAR-IQA-S tasks.

test the robustness of this approach, we conduct 10 tests with different combinations of patches randomly under the FR-IQA and NAR-IQA-S settings. To control the singleness of random variables, we fixed the rescaling ratio of the reference image to 1.05 and the rotation angle to 5° under the NAR-IQA-S setting. As shown in Fig. 8, the fluctuations in the results of multiple experiments are very small, indicating that the method of random cropping is completely feasible.

12) Evaluation with Different Number of Reproducing Kernel: In MMD, we generally use multiple reproducing kernels for projection. We use different numbers of kernels to test on the CSIQ dataset, and the result is shown in the Fig. 9. For FR-IQA, the number of kernels will not have much impact on the result, and there is not much difference between the results of using different numbers of kernels. For NAR-IQA-S, as the number of kernels increases, the SRCC gradually increases, but when the number of kernels increases to 7 the SRCC remains basically unchanged when the number of kernels is increased. For NAR-IQA-V, as the number of kernels increases, the SRCC also increases gradually, but when the number of kernels increases to 5 the SRCC remains basically unchanged when the number of kernels is increased. Since the number of kernels increases, the time-consuming will also increase, so we use 5 kernels in our method.

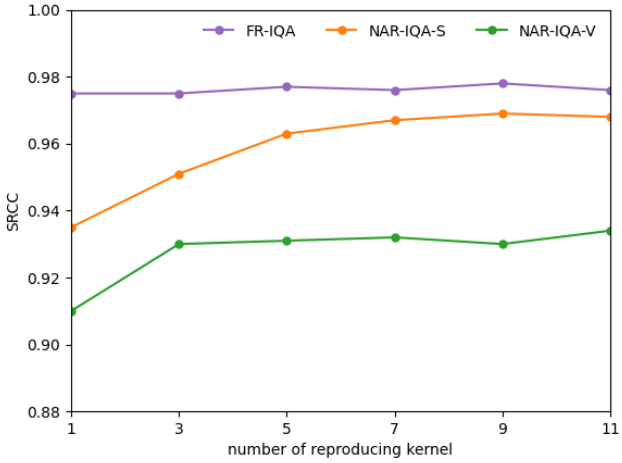

Fig. 9. SRCC of CSIQ while the number of reproducing kernel is different.

TABLE IX
THE SRCC AND PLCC RESULTS WITH DIFFERENT QUALITIES NAR IMAGES. THE MODELS ARE FIXED, AND WE ONLY CHANGE THE TYPES OF REFERENCE IMAGES. THE LAST ROW IS THE RESULT USING THE ORIGINAL REFERENCE IMAGES.

Distortion Type	CSIQ		KADID-10k	
	SRCC	PLCC	SRCC	PLCC
Downsample: $\times 2$	0.918	0.920	0.901	0.915
Downsample: $\times 4$	0.896	0.899	0.881	0.887
Gussian Blur: 3×3	0.924	0.926	0.913	0.919
Gussian Blur: 5×5	0.914	0.921	0.910	0.916
-	**0.931**	**0.940**	**0.921**	**0.927**

13) Evaluation on NAR with Different Quality: Although high-quality images are easy to obtain, it is still necessary to test the results of using different quality images as NAR. As shown in the Table IX, we subject the images in the DIV2K-HR test set to various forms of distortion to varying degrees and test them as reference images. It can be found that slight distortion, especially the reduction of resolution, reduces the accuracy of the evaluation. So we still need to use high-quality NARs as much as possible.

V. CONCLUSION

In this work, we propose a new unified method for IQA. Our proposed MDFE uses classical convolution and Fourier convolution to extract local and global features at the same time, which provide comprehensive references for the degree of image distortion. Furthermore, we analyzes the statistical distance between distorted and reference images using MMD which is more suitable for relaxed renference. Extensive experiments show that our mehtod outperforms other methods in both FR-IQA and NAR-IQA. Remarkably our mehtod using non-aligned references performs even better than other methods using aligned references. In addition, there is no doubt that NAR-IQA is a more convenient form of IQA, but different selection of references will impact the prediction a lot. Therefore, how to select references scientifically is worth studying.

REFERENCES

[1] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.

[2] G. Yin, W. Wang, Z. Yuan, C. Han, W. Ji, S. Sun, and C. Wang, "Content-variant reference image quality assessment via knowledge distillation," *arXiv preprint arXiv:2202.13123*, 2022.

[3] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Trans. on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[4] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[5] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, and C. W. Chen, "Screen content quality assessment: overview, benchmark, and beyond," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–36, 2021.

[6] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[7] Z. Luo, H. Huang, L. Yu, Y. Li, H. Fan, and S. Liu, "Deep constrained least squares for blind image super-resolution," in *Proc. CVPR*, 2022, pp. 17 642–17 652.

[8] S. Cheng, Y. Wang, H. Huang, D. Liu, H. Fan, and S. Liu, "Nbnet: Noise basis learning for image denoising with subspace projection," in *Proc. CVPR*, 2021, pp. 4896–4906.

[9] Z. Liu, W. Lin, X. Li, Q. Rao, T. Jiang, M. Han, H. Fan, J. Sun, and S. Liu, "Adnet: Attention-guided deformable convolutional network for high dynamic range imaging," in *Proc. CVPR*, 2021, pp. 463–470.

[10] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[11] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *Proc. ECCV*, 2016, pp. 3–18.

[12] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. CVPR*, 2021, pp. 10 242–10 251.

[13] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proc. CVPR*, 2021, pp. 433–442.

[14] H. Pratt, B. Williams, F. Coenen, and Y. Zheng, "Fcnn: Fourier convolutional neural networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 786–798.

[15] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Proc. NeurIPS*, vol. 33, pp. 4479–4488, 2020.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *Proc. CVPR*, 2018, pp. 1808–1817.

[18] Y. Cao, Z. Wan, D. Ren, Z. Yan, and W. Zuo, "Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment," in *Proc. CVPR*, 2022, pp. 5851–5861.

[19] H. R. Sheikh, "Image and video quality assessment research at live," *http://live. ece. utexas. edu/research/quality*, 2003.

[20] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, p. 011006, 2010.

[21] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.

[22] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.

[23] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[24] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Trans. on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[25] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. CVPR*, 2014, pp. 1733–1740.

[26] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang, "Blind predicting similar quality map for image quality assessment," in *Proc. CVPR*, 2018, pp. 6373–6382.

[27] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 789–797.

[28] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proc. CVPR*, 2020, pp. 14 143–14 152.

[29] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proc. CVPR*, 2022, pp. 2149–2159.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[32] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. CVPR*, 2018, pp. 586–595.

[34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[35] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of vision*, vol. 17, no. 1, pp. 32–32, 2017.

[36] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[37] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2016.

[38] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[39] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal processing magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[40] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[41] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. CVPR*, 2020, pp. 3667–3676.

[42] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. CVPRW*, 2017, pp. 126–135.