Image Quality Assessment: Measuring Perceptual Degradation via Distribution Measures in a Deep Feature Space

Journal:	*Transactions on Pattern Analysis and Machine Intelligence*
Manuscript ID	TPAMI-2022-11-2222
Manuscript Type:	Regular
Keywords:	Image quality assessment, distribution measures, perceptual degradation

Image Quality Assessment: Measuring Perceptual Degradation via Distribution Measures in a Deep Feature Space

Xingran Liao, Xuekai Wei, Mingliang Zhou, Zhengguo Li, *Senior Member, IEEE*, and Sam Kwong, *Fellow, IEEE*

Abstract—Many deep network-based full-reference image quality assessment (FR-IQA) models use pixelwise measures to compare deep features, determine the severity of perceptual information contamination, and then output objective quality scores. However, the pixelwise comparison process often ignores the pixel correlations among deep features, thus shifting from subjective evaluations of specific distortions. Herein, we propose that deep-based FR-IQA measures should consider pixel correlation fidelity, which compels us to view deep features as statistical distributions and introduce distribution measures to address the pixel correlation issue. Specifically, reference and distorted images are first projected into deep features by a Visual Geometry Group (VGG) network. Then, we test three distribution measures for comparing deep features: the Wasserstein distance (WSD), the symmetric Kullback–Leibler divergence (SKLD), and the Jensen–Shannon divergence (JSD). The experimental results show that these deep network-based distribution measures require no training but are closely related to human subjective evaluations based on both conventional and recently published challenging IQA datasets, indicating that these measures have advanced robustness to different distortions and possess clear interpretability. Moreover, we discuss applications of the proposed measures in image reconstruction tasks and demonstrate their prominent usage in perceptual optimization. The code will be publicly available upon acceptance.

Index Terms—Image quality assessment, distribution measures, perceptual degradation, deep neural network image representations, perceptual optimization.

1 Introduction

With the development of digital devices, the demand for presenting high perceptual quality images is growing. However, digital images always suffer from distortions during the process of capturing, compressing, and transmission; thus, monitoring perceptual quality degradations via automatic measures becomes increasingly necessary. Full-reference image quality assessment (FR-IQA) serves as a perceptual quality measure, providing an automatic method for image quality evaluation. Some FR-IQA measures are even used as perceptual losses and applied in different image reconstruction tasks, such as denoising, dehazing, and image superresolution adjustment [1]–[3], which always provide more visually satisfactory reconstruction results.

An effective FR-IQA measure should be sensitive to the degradation of perceptual quality and correspond to the human visual system (HVS). This issue requires a comparison between the reference and the distorted images to be conducted in a perceptual manner [7]. The traditional mean square error (MSE) fails to be a perceptual metric because it measures the difference through a pixelwise comparison in the pixel domain, and it is widely accepted that comparing images pixel by pixel does not lead to a perceptual model [8], [9]. According to research in natural scene statistics [10], a direct way to establish an effective FR-IQA model is to compare the differences among perceptual features, such as luminance, contrast, and structure. A notable example is the structural similarity index (SSIM) [11], which compares the similarities among these perceptual features and is more perceptual than the MSE in terms of subjective quality score predictions. However, the SSIM [11] is still not a perceptual measure for some distortion types, and people tend to explore various perceptual features to construct more robust FR-IQA measures. As such, the given raw images are projected into spatial, frequency, or deep feature domains via different transformations. These features are usually compared via the MSE or the similarity index [11]. Currently, building a perceptual FR-IQA measure by comparing deep features is flourishing. Existing works [12], [13] have shown that FR-IQA measures based on deep feature comparisons are more robust than those based on spatial or frequency feature comparisons. The underlying reason for this finding may be that deep neural networks capture the perceptual features that govern the perceptual quality degradations caused by various distortions, as shown in Fig. 1. Specifically, distortions induced in the pixel domain are reflected in the deep feature domain, and these deep features contain perceptual information, which is important

- *Xingran Liao and Sam Kwong are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: Xingrliao2-c@my.cityu.edu.hk, cssamk@cityu.edu.hk).*
- *Xuekai Wei is with the State Key Laboratory of Internet of Things for Smart City and the Department of Electrical and Computer Engineering, University of Macau, Macao, China (e-mail: xuekaiwei2-c@my.cityu.edu.hk).*
- *Mingliang Zhou is with the Department of Computer Science, Chongqing University, Chongqing China (e-mail: mingliangzhou@cqu.edu.cn).*
- *Zhengguo Li is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632, Singapore (e-mail: ezgli@i2r.a-star.edu.sg).*

Manuscript received XXXXXX; revised XXXXXX.

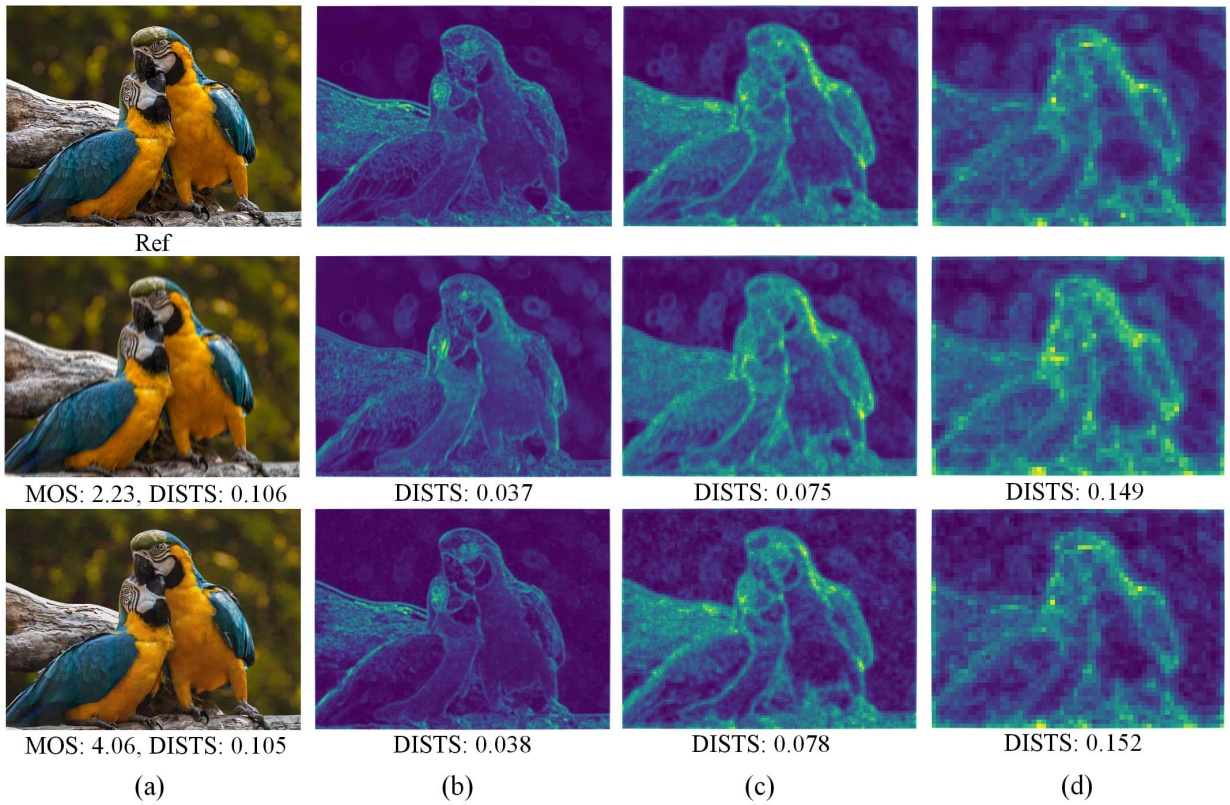

Fig. 1. Visualization of the deep network features derived from an original image and two distorted images from the KADID-10k datasets [4]. The second-row images are contaminated by motion blur distortions, and the third-row images are contaminated by denoising distortions. (a) The original image and the distorted images with their mean opinion scores (MOS) and deep image structure and texture similarity (DISTS) scores listed below. (b)(c)(d) The first-, second-, and third-stage deep features of the VGG16 network [5] with the deep-stage image structures and texture similarity scores [6] listed below.

for conducting subjective evaluations. For instance, Ding *et al.* [6], [14] empirically concluded that the deep features of the Visual Geometry Group 16 (VGG16) [5] network contain structure and texture information, which plays a significant role in subjective quality assessments. However, because the pixels in deep features still have strong correlations, as shown in Fig. 1, the use of pixelwise measures such as the MSE may lead to inferior results in the quality score prediction task. Moreover, Liao *et al.* [15] also reported that using an SSIM variant [16] to compare deep features is not a robust method for some distortion samples, as shown in Fig. 1, and leads to wave-like artefacts in perceptual optimization cases. Therefore, it is urgent to consider applying more powerful metrics in measuring the perceptual quality degradations of the deep network features and avoiding generating artefacts in perceptual optimization. However, works on which kind of measures are good for quality assessment and perceptual optimization are rather barren, which poses a restriction on the further development of the FR-IQA measures.

To address this problem, we focus on designing a new FR-IQA measure that addresses both the pixel correlation issue and perceptual optimization [17]. The core issues in our FR-IQA design are to explore the determined kinds of measures that are suitable for deep feature comparisons and to reveal the underlying reasons behind our approach. Specifically, we show that even without a training process, statistical distribution-based measures can provide better perceptual comparisons and attain superior prediction results in deep feature comparisons. We believe the reason is that these measures take the pixel correlation of deep features into account, and in our approach, we first project the reference and distorted images into deep feature domains via the VGG19 network [5]. Then, we test three widely used distribution measures for comparing deep features: the Wasserstein distance (WSD) [18], the Jensen–Shannon divergence (JSD) [19], and the symmetric Kullback–Leibler divergence (SKLD) [20]. Such a strategy leads to three independent deep network-based FR-IQAs, which are deep network-based WSD (DeepWSD), deep network-based JSD (DeepJSD), and deep network-based SKLD (DeepSKLD). Finally, a training-free logarithm function is used to pool the difference scores among various stages to form a perceptual quality score. In conclusion, our contributions are threefold.

- We propose a new design philosophy for a deep network-based FR-IQA model that pursues pixel correlation fidelity in the deep feature domain. This philosophy is closely related to the efficient coding theory of the HVS and compels us to introduce distribution measures, *i.e.*, the WSD, the JSD, and the SKLD, to compare deep features. Such a strategy also enables the model to become more robust to different distortion types.
- We propose FR-IQA measures that utilize distribution measures to estimate perceptual degradation in

the VGG19 deep feature domain; this approach can tackle the pixel correlation issue when comparing deep features. Moreover, we design the proposed measures without a training process but can achieve satisfactory quality assessment results on several IQA datasets, which indicates that these measures are closely related to human subjective evaluations and are robust to various distortions with clear interpretability. We also perform different perceptual experiments to show that the proposed methods are more perceptual than many existing methods, achieving state-of-the-art quality prediction and maximum differential competition (MDC) [21] results.

- We also demonstrate that the proposed measures can be applied in some image reconstruction tasks to obtain more visually satisfactory results, which reveals the great potential for the proposed measures to act as perceptual losses.

The rest of the thesis is structured as follows. In **Section 2**, we review the background of FR-IQA and various distribution measures. In **Section 3**, we present the design philosophy for the deep-based FR-IQA measures. In **Section 4**, we explain the methodology of DeepWSD, DeepSKLD, and DeepJSD in detail. In **Section 5**, we present the experimental results, including the quality prediction results, MDC experimental results, and ablation study results. In **Section 6**, we discuss the application of the proposed approach in image superresolution and image denoising tasks. In **Section 7**, we conclude this work.

2 Related work

2.1 FR-IQA model

Driven by different philosophies, considerable efforts have been expended to build FR-IQA measures that are more perceptual [22]. Beyond comparing the structural similarity, the DISTS index [6] compares the texture similarity through feature projection via the VGG16 network [5] and has become more perceptual. Another way to build a perceptual FR-IQA measure is to transform an image into the frequency domain and compare differences between the subband coefficients. The underlying philosophy involves mimicking the shallow visual process of the HVS [22]. Many frequency transformations, such as the log-Gabor filter [23], [24] and wavelet transformation [25], have been proven to be in accord with the mechanisms of some forehead brain lobes. The feature similarity index measure (FSIM) [26] adopts the log-Gabor filter to detect local phase differences and compare gradient magnitude differences on a global scale, thereby attaining satisfactory quality prediction results. To explore the relationship between distortion strength and perceptual quality degradation, the image information, visual quality index (VIF) [27] and information fidelity criterion (IFC) [28] are used to predict perceptual quality by gauging the perceptual information losses caused by distortions. The input images are built as Gaussian-scale mixture models and decomposed by wavelet transformations. Then, the mutual information is used to quantify the perceptual information losses among the wavelet subbands and output the final score. To fully utilize the low-level image features and achieve enhanced computational efficiency, the gradient magnitude similarity deviation (GMSD) [29] computes the gradient difference between the reference and distorted images and has achieved satisfactory results on the conventional IQA datasets, which indicates that low-level perceptual features play a significant role in quality assessment.

The design philosophies of these FR-IQA milestones contain two important parts: the feature decomposition part and the feature comparison part. The feature decomposition part is extremely important in the development of the FR-IQA task, and many works have concentrated on extracting essential perceptual features that control perceptual quality. However, the feature comparison part has rarely been explored, and no answer has been provided for determining the proper measures in the deep feature comparisons. As such, the Euclidean norm and the similarity index [11] are still widely used for comparing perceptual features, but inferior perceptual quality prediction and perceptual optimization results are obtained, especially when comparing deep network features. In this paper, we attempt to build an FR-IQA measure by using a new kind of measure to attain more perceptual results in both quality assessment and perceptual optimization cases.

2.2 Distribution measures

Distribution measures are widely used in many visual tasks to compare image features in a broad sense. In the reduced-reference IQA (RR-IQA) task, distribution measures are always used to compare histograms of different statistics. Wang *et al.* [30] used the KLD measure to compare the wavelet coefficient histograms of distorted and reference images and obtained satisfactory score prediction results. However, the KLD is not a complete distance measure, and some of its properties might lead to inferior comparison results. KLD cannot achieve higher prediction accuracy due to its asymmetric property; thus, Liu *et al.* [31] used the JSD to compare the pixel histograms of distorted and reference images and achieved better quality assessment results. In some non-reference IQA (NR-IQA) tasks [32], [33], the KLD and JSD have also been viewed as regular measures for comparing different natural image statistics. This is because these distribution measures can be used to determine significant differences rather than focus on a pixelwise difference. In the image retrieval task [34], the WSD is used to compare the differences between the colour, texture, and outline histograms of images, and this approach yielded higher retrieval accuracy than that of the traditional Euclidean norm and chi-square distance [35]. Moreover, in the image generation task, the WSD alleviates the mode collapse problem of the generative adversarial network and could reduce the generation of some artefacts [36]. All of these measures perform well in different visual tasks, even though the features they compare are quite different, indicating that setting distribution measures as image difference measures can be an effective technique. However, the use of these distribution measures in FR-IQA models as fidelity measures has been insufficiently explored. Moreover, ignoring the pixel correlation in deep features, such as the use of MSE or SSIM variants, leads to inferior results in both

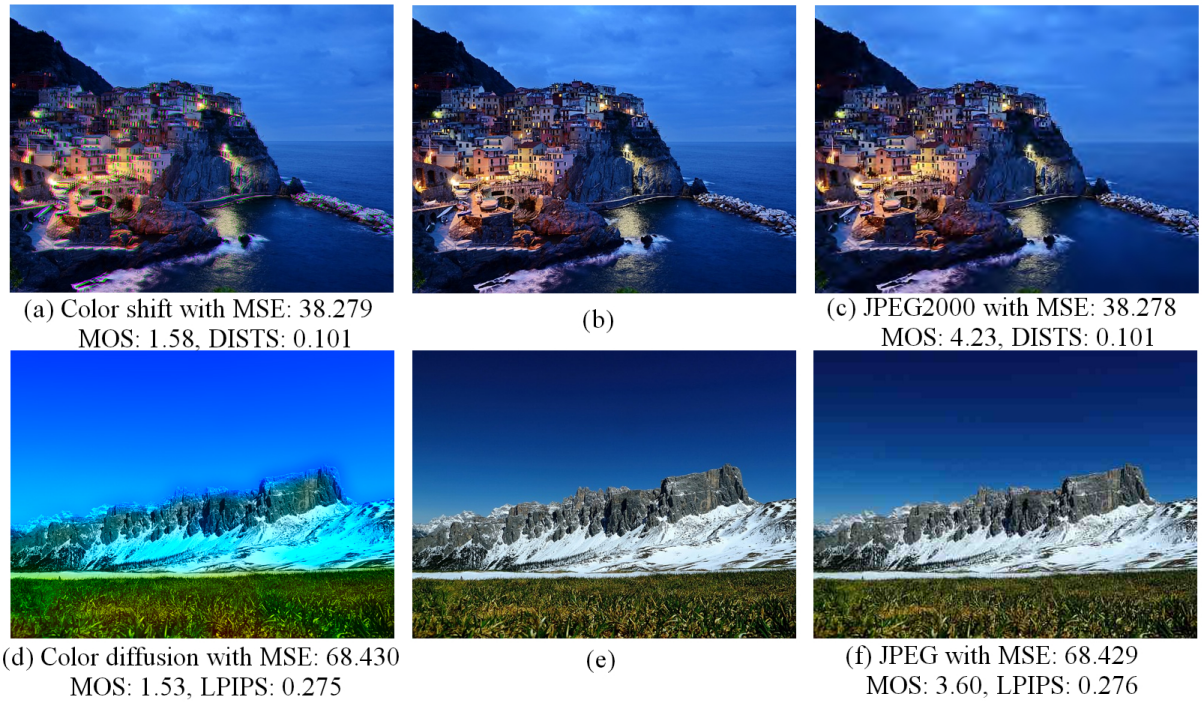

Fig. 2. Distortion samples from the KADID-10k [4] dataset with the same MSE score but different MOS scores. (a) Colour shift distortion. (c) JPEG2000 distortion. (d) Colour diffusion distortion. (f) JPEG distortion. (b) and (e) Reference images for the corresponding distorted images.

quality assessment tasks and perceptual optimization tasks. As such, it is urgent to rethink the essence of deep network features and apply appropriate measures to measure perceptual degradations.

Herein, we show that the contribution of our work is not merely the application of the abovementioned measures to the FR-IQA task. For the deep-based FR-IQA task, we propose that the pursuit of pixel correlation fidelity is more important than the pursuit of pixelwise fidelity, and the utilized distribution measures are suitable for deep network feature comparisons. We also empirically show that the optimization process of the three proposed deep network-based distribution measures is more perceptual than that of other state-of-the-art FR-IQA measures, which is important in both IQA and perceptual optimization tasks.

3 Problem Analysis

3.1 Perceptual information fidelity in the HVS

An ongoing issue regarding the FR-IQA task involves estimating the fidelity between distorted and reference images, but pixelwise fidelity measures such as the MSE fail to serve as perceptual measures. Moreover, explorations and comparisons of deep features can lead to better FR-IQA measures, which compels us to reconsider the type of fidelity pursued by the HVS. According to the efficient coding theory [37], the HVS captures perceptual degradation by sensing the mismatches between the encoded perceptual information and natural scene statistics, such as the mismatch between luminance, contrast, structure, texture and color *etc.*. Such a theory is also the foundation of many NR-IQA measures. Moreover, according to the free-energy principle [38], [39], the HVS is able to evaluate the perceptual quality of the perceived images by using previous visual experience, which indicates that our visual system continues recording the natural scene statistics we have seen and uses them to estimate the perceptual degradation through finding the mismatch [8], [40].

When the visual system receives two stimuli, *i.e.*, a reference image and a distorted image, the comparison between the perceived distorted image and the intrinsic natural scene statistics may no longer be necessary, but the pursuit of perceptual information fidelity should always be the main theme. Such fidelity is very different from pixelwise fidelity because what we sense is not a group of pixels but the perceptual information they contain. Moreover, once perceptual information fidelity is eliminated, even two distortion types with the same strength lead to different perceptual quality degradations, as shown in Fig. 2. Each row of Fig. 2 shows two distortion types with the same strength, *i.e.*, they have the same MSE score, but their MOS scores are quite different because the perceptual information fidelity is ruined to different extents. For (a), colour channel shifting leads to structure ghosting, causing both an unsatisfactory visual experience and structure fidelity disruptions, thus leading to a poor MOS score. For (d), the colour diffusion distortion completely destroys the colour channel fidelity, thus leading to an unacceptable MOS score. On the other hand, the JPEG2000 distortion (c) and JPEG distortion (f) affect the pixelwise fidelity but maintain the perceptual information fidelity to some extent, so their perceptual qualities are acceptable. However, DISTS [6] and learned perceptual image patch similarity (LPIPS) [12] fail to evaluate these image pairs perceptually, indicating that they do not pursue perceptual information fidelity well. In conclusion, the HVS pursues perceptual information fidelity when receiving two

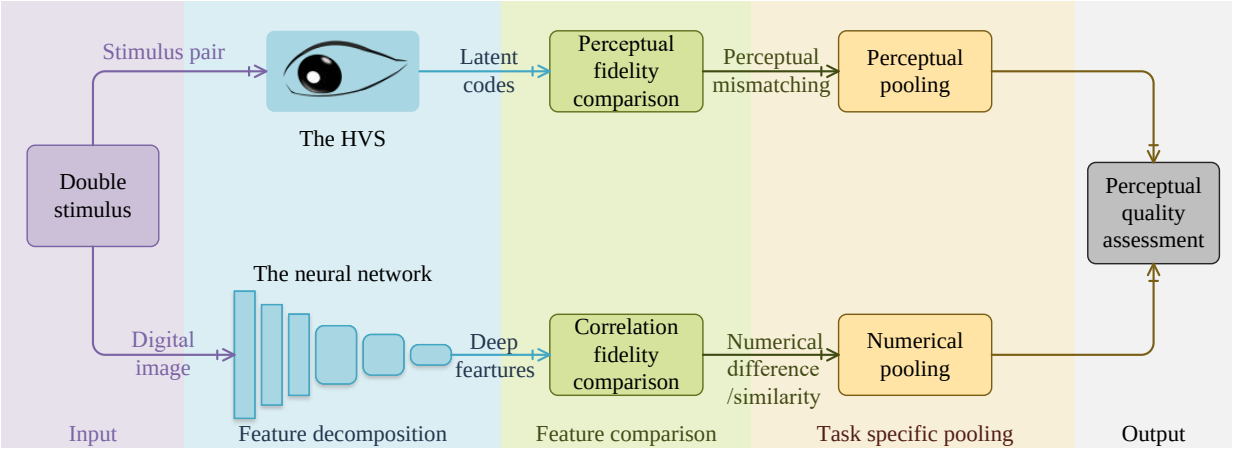

Fig. 3. Overall philosophy of our proposed FR-IQA measures.

stimuli. Such perceptual information is processed by the HVS, and a perceptual FR-IQA measure should consider the procedure.

3.2 Pixel correlation fidelity in deep-based FR-IQA

The quality assessment mechanism in the HVS contains two important parts. First, the reference and distorted image signals are encoded into the underlying perceptual information code. Second, these codes are compared to find mismatches [41]. Such a process is imitated by many FR-IQA measures, which decompose and compare perceptual features. Existing works [6], [12], [42] have shown that it is necessary to build a perceptual FR-IQA measure by comparing the features of deep neural networks because deep features contain perceptual information. Herein, we propose that the pursuit of pixel correlation fidelity is a correct choice for deep feature comparison because pixels in deep features always have strong correlations, and such correlations form the perceptual information we can read, such as structure or texture information. Specifically, for digital images, the correlation between pixels is the basic unit of forming perceptual information, and when distortions occur, pixel correlation degradation in deep features will be affected to different extents, reflecting how severely the perceptual information is lost and finally determining the perceptual quality. In conclusion, the contamination of pixel correlations in deep features is closely related to perceptual information degradation thus applying the metrics that involve comparing pixel correlations is necessary.

4 Methodology

The overall philosophy of our FR-IQA design is shown in Fig. 3, where we divide the whole process of the deep-based FR-IQA measure into 3 parts: a feature decomposition part, a feature comparison part, and a task-specific pooling part. We emphasize that our main work is the feature comparison part. Specifically, we focus on approaching the mechanism of perceptual information fidelity comparison in the HVS by employing a deep feature correlation comparison and introducing distribution measures. In our approach, we regard deep features as a type of distribution and then use distribution measures to estimate pixel correlation fidelity.

4.1 Preliminaries of distribution measures

Distribution measures aim to quantify the distance between two probability distributions in the sense of statistics, and we mainly test the WSD [18], JSD [19], and SKLD [20] in terms of deep feature comparisons, which lead to three independent FR-IQA measures. We first present the definitions of the KLD and WSD as follows:

$$KLD(\mathcal{X},\mathcal{Y}) = \int F_{\mathcal{X}}(x)(\log F_{\mathcal{X}}(x) - \log F_{\mathcal{Y}}(x))\,dx, \quad (1)$$

$$WSD_l(\mathcal{X},\mathcal{Y}) = \left(\inf_{J\in\mathcal{J}(\mathcal{X},\mathcal{Y})} \int ||x-y||_l \, dJ(x,y)\right)^{1/l}. \quad (2)$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the statistical distributions and x and y are their masses, respectively. $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$ are the probability density functions of $\mathcal{X}$ and $\mathcal{Y}$, respectively. $\mathcal{J}(\mathcal{X},\mathcal{Y})$ is the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$, and l is the exponential index. In our work, we set $l = 2$.

The definitions of the JSD and SKLD are based on the KLD and are

$$JSD(\mathcal{X},\mathcal{Y}) = \frac{1}{2}KLD(\mathcal{X},\mathcal{M}) + \frac{1}{2}KLD(\mathcal{Y},\mathcal{M}) \quad (3)$$

$$SKLD(\mathcal{X},\mathcal{Y}) = \frac{1}{2}KLD(\mathcal{X},\mathcal{Y}) + \frac{1}{2}KLD(\mathcal{Y},\mathcal{X}) \quad (4)$$

where $\mathcal{M} = \frac{1}{2}(\mathcal{X}+\mathcal{Y})$. The relationships among the WSD, JSD, and SKLD involve two key points. First, only the WSD is a complete distance, while the SKLD and JSD are merely divergences and do not satisfy the triangle inequality of the complete distance requirement. Second, the SKLD and JSD are two symmetric forms of the KLD that aim to erase the ambiguity encountered when applying them as FR-IQA measures and perceptual losses.

When $\mathcal{X}$ and $\mathcal{Y}$ are one-dimensional discrete statistical distributions and $l = 2$, the computation of the WSD can be reduced to determining the difference between the empirical statistical distributions of $\mathcal{X}$ and $\mathcal{Y}$. Specifically, Eq. 2 can be computed by

$$WSD_1(\mathcal{X},\mathcal{Y}) = \int_{\mathbb{R}} (\hat{F}_{\mathcal{X}}(t) - \hat{F}_{\mathcal{Y}}(t))^2 dt \quad (5)$$

where $\hat{F}_{\mathcal{X}}$ and $\hat{F}_{\mathcal{Y}}$ are the empirical cumulative density functions of $\mathcal{X}$ and $\mathcal{Y}$, respectively. t is the parameter used for integration.

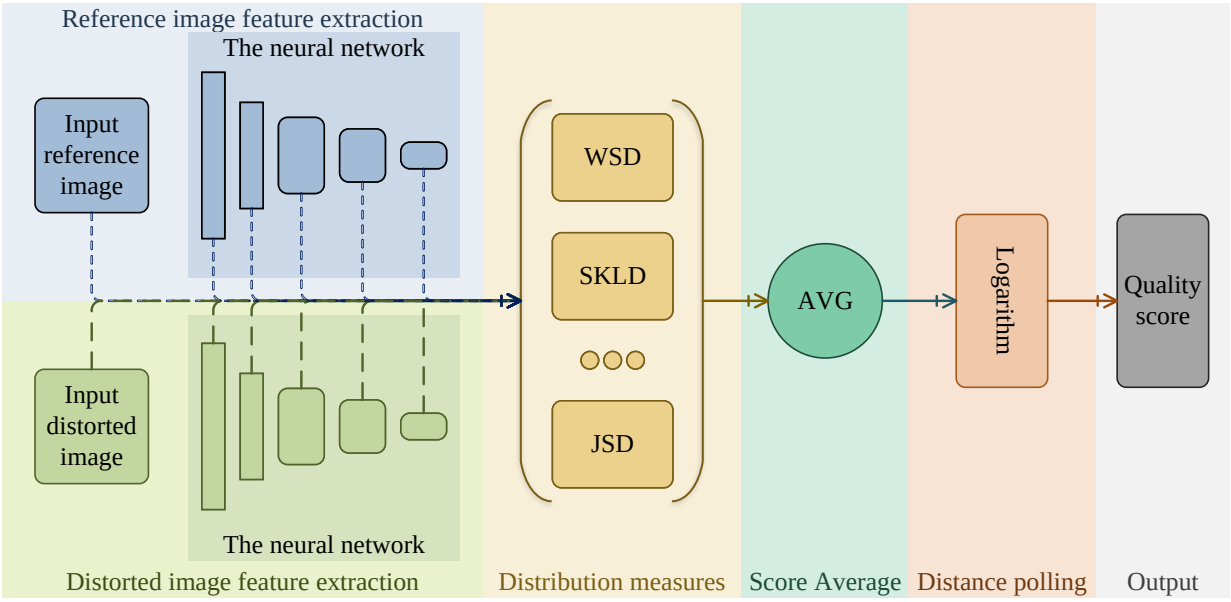

Fig. 4. Structures of the three proposed deep-based distribution measures. We compare the differences between 5-stage VGG19 [5] features and raw images to estimate the pixel correlation fidelity loss. Then, 6 difference scores are averaged and sent to the logarithm function to output the final quality score.

The comparison philosophy of the three distribution measures is entirely different from that of the pixelwise measures, *i.e.*, the MSE and the l_p norm. A trivial example involves comparing the distances between two one-dimensional Gaussian distributions, *i.e.*, $\mathcal{X} \sim \mathcal{N}_x(\mu_x, \sigma_x^2)$ and $\mathcal{Y} \sim \mathcal{N}_y(\mu_y, \sigma_y^2)$, where μ_x, μ_y and σ_x^2, σ_y^2 are the means and variances of two Gaussian distributions, respectively. Specifically, the following theorem shows that the three distribution measures are subject to comparing the similarities between the means and variances, which in fact involves comparing the central locations and the dispersion degrees of the Gaussian distributions. The proof of the theorem is presented in the appendix.

Theorem 4.1. *When $\mathcal{X} \sim \mathcal{N}_x(\mu_x, \sigma_x^2)$ and $\mathcal{Y} \sim \mathcal{N}_y(\mu_y, \sigma_y^2)$, then the WSD, JSD and SKLD have a unified upper bound for comparing $\mathcal{X}$ and $\mathcal{Y}$:*

$$D(\mathcal{X},\mathcal{Y})^2 \leqslant C_1(1-\frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2})+C_2(1-\frac{2\sigma_{xy}}{\sigma_x^2+\sigma_y^2})+C_3, \quad (6)$$

where D can be set as the WSD, JSD, or SKLD. C_1, C_2, and C_3 are constants, and σ_{xy} is the covariance of the distributions $\mathcal{X}$ and $\mathcal{Y}$.

4.2 Deep feature space distribution measures

Given a distorted image P and a reference image Q, we first use the VGG19 network [5] to project them into deep features $\{\tilde{P}_i\}_{i=1,...5}$ and $\{\tilde{Q}_i\}_{i=1,...5}$, where $i = 1,...5$ represents the 5 stages of the VGG19 network [5]. Then, we use the WSD, JSD, and SKLD to compare these deep features and raw images. Specifically, the pixel strength at each position is treated as the probability density; then, the overall shape of the pixel strength in the deep features and raw images forms a kind of distribution. In other words, if the pixel strength at a particular location is represented as (m, n, s), where (m, n) indicates the location and s is the pixel strength, then the variation of s across different locations (m, n) forms the pixel strength distribution. Then, distribution measures can be used to evaluate the variation in the pixel strength distribution, *i.e.*, to compare how the pixel strength is concentrated or dispersed at different locations which is the key to generating perceptual information. The complete procedures for DeepWSD, DeepSKLD, and DeepJSD can be formulated as follows:

$$\mathcal{D}_W(P,Q) = WSD_2(P,Q) + \sum_{i=1}^{5} WSD_2(\tilde{P}_i, \tilde{Q}_i) \quad (7)$$

$$\mathcal{D}_J(P,Q) = JSD(P,Q) + \sum_{i=1}^{5} JSD(\tilde{P}_i, \tilde{Q}_i) \quad (8)$$

$$\mathcal{D}_S(P,Q) = SKLD(P,Q) + \sum_{i=1}^{5} SKLD(\tilde{P}_i, \tilde{Q}_i) \quad (9)$$

Inspired by the most apparent distortion (MAD) FR-IQA measure [43], beyond comparing the pixel correlation fidelity levels, we also introduce a weighted Euclidean norm for each distribution measure:

$$\mathcal{D}_{eul}(P,Q) = g(\mathcal{D}(P,Q)) \times ||P-Q||_2 \quad (10)$$

$$+ \sum_{i=1}^{5} (g(\mathcal{D}(\tilde{P}_i, \tilde{Q}_i)) \times ||\tilde{P}_i - \tilde{Q}_i||_2), \quad (11)$$

where $\mathcal{D}(P,Q)$ is set as the corresponding distribution measure, and the adaptive weight $g(s)$ is set as

$$g(s) = \frac{1}{(s+10)^2\sqrt{\exp(-1/(s+10))}}. \quad (12)$$

Two factors support the introduction of the weighted Euclidean norm. First, the HVS is able to evaluate the perceptual quality at different scales; in other words, we can compare images not only in a global perceptual information

Fig. 5. Distortion samples from the KADID-10k [4] dataset. (a) Mean shift distortion. (c) Colour block distortion. (d) White noise distortion. (f) Pixelwise distortion. (b) and (e) Reference images for the corresponding distorted images. Note that DISTS is trained with the KADID-10K [4] dataset.

comparison manner but also in a local pixelwise comparison manner. Such a mechanism is imitated by our FR-IQA measures. Moreover, regarding the MAD [43], Larson *et al.* reported a notable behaviour exhibited by testers in their subjective evaluation experiment; that is, testers first sensed the distortion on a global scale and directly gave low scores once the perceptual degradation was apparent. If the distortion was imperceptible, they zoomed in to find the distortion in a patch-by-patch or even a pixel-by-pixel manner and then gave a quality score. The weighted Euclidean norm mimics such behaviour, and when the values of distribution measures are large, the value of the weighted Euclidean norm is suppressed by the weight $g(s)$. Second, the weighted Euclidean norm also benefits the proposed measures in terms of perceptual optimization by prohibiting the generation of artefacts and accelerating the convergence, which makes FR-IQA more complete as a perceptual loss. After comparing the features of each network stage, we compute the average score of each stage and send it to a training-free logarithm function to output the final perceptual quality score. The whole process is formulated as follows, and we also present the structures of DeepWSD, DeepSKLD, and DeepJSD in Fig. 4.

$$DeepWSD(P,Q) = (\log(\frac{1}{6}\mathcal{D}_W(P,Q) + \frac{1}{6}\mathcal{D}_{eul}(P,Q)))^{\frac{1}{4}} \quad (13)$$

$$DeepSKLD(P,Q) = (\log(\frac{1}{6}\mathcal{D}_S(P,Q) + \frac{1}{6}\mathcal{D}_{eul}(P,Q)))^{\frac{1}{4}} \quad (14)$$

$$DeepJSD(P,Q) = (\log(\frac{1}{6}\mathcal{D}_J(P,Q) + \frac{1}{6}\mathcal{D}_{eul}(P,Q)))^{\frac{1}{4}} \quad (15)$$

One notable superiority exhibited by the three proposed deep-based distribution measures is that they do not require training and do not contain many empirical hyperparameters, but they can achieve satisfactory results on several conventional IQA datasets, which indicates that the pursuit of pixel correlation fidelity for deep-based FR-IQA measures is an effective way to imitate the evaluation mechanism in the HVS. Additionally, the proposed FR-IQA measures even produce competitive quality assessment results on some of the latest challenging IQA datasets, which also reveals the generality of our theory. In addition, the three proposed measures can also be used for perceptual optimization in many image reconstruction tasks, providing broad application prospects.

4.3 Connection with existing methods

The structures of the three deep network-based distribution measures are similar to those of many existing models. Herein, we explain their similarities and differences to clarify our contributions.

4.3.1 Connection with other deep network-based FR-IQA methods

Existing deep network-based FR-IQA measures tend to apply a deep convolutional neural network to extract essential perceptual features and train an extra score predictor to regress MOS scores on several datasets. Due to the perceptual relevance of deep network features, our proposed FR-IQA measures are also built on deep network features. However, unlike the previous methods, we mainly work on determining how to approach the comparison mechanism in the HVS. Moreover, existing deep-based FR-IQA measures always need to train an extra score predictor, while our proposed measures only use a logarithmic function to output perceptual quality scores; this also demonstrates that comparing deep features through distribution measures is closely related to subjective evaluations.

TABLE 1
Detailed information about the 7 test IQA datasets. For the PIPAL dataset, we use the training set because the labels of the test set are unavailable

Name	No. of Ref	No. of Dist	No. and type of Dist	Image types	Human judgements	Evaluation criterion
TID2013	25	3000	24 (traditional)	Full image	524k	MOS (Swiss system)
LIVE	29	800	5 (traditional)	Full image	25k	DMOS (5 scores)
CSIQ	30	800	6 (traditional)	Full image	5k	MOS (Direct ranking)
IVC	10	235	4 (traditional)	Full image	5k	MOS (5 scores)
KADID-10k	81	10, 125	25 (Synthesis)	Full image	300k	MOS (5 scores)
Live-MultiDist	15	450	2 (Mixed)	Full image	25k	DMOS (5 scores)
PIPAL(train)	200	23200	40 (Algorithm)	image patches	1.13m	MOS (Elo rating system)

4.3.2 Connection with feature similarity measures

From a statistical viewpoint, SSIM [11] can be treated as a distribution similarity measure that measures the similarities among the first-order moments, the second-order moments, and the overall shapes of pixel strength distribution. Moreover, through Eq. (6), with a Gaussian hypothesis, the upper bound of three more general distribution measures can be reduced to a unified form that is very similar to the SSIM. The key reason for this is that the Gaussian distribution can be completely dominated by the first two moments, which are the mean and the variance. In the pixel domain, the means and variances of image patches have clear perceptual meanings, which are their luminance and contrast levels. However, deep features are usually not Gaussian, and their means and variances do not have specific perceptual meanings; thus, directly applying the SSIM or its variants to compare deep features may lead to inferior quality assessment results, as shown in Fig. 5. We elaborately pick two distortion sample pairs that have different MOS scores from the Konstanz Artificially Distorted Image Quality Database 10k (KADID-10k) [4] dataset. Note that DISTS is trained on this dataset, but it returns the same quality score for two distortion sample pairs and fails to evaluate them perceptually. On the other hand, even without training, the three proposed measures are able to reflect the perceptual difference between the distortion sample pairs and return quite different objective scores, which indicates that distribution measures are more suitable for deep feature comparison tasks.

4.3.3 Connection with the perceptual loss

Similar to the widely used perceptual loss [51], [52], our proposed deep-based distribution measures also utilize and compare the features of the VGG19 [5] network. However, our FR-IQA measures are based on distribution measures and consider all 5 stage features and raw images to perform the quality assessment and perceptual optimization tasks. The existing perceptual loss function [51], [52] only compares part of the observed features through the Euclidean norm. Works completed by Delbracio *et al.* [53] and Cao *et al.* [13], which are very similar to ours, compare the VGG16 features through the projected WSD [54] to alleviate the computational burden. In contrast, we do not project deep features because we believe that projection introduces information loss. Specifically, we directly reshape the deep features to one-dimensional forms and treat them as one-dimensional distributions. Then, a sliding window is used to compare these 1D distributions in a patch-by-patch manner through three different distribution measures.

5 Experimental results

5.1 Databases and experimental settings

To demonstrate the effectiveness of the three proposed FR-IQA measures, we test them on 7 IQA datasets: the Tampere Image Database (TID2013) [44], Laboratory for Image & Video Engineering (LIVE) [45], Categorical Image Quality (CSIQ) [43], Image Quality Database (IVC) [46], KADID-10k [4], LIVE Multiply Distorted (LIVE-MultiDist) [50] and Perceptual Image Processing Algorithms (PIPAL) [49] datasets. The details of these FR-IQA datasets are shown in Table 1. Among these FR-IQA datasets, the PIPAL dataset is a recent challenging IQA dataset from the New Trends in Image Restoration and Enhancement (NTIRE 2021) challenges on perceptual IQA and contains new distortion types generated by image reconstruction algorithms. The TID2013 [44], LIVE [45], CSIQ [43] and IVC [46] datasets are conventional IQA datasets that contain traditional distortion types such as Gaussian white noise distortions. The KADID-10k [4] dataset contains various synthetic distortion types that are different from those of the four conventional IQA datasets, and LIVE-MultiDist [50] contains mixed distortions such as JPEG distortion plus blur, *among others*.

We also compare the performance of the three proposed methods with several FR-IQA milestones, which are the peak-signal-to-noise ratio (PSNR), SSIM [11], multiscale SSIM (MS-SSIM) [47], GMSD [29], VIF [27], MAD [43], FSIM [26], complex wavelet structural similarity (CW-SSIM) [48], LPIPS [12], the perceptual image error assessment through the pairwise preference (PieAPP) FR-IQA measure [42], DISTS [6] and space warping difference IQA network (SWD) [49]. Among them, the PSNR, SSIM [11], MS-SSIM [47] and GMSD [29] are FR-IQA measures that operate in the pixel domain. The VIF [27], MAD [43], FSIM [26] and CW-SSIM [48] contain frequency feature comparisons, while the PieAPP [42], LPIPS [12], DISTS [6] and SWD [49] are FR-IQA measures that work in the deep feature domain. All the compared FR-IQA measures use settings that are in line with their open source forms. For LPIPS [12] and SWD [49], we take their VGG16 forms because such forms can lead to better quality assessment results. For DeepWSD, DeepJSD, and DeepSKLD, the only flexible parameter is the patch size for the deep feature comparison, and we set the

TABLE 2
Comparison among several FR-IQA methods on four classic datasets. The bold numbers are the three highest scores in each column.

Method	TID2013 [44]			LIVE [45]			CSIQ [43]			IVC [46]		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	0.6773	0.6876	0.4963	0.8648	0.8726	0.6773	0.8195	0.8101	0.6014	0.6804	0.6499	0.4781
SSIM [11]	0.7767	0.7271	0.5454	0.9341	0.9479	0.7963	0.8523	0.8656	0.6807	0.7708	0.9018	0.7223
MS-SSIM [47]	0.8292	0.7887	0.6049	0.9399	0.9513	0.8048	0.8895	0.9133	0.7393	0.7913	0.8980	0.7203
GMSD [29]	0.8548	0.8043	0.6324	0.9574	0.9608	0.8271	0.9452	0.9503	0.8043	0.8645	0.8548	0.6624
VIF [27]	0.8672	0.8429	0.6526	0.9343	0.9603	0.8284	0.9132	0.9121	0.7432	0.7391	0.7273	0.5590
MAD [43]	0.8267	0.7680	0.6154	0.9682	**0.9669**	0.8425	0.9505	0.9468	0.7975	0.8704	0.8698	0.6671
FSIM [26]	0.8233	0.8549	0.6549	0.9608	**0.9672**	**0.8814**	0.9187	0.9379	0.7683	0.8161	0.9263	0.7537
CW-SSIM [48]	0.6295	0.7560	0.5580	0.8402	0.9082	0.7142	0.7687	0.7652	0.5683	0.5936	0.5834	0.4072
PieAPP [42]	0.8501	0.8479	0.6828	0.9079	0.9279	0.8266	0.9300	0.9369	0.7721	0.9341	0.9306	0.7698
LPIPS [12]	0.7324	0.6696	0.4970	0.9343	0.9324	0.7782	0.8936	0.8758	0.6893	0.8715	0.9044	0.7386
DISTS [6]	0.8624	0.8483	0.6574	0.9560	0.9542	0.8112	0.9284	0.9289	0.7675	0.8993	0.9138	0.7267
SWD [49]	0.8377	0.8153	0.6260	0.8731	0.8832	0.6894	0.9222	0.9154	0.7357	0.8826	0.9012	0.7132
DeepWSD	**0.9001**	**0.8806**	**0.7003**	**0.9720**	0.9654	**0.8447**	**0.9629**	**0.9646**	**0.8279**	**0.9420**	**0.9356**	**0.7767**
DeepJSD	**0.9000**	**0.8790**	**0.6973**	**0.9717**	0.9653	0.8445	**0.9630**	**0.9670**	**0.8343**	**0.9424**	**0.9362**	**0.7776**
DeepSKLD	**0.9012**	**0.8783**	**0.6962**	**0.9720**	**0.9659**	**0.8463**	**0.9652**	**0.9683**	**0.8375**	**0.9425**	**0.9358**	**0.7772**

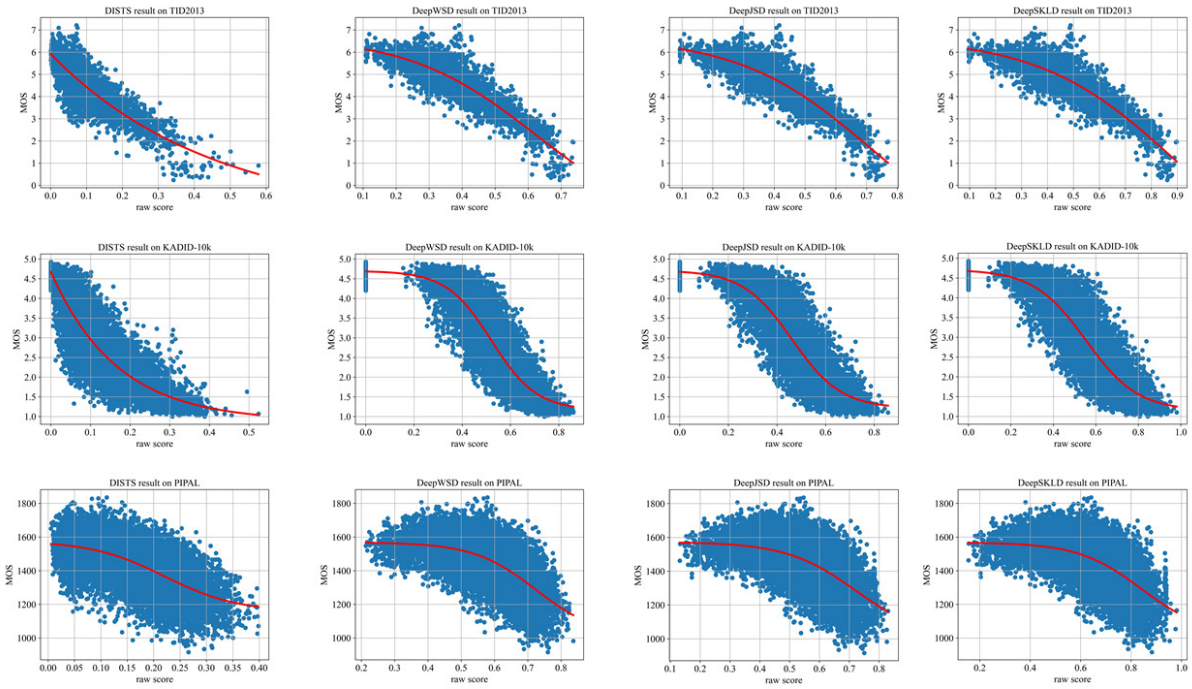

Fig. 6. Scatter plots of DISTS, DeepWSD, DeepJSD, and DeepSKLD based on the TID2013, KADID-10k, and PIPAL datasets. The x-axis denotes the raw score of each tested IQA measure, and the y-axis represents the MOS label for each dataset.

patch size to 8×8. The effect of the patch size is thoroughly researched in an ablation study.

To quantify the score prediction results of these FR-IQA measures, we use Pearson's linear correlation coefficient (PLCC), the Spearman rank-order correlation coefficient (SRCC), and the Kendall rank-order correlation coefficient (KRCC). To better map the scores of the FR-IQA measures to the human evaluation scores, we use a 4-parameter regression model before computing the PLCC, whose definition is

$$\bar{D} = (\beta_1 - \beta_2)/(1 + \exp(-(D - \beta_3)/|\beta_4|) + \beta_2), \quad (16)$$

where D is the raw score and $\bar{D}$ is the regression score. $\{\beta_i\}_{i=1,\ldots 4}$ represents the 4 parameters.

Beyond the quality assessment experiment, we also perform the maximum differential competition (MDC) experiment [21], which provides another approach for evaluating the performance of two given IQA measures Q_1 and Q_2. The basic philosophy is to search for a distorted image pair $\{I_1, I_2\}$ according to a reference image I, where the perceptual qualities of I_1 and I_2 are quite different. If such a pair can fool Q_1 but not Q_2, then we say that Q_2 is more perceptual than Q_1 on this image pair. If we can find this kind of image pair as much as possible but cannot demon-

TABLE 3
Comparison among several FR-IQA methods on three classic datasets. The bold numbers are the three highest scores in each column. Note that DISTS is trained with the KADID-10k dataset, so we use * to highlight its score. For PIPAL, we use PIPAL (train) to denote that we only use the training set.

Method	KADID-10k [4]			LIVE-MultiDist [50]			PIPAL (train) [49]		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	0.6747	0.6665	0.4836	0.7622	0.7664	0.5830	0.4374	0.4065	0.2762
SSIM [11]	0.7217	0.7325	0.5572	0.8652	0.8255	0.6157	0.5231	0.4987	0.3452
MS-SSIM [47]	0.8012	0.8029	0.6088	0.8826	0.8795	0.6431	0.5796	0.5510	0.3884
GMSD [29]	0.8315	0.8308	0.6357	0.9068	0.8923	0.7144	0.5924	0.5703	0.4029
VIF [27]	0.7927	0.7906	0.6014	0.8947	0.8874	0.5428	0.5701	0.5591	0.3960
MAD [43]	0.8232	0.7965	0.6210	0.8542	0.8712	0.6943	**0.6508**	0.6161	**0.4397**
FSIM [26]	0.8511	0.8542	0.6648	0.8386	0.8932	0.6961	0.6321	0.6065	0.4312
CW-SSIM [48]	0.4977	0.4995	0.3521	0.4928	0.4160	0.2930	0.3543	0.2598	0.1751
PieAPP [42]	0.8656	0.8647	0.6736	0.8763	0.8881	0.6897	**0.6972**	**0.6982**	**0.5090**
LPIPS [12]	0.7003	0.7200	0.5313	0.8350	0.8078	0.6083	0.6109	0.5730	0.4041
DISTS [6]	0.8887*	0.8893*	0.7100*	0.8951	**0.9330**	**0.7478**	0.6354	**0.6172**	0.4382
SWD [49]	0.7894	0.7903	0.5883	0.8566	0.8534	0.6516	0.6219	0.6031	0.4264
DeepWSD	**0.9008**	**0.9016**	**0.7254**	**0.9190**	**0.9191**	**0.7483**	**0.6434**	**0.6256**	**0.4431**
DeepJSD	**0.8929**	**0.8935**	**0.7143**	**0.9111**	0.8988	0.7189	0.6124	0.5983	0.4204
DeepSKLD	**0.8957**	**0.8963**	**0.7183**	**0.9137**	**0.9011**	**0.7231**	0.6211	0.6068	0.4274

strate conversely, then Q_2 dominates Q_1 in the perceptual sense.

Specifically, given an error bound ε, the required image pair $\{I_1, I_2\}$ with quite different subjective perceptual quality scores must satisfy

$$\begin{cases} |Q_1(I_1, I) - Q_1(I_2, I)| < \varepsilon, \\ |Q_2(I_1, I) - Q_2(I_2, I)| > \varepsilon. \end{cases} \tag{17}$$

Then, Q_2 is more perceptual than Q_1 on $\{I_1, I_2\}$. A similar definition for Q_1 dominating Q_2 can be derived in a parallel manner, and in our experiment, we set $\epsilon = 0.001$ and pick image pairs $\{I_1, I_2\}$ with MOS score differences that are larger than 2.

The experimental results are structured as follows. In **Section 5.2**, we show the quality assessment results. In **Section 5.3**, we present the MDC experiment results. In **Section 5.4**, we analyse the sensitivity of parameters and show the ablation study results.

5.2 Quality prediction results

The quality prediction results on TID2013 [44], LIVE [45], CSIQ [43] and IVC [46] are shown in Table 2. These IQA datasets have been used for many years, which may cause some FR-IQA measures to unintentionally overadapt. Nonetheless, the three proposed FR-IQA measures attain state-of-the-art performance, which reveals their advanced score prediction abilities. Additionally, we also test the performance of our measures on other datasets and present the results in Table 3. Note that DISTS [6] is trained with the KADID-10k [4] dataset, but the three proposed methods still attain better results than DISTS. Such results strongly support the effectiveness of comparing the pixel correlation difference in deep features. Moreover, on the challenging PIPAL [49] dataset, the three proposed FR-IQA measures also attain satisfactory results, and DeepWSD obtains one of the three highest scores, which indicates that the pursuit of pixel correlation fidelity is also effective for the distortion types generated by some reconstruction algorithms. In addition, we also present scatter plots of the three proposed FR-IQA measures based on the TID2013 [44], KADID-10k [4] and PIPAL [49] datasets to visualize their score prediction results. In Fig. 6, we find that the raw scores of the three proposed measures are nearly linear with the subjective evaluations in the TID2013 dataset.

The success of the three proposed measures can be mainly attributed to the pixel correlation comparison. Beyond that, the three methods perform differently on each dataset. DeepWSD performs better than DeepJSD and DeepSKLD on large-scale datasets such as the TID2013 [44], KADID-10k [4] and PIPAL [49] datasets, indicating that it is more generally applicable to various kinds of distortions. On the other hand, DeepJSD and DeepSKLD are better than DeepWSD on the LIVE [45], CSIQ [43] and IVC [46] datasets. These datasets contain types of distortions related to signal compression and transmission, such as JPEG2000 distortion and Gaussian blur. We empirically conclude that these two measures are more specialized for the quality evaluation of distortions related to compression. The latent reason for this phenomenon may originate from the definitions of the three proposed measures. Specifically, the logarithmic terms in the SKLD and JSD may unintentionally measure information loss that is more important in image compression and transmission, making these measures more specialized for compression distortion. The WSD considers the overall geometry of the pixel strength distribution; thus, it is more generally applicable to various kinds of distortions, such as the distortions generated by image reconstruction algorithms.

5.3 Maximum Differential Competition results

We compare the FSIM [26], MAD [43], PieAPP [42], DISTS [6] and the three proposed FR-IQA based on the

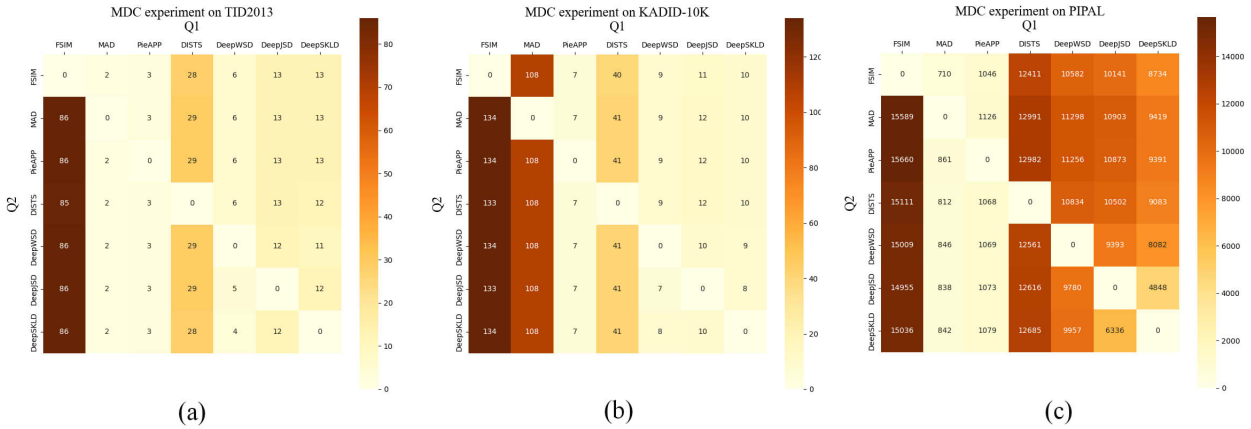

Fig. 7. MDC results obtained on 3 large IQA datasets. (a), (b), and (c) Results obtained on the TID2013 [44], KADID-10k [4] and PIPAL [49] datasets, respectively.

Fig. 8. Toy optimization results of Eq. (18).

TABLE 4
Ablation study results with respect to the patch size and $g(s)$. The bold numbers are the three highest scores in each column. '*w*/P4' means that a window size of 4×4 is used. '*w/o* $g(s)$' means that the adaptive weight $g(s)$ is not used. Bold font signifies the best results obtained by each block on the corresponding dataset. For PIPAL, we use PIPAL (train) to denote that we only use the training set.

Method	TID2013 [44]			KADID-10k [4]			PIPAL (train) [49]		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
DeepWSD w/P4	**0.9024**	**0.8817**	**0.7014**	0.8962	0.8970	0.7186	0.6103	0.6013	0.4229
DeepWSD w/P8	0.9001	0.8806	0.7003	**0.9008**	**0.9016**	**0.7254**	0.6434	0.6256	0.4431
DeepWSD w/P16	0.8872	0.8692	0.6849	0.8962	0.8971	0.7190	**0.6620**	**0.6437**	**0.4588**
DeepWSD w/o $g(s)$	0.8994	0.8801	0.6992	0.8952	0.8947	0.7162	0.6125	0.5974	0.4196
DeepJSD w/P4	0.8990	0.8776	0.6951	**0.8933**	**0.8939**	**0.7148**	0.6141	0.5991	0.4212
DeepJSD w/P8	0.9000	0.8790	0.6973	0.8929	0.8935	0.7143	0.6124	0.5983	0.4204
DeepJSD w/P16	**0.9005**	**0.8797**	**0.6985**	0.8928	0.8935	0.7141	**0.6176**	**0.6073**	**0.4278**
DeepJSD w/o $g(s)$	0.8927	0.8710	0.6863	0.8898	0.8901	0.7103	0.6120	0.5969	0.4192
DeepSKLD w/P4	0.8955	0.8777	0.6950	0.8952	0.8957	0.7179	0.6195	0.6058	0.4268
DeepSKLD w/P8	**0.9012**	**0.8783**	**0.6962**	**0.8957**	**0.8963**	**0.7183**	0.6211	0.6068	0.4274
DeepSKLD w/P16	0.8953	0.8770	0.6947	0.8928	0.8937	0.7147	**0.6218**	**0.6150**	**0.4337**
DeepSKLD w/o $g(s)$	0.8831	0.8678	0.6903	0.8946	0.8950	0.7160	0.6123	0.5973	0.4195

TID2013 [44], KADID-10k [4] and PIPAL [49] datasets by searching for the image pairs. We visualize the MDC experimental results in heatmaps in Fig. 7. In the heatmaps, each element indicates the number of image pairs on which Q_1 is fooled while Q_2 is not. To compare the evaluation abilities of the two models, we can compare the values on the two

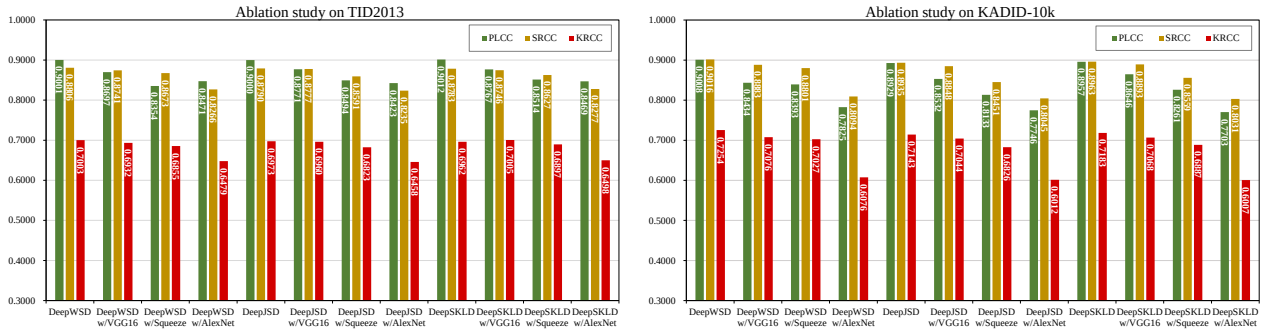

Fig. 9. The influences of network structures on the TID2013 and KADID-10k datasets.

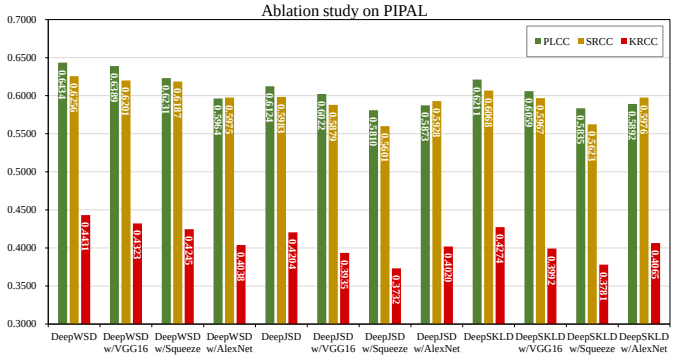

Fig. 10. The influence of the network structure on the PIPAL dataset.

sides of the heatmap diagonal at the corresponding location. We find that some FR-IQA measures are not as good as we expected, even though they have quite high PLCC, SRCC, and KRCC scores. For example, the FSIM [26] has high SRCCs for the TID2013 [44] and KADID-10k [4] datasets, but many image pairs are not properly evaluated. In contrast, PieAPP [42] does not perform well in terms of the PLCC, SRCC, and KRCC scores obtained on these datasets, but we can hardly find image pairs to fool it, and the performance of PieAPP is even better than the performance of DeepWSD, DeepJSD, and DeepSKLD.

The MDC experiment [21] reflects whether the quality isoline of the proposed FR-IQA measures is perceptual, which is very important for perceptual optimization. In many image reconstruction tasks, we always reconstruct images with poor perceptual quality into those with better perceptual quality; thus, a perceptual quality isoline is one of the determining factors that control the visual satisfaction of the final results. The MDC results show that MAD [43], PieAPP [42] and the three proposed deep distribution measures have more perceptual quality isolines. However, when we apply them in a toy optimization example, we find that the optimization process of MAD [43] suffers from nonconvex issues and that PieAPP [42] is not optimizable, as shown in Fig. 8. The toy example is defined as follows:

$$\hat{x} = \arg\min_{x} D(x, y), \tag{18}$$

where y is a given reference image and x is a distorted image. The FR-IQA measure $D(x, y)$ should reconstruct x to y with high perceptual quality, whose result is denoted as $\hat{x}$. We set two initial images for x; one is a white image, and the other is a blurred image. With different initial images, MAD [43] cannot reconstruct the same perceptual results, indicating that it induces a nonconvex loss. On the other hand, PieAPP [42] is not optimizable and cannot reconstruct the final results. In contrast, with different initial images, DeepWSD, DeepJSD, and DeepSKLD all nicely reconstruct highly perceptual quality results, which indicates that they can serve as optimizable perceptual losses with perceptual quality isolines. We argue that this is because WSD [18], JSD [19] and SKLD [20] are all well-defined convex measures so they can naturally address the optimization task.

5.4 Ablation study and parameter sensitivity analysis

Influence of the patch size. We present the influence of the patch size on three IQA datasets, the TID2013 [44], KADID-10K [4] and PIPAL [49] datasets, in Table 4. Different patch sizes have different influences on the three measures for the TID2013 [44] and KADID-10k [4] datasets. Specifically, DeepWSD can achieve better accuracy for the TID2013 [44] dataset with a small patch size, while for the KADID-10k [4] dataset, a larger patch size leads to better accuracy. Such phenomena also exist in DeepJSD, but the influence of the patch size is the opposite of that experienced by DeepWSD. In contrast, all three measures can achieve better results on the PIPAL [49] datasets with larger patch sizes. We argue that this is because the distortion types of the three datasets are quite different. Specifically, the TID2013 [44] and KADID-10k [4] datasets contain distortion types such as Gaussian white noise, which can be better detected with a small receptive field. However, PIPAL contains distortions generated by many image reconstruction or image generation algorithms, which must be detected with a larger receptive field. Considering the overall performance of the three measures, we finally set the default patch size to 8×8, which is a compromise.

Influence of the adaptive weight $g(s)$**.** The adaptive weight is necessary for the quality assessment task. To demonstrate this, we compare the performances of DeepWSD, DeepJSD, and DeepSKLD without $g(s)$ on three IQA datasets and present the results in Table 4. Without $g(s)$, the performances of the three IQA measures deteriorate, which further demonstrates that the pursuit of pixel correlation fidelity is more important when comparing deep features.

Influence of the network structure. We test other deep networks to demonstrate the generality of the proposed

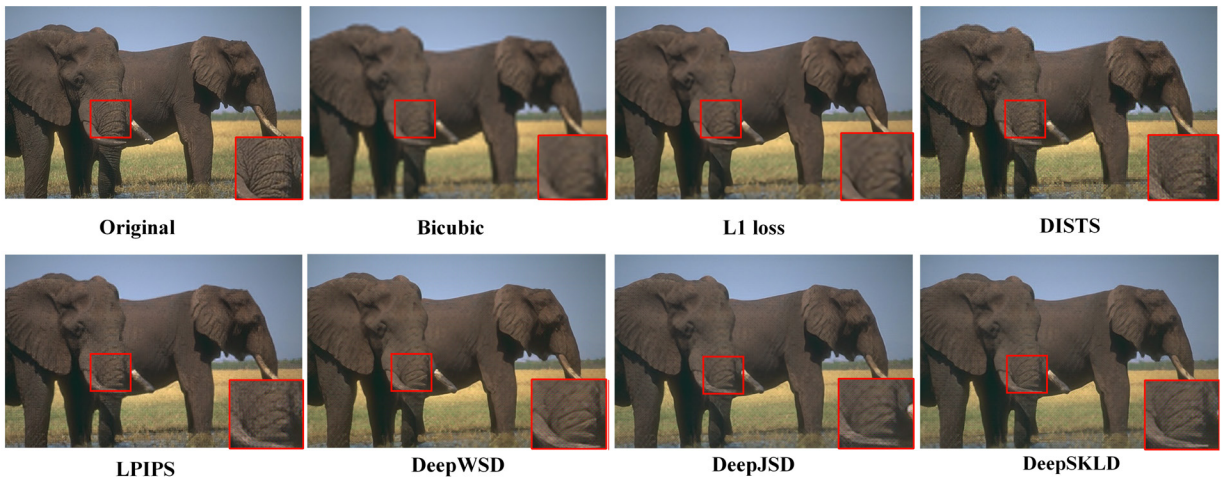

Fig. 11. Superresolution results obtained for a natural image.

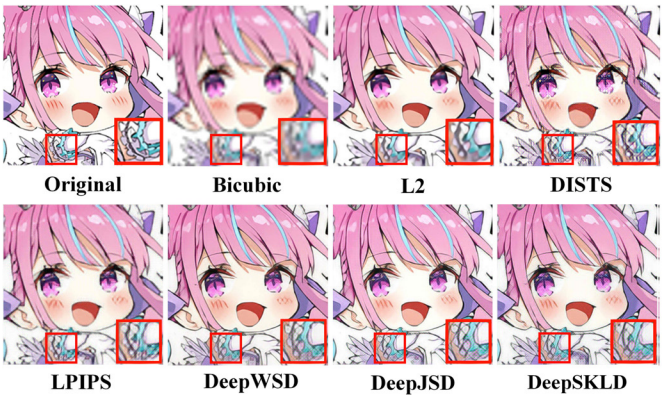

Fig. 12. Superresolution results obtained for the comic image.

TABLE 5
Comparison among several FR-IQA methods as loss functions in the superresolution task. The bold numbers are the three highest scores in each column. HyperIQA↑ means that the higher the value is, the better the perceptual quality.

HyperIQA↑	Set5 [59]	Set14 [59]	Manga109 [60]
Bicubic	35.7844	35.4403	41.4170
L_1	42.5367	49.7005	**67.3087**
LPIPS	40.2194	46.6953	64.6380
DISTS	**44.4219**	50.7813	64.7674
DeepWSD	**46.6713**	**54.7660**	67.2606
DeepJSD	42.5074	**51.4235**	**67.9386**
DeepSKLD	**45.3492**	**54.0678**	**68.1223**

distribution measures in terms of comparing deep features. Specifically, three widely used network structures, SqueezeNet [55], AlexNet [56] and VGG16 [5], are employed, and the results are shown in Fig. 9 and Fig. 10. These networks also generate satisfactory quality assessment results, but the VGG19 network can achieve the best results, so we choose it as the backbone of DeepWSD, DeepJSD, and DeepSKLD.

6 Applications

The proposed FR-IQA measures can be applied as perceptual losses in several image reconstruction tasks. Herein, we show two applications in the image superresolution task and the image denoising task.

6.1 Image superresolution

The image superresolution task aims at reconstructing a low-resolution image into its possible superresolution form, and the result may not be unique. Among all the results, we need to search for the most visually satisfactory result. Typically, different network structures and different training loss functions lead to different reconstruction results. Herein, we focus on the effect of the perceptual loss and fix the network structure as a deep residual channel attention network (RCAN) [57]. Beyond using DeepWSD, DeepJSD, and DeepSKLD to train the network, we also use the L_1 loss, DISTS, and LPIPS for training. We train the RCAN [57] with the DIVerse-2K (DIV2K) dataset [58] for 10 epochs by using the adaptive moment estimation (Adam) optimizer. The maximum number of iterations for each epoch is set to 20k, and the training process is conducted on image patches that are randomly scratched from the original image with a size of 192×192. The batch size is set to 16, and the upscaling factor is 4. We first use bicubic interpolation to downsample the image patches for training. The other settings are the same as the settings in the RCAN [57]. The reconstruction results are shown in Fig. 11. An obvious characteristic of DeepWSD, DeepSKLD, and DeepJSD is that they can retain sharp edges and fine textures. For example, the three proposed methods all retain the unsmooth texture of elephant noses without introducing artefacts. However, LPIPS and the L_1 loss fail to retain such textures, and LPIPS even introduces noise to the results. Moreover, DISTS can also generate unsmooth results, but such results are not visually satisfying. Liao *et al.* [15] also reported that using DISTS as the perceptual loss may risk the introduction of wave-like artefacts. For better interpretation, we present another superresolution result for the comic image in Fig. 12, which does not have complex textures and tends to be smooth;

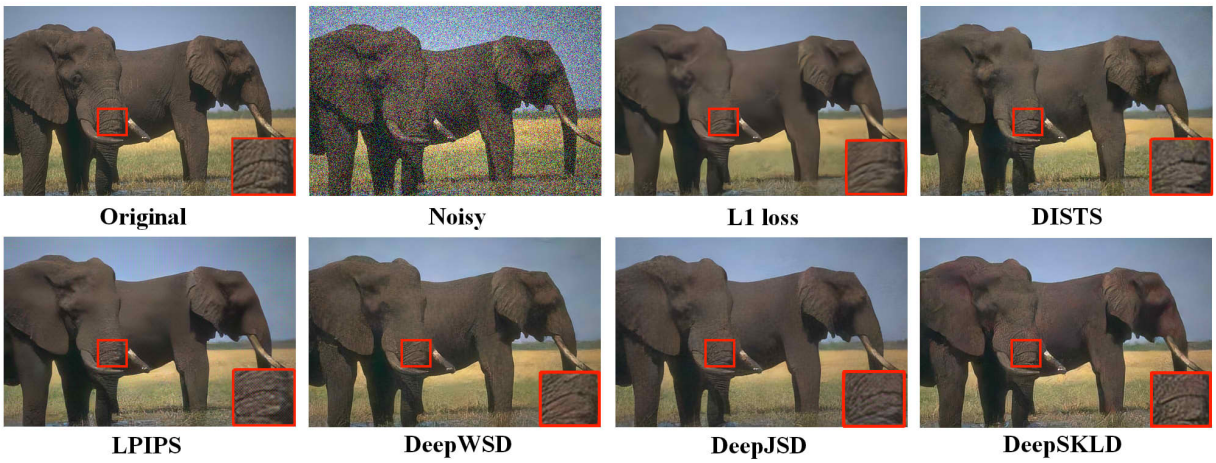

Fig. 13. Denoising results obtained for the elephant image, which is a natural image. The noisy image is contaminated by additive Gaussian noise with a variance of $\sigma = 10$ (with respect to [0,255]).

TABLE 6
Comparison among several FR-IQA methods as loss functions in the image denoising task. The bold numbers are the three highest scores in each column. HyperIQA↑ means that the higher value is, the better the perceptual quality.

HyperIQA↑	Set5 [59]	Set14 [59]	Manga109 [60]
L_1	42.6820	52.6146	**70.3167**
LPIPS	34.4460	51.1156	67.0823
DISTS	43.8713	53.5581	**69.6534**
DeepWSD	**48.7731**	**56.5976**	68.7241
DeepJSD	**45.2379**	**54.7754**	**69.2623**
DeepSKLD	**46.1086**	**54.8391**	68.6905

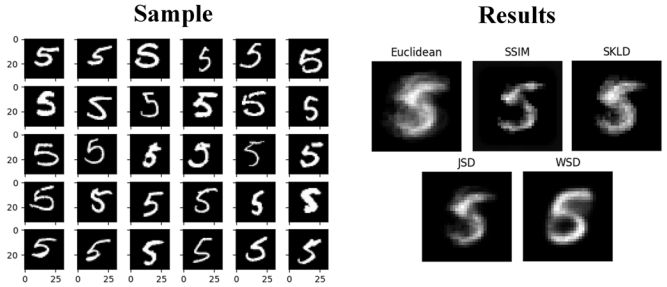

Fig. 14. Average images of the same digit obtained using different measures.

thus, intentionally generating unsmooth results does not lead to visual satisfaction. In contrast, the three proposed FR-IQA measures still retain sharp edges and some fine textures in the comic image, as magnified in the red zone. Moreover, we also provide a quantitative evaluation of three widely used superresolution datasets: Set5 [59], Set14 [59] and Manga109 [60]. The first two datasets contain natural images, and the last dataset contains comic images. For a fair comparison, the quantitative evaluation index is chosen as an NR-IQA approach, HyperIQA [61], whose value is in $[0, 100]$, and the larger the value is, the better. We compute the average score for the images in the datasets as the final performance of each model. The results are shown in Table 5. On different image datasets, the three proposed FR-IQA measures always earn the top three rankings.

6.2 Image denoising

The image denoising task aims at removing noise and reconstructing visually satisfactory noise-free images. Herein, we fix the network structure as the fast and flexible denoising network (FFDNet) [62] and train it with the L1 loss, DISTS [6], LPIPS [12], and the three proposed measures. For training, we introduce Gaussian noise with a strength of $\sigma = 10$ (with respect to [0,255]) and set the patch size to 48. All other settings are the same as the initial settings of FFDNet [62], and we present the natural image reconstruction results in Fig. 13. The results of DeepWSD, DeepJSD, and DeepSKLD contain much texture information that is clear to read, indicating that they have the ability to retain fine texture information during image denoising. In contrast, the result of the L_1 loss is rather vague, and the result of LPIPS [12] contains grid-like artefacts. Such grid-like artefacts may come from the pixelwise measure used in the deep feature domain, as artefacts do not occur in the results of other pixel correlation measures. On the other hand, the results of DISTS [6] also contain satisfactory texture information, and we surmise that the key to retaining texture information is to use the correlation-based measures in the deep feature domain. The quantitative evaluation results are presented in Table 6, among which the three proposed measures always earn the top three ranks; thus, we empirically conclude that the three proposed deep-based distribution measures can serve as perceptual losses in retaining structure and texture and be applied to different image reconstruction tasks.

6.3 Key to generating sharp edges and fine textures

The notable property of DeepWSD, DeepJSD, and DeepSKLD in image reconstruction tasks is that they can maintain sharp structures and fine textures. Herein, we argue that the key reason for this lies in the conformal property [63] of pixel correlation measures. That is, these measures possess the ability to capture the geometric information of several aligned distributions during optimization. We use Fig. 14 to illustrate the main idea. In this figure, we attempt to calculate the average image from digit image samples in the

Modified National Institute of Standards and Technology (MNIST) dataset [64], and the results are shown on the right side of the figure. This problem is defined as

$$\arg\min_{I^* \in \mathcal{P}} \sum_i D(I, S_i), i = 1...n, \tag{19}$$

where I^* is the optimal average image and the S_i are the image samples. $I^* \in \mathcal{P}$ restricts the results to lie in the same image space with the samples. When using the Euclidean norm, we can derive that the closed-form solution to Eq. (19) is to average all samples pixel by pixel, which leads to vague results. However, with pixel correlation measures such as the SSIM, WSD, JSD, and SKLD, the overall geometric shape of the digit will be maintained; thus, a clear structure is obtained. This property is important in image reconstruction tasks. Specifically, there are several local minima for the reconstruction network, and different local minima lead to different reconstructed images. The training loss attempts to find the average good reconstructed image and guide the training of the weights to this average good reconstruction result [65]. As such, networks trained with the L_2 norm will generate vague results, whereas pixel correlation measures attempt to preserve geometric information and thus guide a network to reconstruct images with a clear structure and fine textures.

7 Conclusion

We propose a new philosophy for designing deep network-based FR-IQA measures. Specifically, we point out that the pursuit of pixel correlation fidelity in deep feature comparisons is highly important and distribution measures are suitable for such comparisons. Then, we introduce three distribution measures, the WSD, JSD, and SKLD, to estimate the perceptual degradation in the deep features of the VGG19 [5] network, leading to DeepWSD, DeepJSD, and DeepSKLD. DeepWSD, DeepJSD, and DeepSKLD are superior to other measures in that they do not require training but correlate well with the MOS scores of several IQA datasets, indicating that they have advanced score prediction abilities and are robust to different distortions. Moreover, an MDC experiment and applications of these IQA measures demonstrate that they have perceptual quality isolines and can serve as advanced perceptual losses. Two key reasons support the superiority of these measures. First, the pursuit of pixel correlation fidelity is an effective way to approach the pursuit of perceptual information fidelity in the HVS. Second, the three distribution measures can capture the degradation of the pixel correlations in deep features.

The limitation of the DeepWSD, DeepJSD, and DeepSKLD approaches is that they are too simple and still incapable of presenting the highly nonlinear HVS. Moreover, the patch comparison strategy may also constrain these distribution measures to capture long-range pixel correlations. Nonetheless, the three deep-based distribution measures can incorporate more powerful feature decomposition backbones, such as residual networks and transformers, and perform better in score prediction and perceptual optimization. We hope that the pursuit of pixel correlation fidelity theory can also inspire research on extracting and interpreting the meaning of deep network features from a new point of view.

References

[1] S. Gao and X. Zhuang, "Bayesian image super-resolution with deep modeling of image statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[2] H. L. Tan, Z. Li, Y. H. Tan, S. Rahardja, and C. Yeo, "A perceptually relevant mse-based image quality metric," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4447–4459, 2013.
[3] C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[4] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2019.
[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image refcognition," *arXiv: Computer Vision and Pattern Recognition*, vol. abs/1409.1556, 2015.
[6] 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*, vol. abs/2004.07728, 2020.
[7] Z. Wang and A. C. Bovik, "Modern image quality assessment," vol. Synthesis Lectures on Image, Video, and Multimedia Processing, pp. 1–156, 2006.
[8] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.
[9] Z. Wang and K. Ma, "Active fine-tuning from gmad examples improves blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4577–4590, 2022.
[10] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.
[11] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, jun 2018.
[13] 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," *IEEE Conference on Computer Vision and Pattern Recognition 2022*, vol. abs/2204.08763, 2022.
[14] K. Ding, Y. Liu, X. Zou, S. Wang, and K. Ma, "Locally adaptive structure and texture similarity for image quality assessment," *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2483–2491, 2021.
[15] X. Liao, B. Chen, H. Zhu, S. Wang, M. Zhou, and S. Kwong, "Deepwsd: Projecting degradations in perceptual space to wasserstein distance in deep feature space," *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
[16] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488–1499, 2012.
[17] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1258–1281, 2021.
[18] D. Johnson and S. Sinanovic, "Monge's optimal transport distance for image classification," *arXiv: Computer Vision and Pattern Recognition*, 2016.
[19] R. Kumari and D. Sharma, "Generalized 'useful' ag and 'useful' js-divergence measures and their bounds," *International Journal of Engineering, Science and Mathematics*, vol. 7, no. 1, pp. 441–450, 2018.
[20] M. Snow and J. V. Lent, "Symmetrizing the kullback-leibler distance," *IEEE Transactions on Information Theory*, 2001.
[21] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual discriminability," *Journal of Vision*, vol. 8, pp. 1–13, Sep 2008.
[22] Z. Duanmu, W. Liu, Z. Wang, and Z. Wang, "Quantifying visual image quality: A bayesian view," *Annual Review of Vision Science*, vol. 7, 01 2021.

[23] S. Fischer, F. Sroubek, L. Perrinet, R. Redondo, and G. Cristobal, "Self-invertible 2d log-gabor wavelets," *International Journal of Computer Vision*, vol. 75, pp. 231–246, 08 2007.

[24] C. Thillou and B. Gosselin, "Character segmentation-by-recognition using log-gabor filters," *International Conference on Pattern Recognition*, vol. 2, pp. 901–904, 01 2006.

[25] M. Morrone and D. Burr, "Feature detection in human vision: a phase-dependent energy model.," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 235, pp. 221–245, Dec. 1988.

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

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

[28] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[29] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[30] Z. Wang and E. Simoncelli, "Reduce-reference image quality assessment using a wavelet-domain natural image statistic model," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5666, 03 2005.

[31] M. Liu, K. Gu, G. Zhai, P. Le Callet, and W. Zhang, "Perceptual reduced-reference visual quality assessment for contrast alteration," *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 71–81, 2017.

[32] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4559–4565, 2017.

[33] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1301–1313, 2018.

[34] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, p. 2000, 2000.

[35] W. Yang, L. Xu, X. Chen, F. Zheng, and Y. Liu, "Chi-squared distance metric learning for histogram data," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–12, 04 2015.

[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *the 34th International Conference on Machine Learning - Volume 70*, p. 214–223, 2017.

[37] V. Ming and L. Holt, "Efficient coding in human auditory perception," *The Journal of the Acoustical Society of America*, vol. 126, pp. 1312–20, 10 2009.

[38] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology-Paris*, vol. 100, no. 1-3, pp. 70–87, 2006.

[39] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2014.

[40] J. Liu, W. Zhou, X. Li, J. Xu, and Z. Chen, "Liqa: Lifelong blind image quality assessment," *IEEE Transactions on Multimedia*, pp. 1–16, 2022.

[41] C. R. Sims, "Efficient coding explains the universal law of generalization in human perception," *Science*, vol. 360, no. 6389, pp. 652–656, 2018.

[42] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[43] 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, pp. 011006–011006–21, Jan. 2010.

[44] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[45] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database release 2," *Journal of Biomedical Science and Engineering*, June 2010.

[46] A. Ninassi, F. Autrusseau, and P. Le Callet, "Pseudo no reference image quality metric using perceptual data hiding," *Human Vision and Electronic Imaging*, 02 2006.

[47] Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," *Proceedings of the IEEE Asilomar Conference Signals, Systems and Computers*, 02 2004.

[48] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 2009.

[49] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *European Conference on Computer Vision (ECCV) 2020*, pp. 633–651, Springer International Publishing, 2020.

[50] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 1693–1697, 2012.

[51] D. Amir and Y. Weiss, "Understanding and simplifying perceptual distances," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12221–12230, 2021.

[52] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[53] M. Delbracio, H. Talebei, and P. Milanfar, "Projected distribution loss for image enhancement," (Los Alamitos, CA, USA), pp. 1–12, IEEE Computer Society, may 2021.

[54] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced wasserstein distances," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[55] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," *International Conference on Learning Representations,(ICIR)*, Feb. 2017.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, p. 84–90, may 2017.

[57] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020.

[58] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, *et al.*, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[59] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi-Morel, "Low-complexity single image super-resolution based on non-negative neighbor embedding," *Electronic Proceedings of the British Machine Vision Conference 2012*, 09 2012.

[60] K. Aizawa, A. Fujimoto, A. Otsubo, T. Ogawa, Y. Matsui, K. Tsubota, and H. Ikuta, "Building a manga dataset "manga109" with annotations for multimedia applications," *IEEE MultiMedia*, vol. 27, no. 2, pp. 8–18, 2020.

[61] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3664–3673, 2020.

[62] Z. Kai, Z. Wangmeng, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[63] M. Jin, X. Gu, Y. He, and Y. Wang, *Conformal Geometry: Computational Algorithms and Engineering Applications*. Springer International Publishing, 2018.

[64] Y. LeCun and C. Cortes, "Mnist: large-scale handwritten digit database," *Proceedings of the Institute of Radio Engineers*, vol. 86, no. 11, pp. 2278–2323, 1998.

[65] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

Image Quality Assessment: Measuring Perceptual Degradation via Distribution Measures in a Deep Feature Space

Xingran Liao, Xuekai Wei, Mingliang Zhou, Zhengguo Li, *Senior Member, IEEE*, and Sam Kwong, *Fellow, IEEE*

APPENDIX A
PROOF OF THEOREM 4.1.

WSD case. We start from the following lemma [1].

Lemma A.1. *Let $\mathcal{X} \sim \mathcal{N}_x(\mu_x, \sigma_x^2)$ and $\mathcal{Y} \sim \mathcal{N}_y(\mu_y, \sigma_y^2)$ be one-dimensional Gaussian distributions, where μ_x, μ_y and σ_x, σ_y are the means and standard deviations of the corresponding distributions; then,*

$$WSD_2(\mathcal{X}, \mathcal{Y}) = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2, \tag{1}$$

Such a lemma is a trivial one-dimensional corollary of proposition 7 in [1]; we omit its proof and use it to derive our main conclusion.

Proof of Theorem 4.1 for the WSD case. According to Eq. (1),

$$\begin{aligned} WSD_2(\mathcal{X}, \mathcal{Y}) =& (\mu_x^2 + \mu_y^2)(1 - \frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2}) \\ &+ (\sigma_x^2 + \sigma_y^2)(1 - \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}), \end{aligned} \tag{2}$$

For $\sigma_x\sigma_y$,

$$\begin{aligned} \sigma_x\sigma_y &= (\int_\Omega (\mathcal{X} - \mathbb{E}(\mathcal{X}))^2 dP)^{\frac{1}{2}} (\int_\Omega (\mathcal{Y} - \mathbb{E}(\mathcal{Y}))^2 dP)^{\frac{1}{2}} && (3)\\ &\geq \int_\Omega |(\mathcal{X} - \mathbb{E}(\mathcal{X}))(\mathcal{Y} - \mathbb{E}(\mathcal{Y}))| dP && (4)\\ &\geq \int_\Omega (\mathcal{X} - \mathbb{E}(\mathcal{X}))(\mathcal{Y} - \mathbb{E}(\mathcal{Y})) dP && (5)\\ &= \sigma_{xy}, && (6) \end{aligned}$$

where the first and the second equations are the definitions of the standard deviations and $\mathbb{E}(\mathcal{X})$ is the expectation of the random variable $\mathcal{X}$. The third equation is the definition of the expectation, and the fourth inequality holds because of the Hölder inequality, which takes the form

$$\int_\Omega |f(z)g(z)| dz \leq (\int_\Omega f(z)^p dz)^{1/p} (\int_\Omega g(z)^p dz)^{1/p}, \tag{7}$$

where $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. By applying Eq. (6) to Eq. (1), we obtain

$$WSD_2(\mathcal{X}, \mathcal{Y}) \leq C_1'(1 - \frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2}) + C_2'(1 - \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2}), \tag{8}$$

where $C_1' = \mu_x^2 + \mu_y^2$ and $C_2' = \sigma_x^2 + \sigma_y^2$. □

JSD case. Similar to the WSD, we start from a lemma based on KLD comparing the divergence of two Gaussian distributions.

Lemma A.2. *Let $\mathcal{X} \sim \mathcal{N}_x(\mu_x, \sigma_x^2)$ and $\mathcal{Y} \sim \mathcal{N}_y(\mu_y, \sigma_y^2)$ be one-dimensional Gaussian distributions; then, for KLD,*

$$KLD(\mathcal{X}, \mathcal{Y}) = \log \frac{\sigma_y}{\sigma_x} - \frac{1}{2} + \frac{\sigma_x^2 + (\mu_x - \mu_y)^2}{2\sigma_y^2}. \tag{9}$$

Proof of Lemma A.2. Denote the probability density of $\mathcal{X}$ $p_\mathcal{X}(z) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp(-\frac{(z-\mu_x)^2}{2\sigma_x^2})$ and the probability density of $\mathcal{Y}$ $p_\mathcal{Y}(z) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp(-\frac{(z-\mu_y)^2}{2\sigma_y^2})$. According to the definition of KLD,

$$\begin{aligned} &KLD(\mathcal{X}, \mathcal{Y}) && (10)\\ &= \int_z p_\mathcal{X}(z) \log \frac{\frac{1}{\sqrt{2\pi}\sigma_x} \exp(-\frac{(z-\mu_x)^2}{2\sigma_x^2})}{\frac{1}{\sqrt{2\pi}\sigma_y} \exp(-\frac{(z-\mu_y)^2}{2\sigma_y^2})} dz && (11)\\ &= \int_z p_\mathcal{X}(z) (\log \frac{\sigma_y}{\sigma_x} - \frac{(z-\mu_x)^2}{2\sigma_x^2} + \frac{(z-\mu_y)^2}{2\sigma_y^2}) dz. && (12) \end{aligned}$$

Note that

$$\int_z p_\mathcal{X}(z) dz = 1, \tag{13}$$

due to the definition of the probability density function and

$$\int_z p_\mathcal{X}(z)(z - \mu_x)^2 dz = \sigma_x^2, \tag{14}$$

due to the definition of the variance. Additionally, the following equations are used:

$$\begin{aligned} (z - \mu_y)^2 &= (z - \mu_x + \mu_x - \mu_y)^2 && (15)\\ &= (z - \mu_x)^2 + 2(z - \mu_x)(\mu_x - \mu_y) + (\mu_x - \mu_y)^2. && (16) \end{aligned}$$

and it can be derived that

$$\begin{aligned} &\frac{1}{2\sigma_y^2} \int_z p_\mathcal{X}(z)(z - \mu_y)^2 \, dz && (17)\\ &= \frac{1}{2\sigma_y^2} (\sigma_x + (\mu_x - \mu_y)^2). && (18) \end{aligned}$$

Applying Eq. (13), Eq. (14) and Eq. (16) to Eq. (12), we can finally obtain the result. □

Proof of Theorem 4.1 for the JSD case.

$$JSD(\mathcal{X},\mathcal{Y}) = \frac{1}{2}KLD(\mathcal{X},\mathcal{M}) + \frac{1}{2}KLD(\mathcal{Y},\mathcal{M}) \tag{19}$$

$$= \frac{1}{2}\int_z p_{\mathcal{X}}(z)\log\frac{2p_{\mathcal{X}}(z)}{p_{\mathcal{X}}(z)+p_{\mathcal{Y}}(z)}dz + \frac{1}{2}\int_z p_{\mathcal{Y}}(z)\log\frac{2p_{\mathcal{Y}}(z)}{p_{\mathcal{X}}(z)+p_{\mathcal{Y}}(z)}dz \tag{20}$$

$$= \frac{1}{2}\int_z p_{\mathcal{X}}(z)\log\frac{2}{1+p_{\mathcal{Y}}(z)/p_{\mathcal{Y}}(z)}dz + \frac{1}{2}\int_z p_{\mathcal{Y}}(z)\log\frac{2}{p_{\mathcal{X}}(z)/p_{\mathcal{Y}}(z)+1}dz \tag{21}$$

$$= \log 2 - \frac{1}{2}\int_z p_{\mathcal{X}}(z)\log(1+\frac{p_{\mathcal{Y}}(z)}{p_{\mathcal{X}}(z)})dz - \frac{1}{2}\int_z p_{\mathcal{Y}}(z)\log(1+\frac{p_{\mathcal{X}}(z)}{p_{\mathcal{Y}}(z)})dz \tag{22}$$

$$\leq \log 2 - \frac{1}{2}\int_z [p_{\mathcal{X}}(z)\log\frac{p_{\mathcal{Y}}(z)}{p_{\mathcal{X}}(z)} + p_{\mathcal{Y}}(z)\log\frac{p_{\mathcal{X}}(z)}{p_{\mathcal{Y}}(z)}]dz \tag{23}$$

$$= \log 2 - \frac{1}{2}\int_z p_{\mathcal{X}}(z)[\log\frac{\sigma_y}{\sigma_x} - \frac{(z-\mu_y)^2}{2\sigma_y^2} + \frac{(z-\mu_x)^2}{2\sigma_x^2}]dz - \frac{1}{2}\int_z p_{\mathcal{Y}}(z)[\log\frac{\sigma_x}{\sigma_y} - \frac{(z-\mu_x)^2}{2\sigma_x^2} + \frac{(z-\mu_y)^2}{2\sigma_y^2}]dz \tag{24}$$

$$= \log 2 - \frac{1}{2} + \frac{1}{2}\int_z [p_{\mathcal{X}}(z)\frac{(z-\mu_y)^2}{2\sigma_y^2} + p_{\mathcal{Y}}(z)\frac{(z-\mu_x)^2}{2\sigma_x^2}]dz \tag{25}$$

$$= \log 2 - \frac{1}{2} + \frac{1}{2}\int_z \frac{p_{\mathcal{X}}(z)z^2 - 2\mu_y p_{\mathcal{X}}(z)z + \mu_y^2 p_{\mathcal{X}}(z)}{2\sigma_y^2}dz + \frac{1}{2}\int_z \frac{p_{\mathcal{Y}}(z)z^2 - 2\mu_x p_{\mathcal{Y}}(z)z + \mu_x^2 p_{\mathcal{Y}}(z)}{2\sigma_y^2}dz \tag{26}$$

$$= \log 2 - \frac{1}{2} + \frac{1}{2}(\frac{\sigma_x^2+\mu_x^2-2\mu_x\mu_y+\mu_y^2}{2\sigma_y^2} + \frac{\sigma_y^2+\mu_y^2-2\mu_x\mu_y+\mu_x^2}{2\sigma_x^2}) \tag{27}$$

$$= \log 2 - \frac{1}{2} + \frac{\sigma_x^2+(\mu_x-\mu_y)^2}{4\sigma_y^2} + \frac{\sigma_y^2+(\mu_x-\mu_y)^2}{4\sigma_x^2} \tag{28}$$

where in Eq. (28), we use the definitions of variance $\int_z p_{\mathcal{Y}}(z)z^2 = \sigma_y^2+\mu_y^2$ and $\int_z p_{\mathcal{X}}(z)z^2 = \sigma_x^2+\mu_x^2$. Then, using a similar trick in providing the WSD case, we can obtain the upper bound of the JSD case as

$$JSD(\mathcal{X},\mathcal{Y}) \leq \log 2 - \frac{1}{2} + \frac{\sigma_x^4+\sigma_y^4+(\sigma_x^2+\sigma_y^2)(\mu_x-\mu_y)^2}{4\sigma_x^2\sigma_y^2} \tag{29}$$

$$= \log 2 + \frac{(\sigma_x^2-\sigma_y^2)^2+(\sigma_x^2+\sigma_y^2)(\mu_x-\mu_y)^2}{4\sigma_x^2\sigma_y^2} \tag{30}$$

Then, we set $C_1'' = \frac{\sigma_x^2+\sigma_y^2}{4\sigma_x^2\sigma_y^2(\mu_x^2+\mu_y^2)}$, $C_2'' = \frac{(\sigma_x^2+\sigma_y^2)(\sigma_x+\sigma_y)^2}{4\sigma_x^2\sigma_y^2}$, and $C_3'' = \log 2$ and apply $\sigma_{xy} < \sigma_x\sigma_y$; then, we obtain

$$JSD(\mathcal{X},\mathcal{Y}) \leq C_1''(1-\frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2}) + C_2''(1-\frac{2\sigma_{xy}}{\sigma_x^2+\sigma_y^2}) + C_3'' \tag{31}$$

□

SKLD case. The proof of the SKLD case is the direct conclusion of Lemma A.2

Proof of Theorem 4.1 for the SKLD case.

$$SKLD(\mathcal{X},\mathcal{Y}) = \frac{1}{2}KLD(\mathcal{X},\mathcal{Y}) + \frac{1}{2}KLD(\mathcal{Y},\mathcal{X}) \tag{32}$$

$$= -1 + \frac{\sigma_x^2+(\mu_x-\mu_y)^2}{2\sigma_y^2} + \frac{\sigma_y^2+(\mu_x-\mu_y)^2}{2\sigma_x^2} \tag{33}$$

$$= -1 + \frac{\sigma_x^4+(\sigma_x^2+\sigma_y^2)(\mu_x-\mu_y)^2+\sigma_y^4}{2\sigma_x^2\sigma_y^2} \tag{34}$$

$$= \frac{(\sigma_x^2-\sigma_y^2)^2}{2\sigma_x^2\sigma_y^2} + \frac{(\sigma_x^2+\sigma_y^2)(\mu_x-\mu_y)^2}{2\sigma_x^2\sigma_y^2} \tag{35}$$

$$= \frac{1}{2\sigma_x^2\sigma_y^2}[(\sigma_x+\sigma_y)^2(\sigma_x-\sigma_y)^2 + (\sigma_x^2+\sigma_y^2)(\mu_x-\mu_y)^2] \tag{36}$$

$$= \frac{\sigma_x^2+\sigma_y^2}{2\sigma_x^2\sigma_y^2}[(\sigma_x-\sigma_y)^2+(\mu_x-\mu_y)^2] + \frac{(\sigma_x-\sigma_y)^2}{\sigma_x\sigma_y} \tag{37}$$

$$= \frac{(\sigma_x^2+\sigma_y^2)^2}{2\sigma_x^2\sigma_y^2}(1-\frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}) + \frac{(\sigma_x^2+\sigma_y^2)(\mu_x^2+\mu_y^2)}{2\sigma_x^2\sigma_y^2}(1-\frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2}) + \frac{(\sigma_x-\sigma_y)^2}{\sigma_x\sigma_y} \tag{38}$$

Let $C_1''' = \frac{(\sigma_x^2+\sigma_y^2)^2}{2\sigma_x^2\sigma_y^2}$, $C_2''' = \frac{(\sigma_x^2+\sigma_y^2)(\mu_x^2+\mu_y^2)}{2\sigma_x^2\sigma_y^2}$, and $C_3''' = \frac{(\sigma_x-\sigma_y)^2}{\sigma_x\sigma_y}$, and use $\sigma_{xy} \leq \sigma_x\sigma_y$; then, we obtain the upper bound:

$$SKLD(\mathcal{X},\mathcal{Y}) \leq C_1'''(1-\frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}) + C_2'''(1-\frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2}) + C_3''' \tag{39}$$

□

The upper bounds of the WSD, JSD, and SKLD take very similar forms as SSIM variants [2], which are complete metrics that lie in [0,2]. We denote it as $SSIM_{add}$, and its definition is

$$SSIM_{add}(\mathcal{X},\mathcal{Y}) = 2 - \frac{2\mu_x\mu_y+\epsilon_1}{\mu_x^2+\mu_y^2+\epsilon_1} - \frac{2\sigma_{xy}+\epsilon_2}{\sigma_x^2+\sigma_y^2+\epsilon 2} \tag{40}$$

$$\leq 2 - \frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2} - \frac{2\sigma_{xy}}{\sigma_x^2+\sigma_y^2} \tag{41}$$

where the inequality holds because $\frac{a}{b} \leq \frac{a+c}{b+c}$ and in DISTS [3], Ding *et al.* used such an SSIM variant to compare the deep network features and obtained good results.

Herein, the purpose of the theorem and our proof is not to show the equivalence of the WSD, JSD, and SKLD to the SSIM variant. We focus on revealing the comparison philosophy of these measures, that is, comparing the difference in the pixel correlation among a group of pixels, such as comparing the mean and variance. The Gaussian hypothesis forms a bridge to better comprehend such a philosophy, and

the core idea is that we want to capture, comprehend, and compare the correlation in the deep network features; even if we cannot understand it intuitively, we should not ignore the existing correlation.

REFERENCES

[1] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions.," *Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231 – 240, 1984.

[2] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488–1499, 2012.

[3] 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*, vol. abs/2004.07728, 2020.