FROM DISTANCE TO DEPENDENCY: A PARADIGM SHIFT OF FULL-REFERENCE IMAGE QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning-based full-reference image quality assessment (FR-IQA) models typically rely on the feature distance between the reference and distorted images. However, the underlying assumption of these models that the distance in the deep feature domain could quantify the quality degradation does not scientifically align with the invariant texture perception, especially when the images are generated artificially by neural networks. In this paper, we bring a radical shift in inferring the quality with learned features and propose the Deep Image Dependency (DID) based FR-IQA model. The feature dependency facilitates the comparisons of deep learning features in a high-order manner with Brownian distance covariance, which is characterized by the joint distribution of the features from reference and test images, as well as their marginal distributions. This enables the quantification of the feature dependency against nonlinear transformation, which is far beyond the computation of the numerical errors in the feature space. Experiments on image quality prediction, texture image similarity, and geometric invariance validate the appealing performance of our proposed measure, and the implementation will be publicly available.

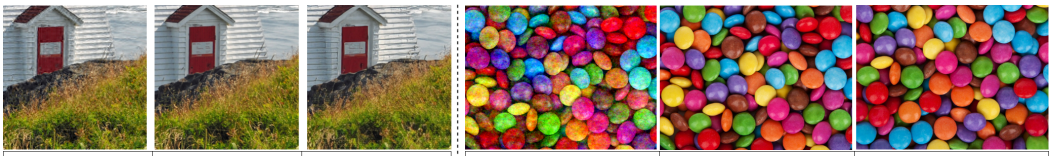
1 INTRODUCTION

The primary target of objective image quality assessment (IQA) is to automatically predict the perceptual visual quality, providing a cost-effective alternative for the cumbersome subjective user study Athar & Wang (2019). Full-reference IQA (FR-IQA) feeds the pristine image $\mathbf{x}$ and the counterpart distorted image $\mathbf{y}$ into different perceptual distance measures. The predicted quality score is used to evaluate the image processing system and optimize various real-world applications, such as image compression, restoration, and rendering. The FR-IQA models can be summarized from a Bayesian perspective Duanmu et al. (2021):

$$p(s|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}, \beta) = \mathcal{N}(s|d(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}), \beta), \tag{1}$$

where s is the subjective quality rating which is assumed to follow a Gaussian distribution with the mean $d(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ and variance β. Herein, we denote the $d(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ as the perceptual distance measured between $\mathbf{x}$ and $\mathbf{y}$, with the θ encoding the prior knowledge of human vision system (HVS). In the last decades, efforts have been mainly devoted to exploring a meaningful and powerful parameter distribution $p(\boldsymbol{\theta})$, aiming to achieve a fully perceptual consistent measure. Those models can be classified into knowledge-driven and data-driven approaches.

The *knowledge-driven approaches* have dominated the FR-IQA models for more than a half-century. Mean squared error (MSE) is one of the most popular error visibility methods owing to its simplicity, clear physical meaning and desired properties for optimization, but it shows poor correlation with the HVS Wang & Bovik (2009). Afterward, methods that correlate better with HVS were developed, such as the structural similarity index (SSIM) Wang et al. (2004), the visual information fidelity (VIF) Sheikh & Bovik (2006), and the normalized Laplacian pyramid distance (NLPD) Laparra et al. (2016). Recently, the *data-driven approaches* are prevailing due to the perceptual meaningful characteristic of deep pre-trained convolution neural network (CNN), denoted as $\tilde{\mathbf{x}} = h(\mathbf{x}, \boldsymbol{\theta}_\mathbf{c})$ and $\boldsymbol{\theta}_c$ is the parameters of network $h(\cdot)$. In the deep learning feature domain, various distance measures have been developed, *i.e.*, $d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \boldsymbol{\theta}_c)$. For the element-wise methods, Johnson *et al*. constructs the



(a) JPEG2000	(b) Reference image	(c) GAN	(d) Pink noise	(e) Reference image	(f) Resampling image
0.621	Human ↑	**0.700**	0.658	Human ↑	**0.871**
22.818	PSNR ↑	20.722	**16.026**	PSNR ↑	9.900
0.521	SSIM ↑	0.440	**0.509**	SSIM ↑	0.101
0.253	LPIPS ↓	0.300	**0.624**	LPIPS ↓	0.631
0.152	DISTS ↓	0.154	0.417	DISTS ↓	**0.110**
0.830	DID ↑	**0.854**	0.371	DID ↑	**0.993**

Figure 1: Illustration of the contrastive preference by human and existing quality measures. Left: Human and DID prefer the image (c) from generative adversarial network (GAN) Navarrete Michelini et al. (2018) over the JPEG2000 distorted image (a), but the distance-based measure (PSNR, SSIM, LPIPS, and DISTS) prefer image (a) over (c); Right: Human, DISTS, and DID prefer the resampling texture image (f) over the pink noise image (d), but the PSNR, SSIM, and LPIPS prefer image (d) over (f). The better quality scores are highlighted in boldface.

perceptual loss by computing the weighted summation of ℓ_2-norm distance in both image and feature domains Johnson et al. (2016). The learned perceptual image patch similarity (LPIPS) Zhang et al. (2018) calculates the weighted MSE result with corresponding deep representations, attempting to account for the "unreasonable" effectiveness. Ding *et al*. computes the global or local deep feature statistics (*i.e*., mean and variance) to unify the structure and texture similarity (DISTS) Ding et al. (2020; 2021). In addition, the internal feature dependency (*i.e*., distribution-wise) comparisons also play an essential role in perceptual visual quality predictions. Representative examples include style loss Gatys et al. (2016), deep self-dissimilarity Kligvasser et al. (2021), and deep Wasserstein distance (DeepWSD) Liao et al. (2022).

The knowledge-driven and data-driven approaches primarily rely on the deterministic comparisons of the images in various domains. The deterministic comparisons based on distance metrics, though faithfully reflecting the fidelity, may fail when the test images are generated instead of physically acquired. This can be attributed to the rooted view that images with perfect quality can be feasibly modeled as the output of a stochastic source. One example can be observed in Fig. 1, where the image (c) with pleasant texture is disfavored by the distance-based measures, due to the large distance caused by the generated textures. In addition, as shown in the right case of Fig. 1, human perception is usually invariant to the texture resampling even the resampling brings distance boosting in both signal space and feature space. Those phenomenons inspire us to bring the shift from the traditional distance-based FR-IQA paradigm, to the dependency modeling from statistical perspective. In particular, though statistical models dominate the no-reference IQA methods, much less work has been dedicated to characterizing the statistics for FR-IQA. More importantly, statistical and perceptual modeling of visual signals are broadly recognized as the dual problems Sheikh et al. (2005). As such, the shift is grounded on the mild assumption that the perception of texture variation can be well reflected by a reliable feature dependency measure.

In this work, we adopt the Brownian Distance Covariance (BDC) as the ideal feature dependency measure for FR-IQA, and propose the Deep Image Dependency (DID) based FR-IQA model. The BDC is defined as the weighted Euclidean distance between the joint characteristic function and the product of the marginal characteristic functions Székely & Rizzo (2009). The proposed DID model presents several desired advantages. First, the model is able to naturally capture the feature dependency against both linear and non-linear transformations. Second, in the deep feature space, the model does not rely on any trainable parameters, demonstrating promising flexibility and generalization capability. Third, the model presents superior performance on texture perception, no matter the texture is artificially generated, randomly reasmpled and geometrically transformed. Extensive experiments based upon classical IQA datasets, texture similarity datasets, and geometric transformation dataset demonstrate that DID achieves state-of-the-art performance according to the correlation with mean opinion scores (MOSs). It is also worth mentioning that though DID obtains competitive performance in quality evaluation tasks, it is independent of the training data (*i.e*., MOSs) and does not contain any controllable parameters.

2 RELATED WORK

2.1 FULL-REFERENCE IMAGE QUALITY ASSESSMENT

The early works for FR-IQA capture the distortion relying on signal fidelity measure *e.g.*, MSE, and peak signal-to-noise ratio (PSNR). Although those works enjoy the calculation simplicity and mathematical convenience, the low consistency with human perception has been widely criticized Lin et al. (2003). In Wang et al. (2004), the SSIM was proposed by introducing three important components, including luminance, contrast, and structural similarity. This work was extended to several more advanced quality measures, such as MS-SSIM Wang et al. (2003), IW-SSIM Wang & Li (2010) and CW-SSIM Wang & Simoncelli (2005). In the deep-learning era, pioneering works concentrate on image comparison in the deep learning feature space. For example, in LPIPS Zhang et al. (2018), the multi-scale features were extracted from the pre-trained VGG Simonyan & Zisserman (2015) network and the image quality is estimated by measuring the feature fidelity loss with Euclidean distance. Analogously, the combination of spatial averages and correlations of the feature maps was adopted in DISTS Ding et al. (2020), aiming for the estimation of texture similarity and structure similarity. The work was further improved by processing the structure and texture information adaptively in A-DISTS Ding et al. (2021). Instead of measuring the feature distance point-by-point, the the Wasserstein distance was utilized in DeepWSD Liao et al. (2022) to capture the quality contamination. Driven by the quality annotated data, the human perception knowledge can also be learned by CNN, such as DeepQA Kim & Lee (2017), WaDIQaM Bosse et al. (2017), PieAPP Prashnani et al. (2018) and JSPL Cao et al. (2022). However, compared with the features extracted from pre-trained networks, the learned models usually suffer from the over-fitting problem due to the limited labeled data.

2.2 DATA DEPENDENCY MEASURE

Classical data dependency is measured only in the linear scenario. For example, the product-moment correlation and covariance are two widely used dependency measures between two random variables. In the case that the joint distribution of two vectors is multivariate normal distribution, the covariance matrix can be used to measure the dependence of different dimensions. This measure which captures the high-order information is also utilized in computer vision tasks, *i.e.*, the style transferring Gatys et al. (2016) and deep self-similarity Kligvasser et al. (2021). For the image stylization, the Gram matrices of the neural activations of different CNN layers are extracted under the view that the channel dependency captured by Gram matrices well represents the artistic style of an image. However, in a more generic scene, the nonlinear or nonmonotone dependence is expected to be effectively captured. In Székely & Rizzo (2009); Székely et al. (2007), the BDC was proposed which is able to measure the data dependency efficiently even in the nonlinear scenario. The BDC is designed based on the construction of joint characteristics of two variables, such that it could be more effective than only the marginal distribution involved. Benefiting from the properties, the BDC was also introduced for few-shot classification tasks Xie et al. (2022), showing its robustness in different settings.

3 METHODOLOGY

3.1 PRELIMINARY OF BROWNIAN DISTANCE COVARIANCE

Let $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ be two random vectors, where p and q are their dimensions. The characteristic functions of X and Y are denoted as f_X and f_Y and their joint characteristic function is f_{XY}. Assuming X and Y have finite first moments, analogous to classical covariance, the BDC measure is defined as follows,

$$\mathcal{V}^2(X, Y; w) = \|f_{X,Y}(t, s) - f_X(t) f_Y(s)\|_w^2 , \tag{2}$$

where $\|\cdot\|_w^2$-norm is defined by,

$$\|\gamma(t, s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(t, s)|^2 w(t, s) dt ds. \tag{3}$$

The $w(t, s)$ is a positive weight function for which the integral above exists. As such,

$$\mathcal{V}^2(X, Y; w) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 \, w(t, s) dt ds. \tag{4}$$

To endow the $\mathcal{V}^2(X,Y;w)$ with the capability to capture the dependence between X and Y, a suitable weight function can be found as follows Székely & Rizzo (2009),

$$w(t,s) = \left(c_p c_q |t|_p^{1+p} |s|_q^{1+q}\right)^{-1}, \tag{5}$$

where

$$c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}, \quad c_q = \frac{\pi^{(1+q)/2}}{\Gamma((1+q)/2)}, \tag{6}$$

and $\Gamma(\cdot)$ is the complete gamma function. From the Eq. (4) and Eq. (5), the BDC measure can be formed by,

$$\begin{aligned} \mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t) f_Y(s)\|^2 \\ &= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{|t|_p^{1+p} |s|_q^{1+q}} dt ds. \end{aligned} \tag{7}$$

Herein, we omit the w in $\mathcal{V}^2(X,Y;w)$ for simplification. In practice, the observations of X and Y are usually discrete. For n i.i.d. observed random vectors $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$, the BDC measure $\mathcal{V}^2(X,Y)$ can be efficiently acquired by Székely & Rizzo (2009); Székely et al. (2007),

$$\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} \mathbf{A}_{kl} \mathbf{B}_{kl}, \tag{8}$$

where $\mathbf{A}_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$ and

$$a_{kl} = \|X_k - X_l\|_z, \quad \bar{a}_{k\cdot} = \frac{1}{n} \sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l}, = \frac{1}{n} \sum_{k=1}^{n} a_{kl}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2} \sum_{k,l=1}^{n} a_{kl}, \tag{9}$$

where $\|\cdot\|_z$ means the z-norm. Analogously, we define $\mathbf{B}_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$, for $k, l = 1, \ldots, n$. The model enjoys several interesting properties which are summarized as follows,

(i) $\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) \geq 0$.

(ii) $\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = 0$, if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

(iii) For all constant vector t_1, t_2, nonzero real number s_1, s_2 and orthogonal matrix $\mathbf{R_1}$, $\mathbf{R_2}$, $\mathcal{V}^2(t_1 + s_1 \mathbf{X} \mathbf{R_1}, t_2 + s_2 \mathbf{X} \mathbf{R_2}) = |s_1 s_2| \, \mathcal{V}^2(\mathbf{X}, \mathbf{Y})$.

The above properties reveal that the BDC is non-negative and able to capture the dependency of the signals well under the translations and orthonormal transformations. Those properties exhibit surprising consistency with human texture perception, which is usually not sensitive to texture resampling and geometric transformations. Incorporating the BDC in our method endows our method with a more powerful quality prediction capability and its effectiveness can be verified by the extensive experiments in Sec. 4.

3.2 BDC BASED FR-IQA MODEL

Given the reference image $\mathbf{x} \in \mathbb{R}^{\mathbf{3}}$ and the test image $\mathbf{y} \in \mathbb{R}^{\mathbf{3}}$, the aim of FR-IQA is to predict the image quality $\mathbf{q} \in \mathbb{R}^{\mathbf{1}}$ of $\mathbf{y}$. Herein, directly calculating the dependency between $\mathbf{x}$ and $\mathbf{y}$ in the pixel space may not be adequate, as human perceptual sensitivity is usually non-uniform Wang & Simoncelli (2008); Berardino et al. (2017). Recently, deep neural networks have shown a surprising power in capturing image distortions, popularly adopted as the quality-aware representation generator Zhang et al. (2018); Prashnani et al. (2018); Ding et al. (2020). Following the vein, we first adopt the VGG16 network Simonyan & Zisserman (2015) to nonlinearly transform the images ($\mathbf{x}$ and $\mathbf{y}$) to the deep representations. Then, the BDC is calculated to evaluate the dependency between the representations of $\mathbf{x}$ and $\mathbf{y}$, deemed that the higher dependency corresponds to the higher quality. The design details are shown in Fig. 2. In particular, the VGG16 network contains five stages in total and is pre-trained on the ImageNet Deng et al. (2009) dataset. We empirically abandon the layers after the fourth stage and use the rest layers as the deep-feature extractor. Supposing the extractor is represented by $\phi(\cdot)$, the deep-features of $\mathbf{x}$ and $\mathbf{y}$ can be obtained by,

$$\mathbf{X} = \phi(\mathbf{x}), \mathbf{Y} = \phi(\mathbf{y}), \tag{10}$$

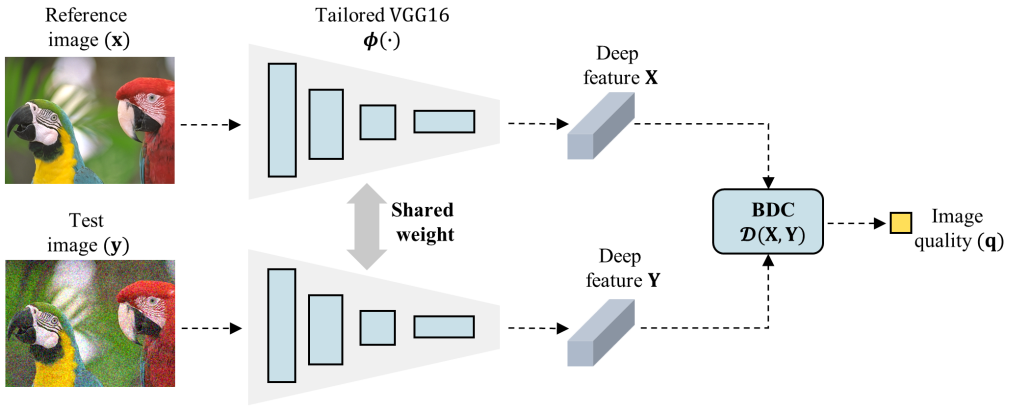


Figure 2: Overall structure of DID. The tailored VGG16 (only the first four stages are reserved) network is adopted as the deep feature extractor. Then the features dependency is measured by the BDC and higher dependency indicates better quality.

Algorithm 1: DID based FR-IQA model.

Input: Reference image $\mathbf{x}$; test image $\mathbf{y}$.
Output: The quality of test image $\mathbf{q}$.

Step 1. Extract the deep representations $\mathbf{X}, \mathbf{Y}$ of $\mathbf{x}, \mathbf{y}$ by the tailored VGG16;
Step 2. Obtain the matrices $\mathbf{A}_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$ and $\mathbf{B}_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ by Eq. (9) and Eq. (12).
Step 3. Select the upper triangular portions of $\mathbf{A}$ and $\mathbf{B}$ as $\mathbf{A_u}$ and $\mathbf{B_u}$.
Step 4. Estimate the test image quality $\mathbf{q} = \mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} \mathbf{A}_{kl}\mathbf{B}_{kl}$ by Eq. (11).

where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{h\times w\times d}$, h and w are the spatial dimensions and d is the channel number. We reshape the $\mathbf{X}$ and $\mathbf{Y}$ into $hw \times d$, *i.e.*, $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{hw\times d}$, then treat each column vector of $\mathbf{X}$ and $\mathbf{Y}$ (denoted as $X \in \mathbb{R}^{hw}$, $Y \in \mathbb{R}^{hw}$) as the observations of random vectors sampled from the marginal distributions of f_X and f_Y, respectively. However, the i.i.d. assumption may not hold for X and Y due to distinct semantic information lying in different channels. As such, instead of directly using the inner product in Eq. (8), suggested by Xie et al. (2022), the cosine similarity is adopted for dependency estimation,

$$\mathbf{q} = \mathcal{D}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{A_u}\mathbf{B_u^T}}{\|\mathbf{A_u}\|_2 \|\mathbf{B_u}\|_2}, \tag{11}$$

where $\mathbf{A_u}$ and $\mathbf{B_u}$ are the flattened results of the upper triangular portions of $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{A}_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$ and $\mathbf{B}_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$. For a_{kl} and b_{kl}, we adopt $z = 2$ in Eq. (9),

$$a_{kl} = \|X_k - X_l\|_2, b_{kl} = \|Y_k - Y_l\|_2. \tag{12}$$

Algorithm 1 summarizes the framework of our method. The $\mathcal{D}(\mathbf{X}, \mathbf{Y})$ depicts the dependency level of $\mathbf{Y}$ with $\mathbf{X}$, ranging from (-1,1). Higher $\mathcal{D}(\mathbf{X}, \mathbf{Y})$ corresponds to better quality of test image. It should be noted that although our DID-based IQA model is a deep learning-based quality measure, it enjoys the learning-free advantage, avoiding over-fitting to a specific dataset.

4 Experiment

In this section, we first describe the experimental setup. Next, we conduct comprehensive experiments to verify the effectiveness of the proposed model, including image quality prediction, texture quality assessment, and invariance of geometric transformation. Finally, the ablation studies are performed.

Method	LIVE		CSIQ		TID2013		KADID-10k		PIPAL	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR	0.873	0.868	0.809	0.815	0.688	0.679	0.676	0.680	0.407	0.415
SSIM	0.931	0.928	0.872	0.868	0.720	0.745	0.724	0.723	0.498	0.505
MS-SSIM	0.931	0.931	0.908	0.896	0.798	0.810	0.802	0.801	0.552	0.590
VIF	0.927	0.925	0.902	0.887	0.690	0.732	0.593	0.602	0.443	0.468
FSIM	**0.965**	**0.961**	0.931	0.919	0.851	0.877	0.854	0.851	0.589	0.615
NLPD	0.914	0.914	0.917	0.911	0.808	0.823	0.810	0.810	0.469	0.509
Style	0.898	0.882	0.853	0.837	0.675	0.681	0.701	0.707	0.339	0.337
PieAPP	0.908	0.919	0.877	0.892	0.850	0.848	0.836	0.836	**0.700**	**0.712**
LPIPS	0.939	0.945	0.883	0.906	0.695	0.759	0.720	0.729	0.573	0.618
DISTS	**0.955**	**0.954**	0.939	0.941	0.848	0.870	**0.890**	**0.889**	0.624	0.644
DSD	0.577	0.552	0.603	0.700	0.548	0.657	0.439	0.527	0.274	0.350
DeepWSD	0.896	0.890	**0.963**	**0.953**	**0.874**	**0.896**	0.888	0.888	0.514	0.517
DID	0.948	0.940	**0.945**	**0.944**	**0.858**	**0.879**	**0.905**	**0.904**	**0.677**	**0.697**

Table 1: Performance comparison of DID against twelve existing FR-IQA models on five standard IQA datasets. The best two results are highlighted in boldface.

4.1 Experimental Setups

The VGG16 is tailored at "ReLU4_3" layer and pre-trained on the ImageNet Deng et al. (2009). Inspired by the SSIM Wang et al. (2004) and DISTS Ding et al. (2020), we resize the shorter side of the input images to 224 while keeping the aspect ratio. We apply the Spearman's rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) to evaluate the monotonicity and linearity. The larger SRCC and PLCC values reflect better quality prediction results. In particular, a five-parameter nonlinear logistic function is fitted to map the predicted scores to the same scale as MOSs when computing PLCC VQEG (2000).

4.2 Performance on Image Quality Prediction

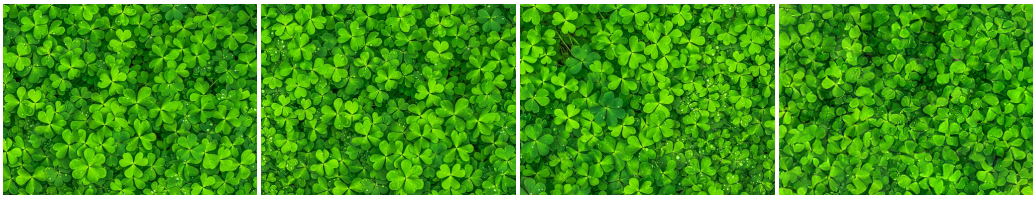
We compare DID with 12 FR-IQA models on five standard IQA datasets, including the LIVE Sheikh et al. (2006) CSIQ Larson & Chandler (2010), TID2013 Ponomarenko et al. (2015), KADID-10k Lin et al. (2019), and PIPAL Jinjin et al. (2020). In particular, LIVE, CSIQ, and TID2013 contain limited image contents and distortion types, and they have been widely used for more than ten years. The KADID-10k and PIPAL are two large-scale IQA datasets with more than ten thousand distorted images. KADID-10k has 81 pristine images, and 25 distortion types with 5 levels are adopted to generate $10,125$ distorted images. PIPAL is so far the largest human-rated IQA dataset with $23,200$ images, which are generated by 200 reference images with 40 distortions types. It is worth noting that PIPAL introduces 19 GAN-based distortions, challenging the existing FR-IQA a lot. In addition, the 12 FR-IQA models cover various design methodologies: the error visibility methods - PSNR, and NLPD Laparra et al. (2016), the structural similarity methods - SSIM Wang et al. (2004), MS-SSIM Wang et al. (2003), and FSIM Zhang et al. (2011); the information-theoretic methods - MAD Larson & Chandler (2010), and VIF Sheikh & Bovik (2006); the learning-based methods - PieAPP Prashnani et al. (2018), LPIPS Zhang et al. (2018), and DISTS Ding et al. (2020); the distribution-based methods - Style Gatys et al. (2016), DSD Kligvasser et al. (2021), DeepWSD Liao et al. (2022). The experimental results are reported in the Table 1, from which we can find DID achieves superior performance on both classical (LIVE, CSIQ, and TID2013) and latest (KADID-10k and PIPAL) IQA datasets. It demonstrates that the dependency-based model is well correlated with human ratings, and the learning-free advantage equips the DID with strong generalization capability. In addition, the knowledge-driven methods (*e.g.*, FSIM) generally perform better on small-scale IQA datasets, indicating the potential over-fitting problem because of the extensive parameter tuning. Moreover, DeepWSD outperforms most learning-based methods on synthetic distortions, which further reflects the success of the joint distribution based FR-IQA model. Finally, though PieAPP obtains the best performance on the PIPAL dataset, it requires plenty of the human-rated images to train the model Prashnani et al. (2018).

Method	PIPAL (GAN distortion)	
	SRCC	PLCC
SSIM	0.322	0.472
MS-SSIM	0.387	0.615
VIF	0.324	0.543
FSIM	0.410	0.621
NLPD	0.341	0.570
LPIPS	0.486	0.617
PieAPP	**0.553**	**0.632**
DISTS	0.549	0.607
DeepWSD	0.397	0.560
FID	0.413	0.496
DID	**0.5742**	**0.6403**

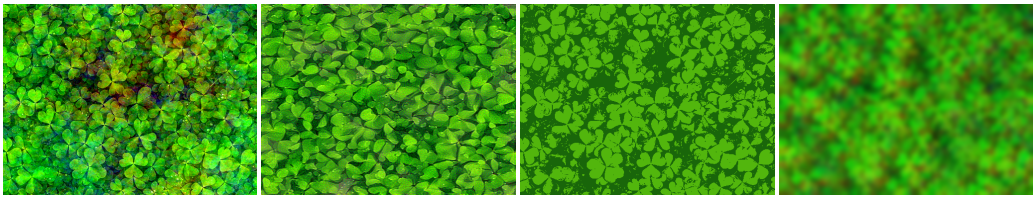
Table 2: Performance comparison of DID against state-of-the-art methods on the GAN distortion of PIPAL dataset. The measure specifically designed for GAN images is represented in italics.

Method	SynTEX		TQD	
	SRCC	PLCC	SRCC	PLCC
SSIM	0.579	0.598	0.352	0.418
VIF	0.606	0.697	0.549	0.614
FSIM	0.081	0.115	0.386	0.272
NLPD	0.606	0.607	0.409	0.457
LPIPS	0.788	0.788	0.203	0.188
PieAPP	0.715	0.719	0.718	0.721
DISTS	**0.923**	**0.901**	**0.910**	**0.903**
$DISTS_s$	0.877	0.868	0.795	0.780
STSIM	0.643	0.650	0.408	0.422
NPTSM	0.496	0.505	0.679	0.678
ISGTQA	0.820	0.816	0.802	0.804
DID	**0.896**	**0.874**	**0.889**	**0.917**

Table 3: Performance comparison of DID against state-of-the-art methods on two texture quality datasets. The texture similarity models are represented in italics.



(a) MOS ↑ / DID ↑ (b) 0.873 / 0.991 (c) 0.727 / 0.980 (d) 0.693 / 0.931



(e) 0.528 / 0.821 (f) 0.481 / 0.758 (g) 0.437 / 0.506 (h) 0.401 / 0.210

Figure 3: Texture images sampled from TQD Ding et al. (2020). (a) Reference image. (b) Resampling image. (c) Resampling image. (d) Texture synthesis method Snelgrove (2017). (e) Pink noise. (f) Texture synthesis method Gatys et al. (2015). (g) Color quantization. (h) Chromatic aberration.

We further compare DID with the FR-IQA models on the GAN-generated images of the PIPAL dataset. As shown in Table 2, most distance-based DR-IQA models present poor performance, and the underlying reason may lie in that synthesized textures which not appear in reference images are usually introduced by the generation networks. Although the Fréchet Inception Distance (FID) Heusel et al. (2017) is designed especially for the quality evaluation of GAN models, the poor performance reveals its limitations and the great challenges of GAN-generated IQA. Compared with those methods, our model achieves the best result, the dependency rather than the distance mitigates the strict requirement of point-by-point alignment during feature comparison.

4.3 Performance on Texture Similarity

To verify the effectiveness of our method on texture quality prediction, we conduct experiments on SynTEX Golestaneh et al. (2015) and TQD Ding et al. (2020) datasets. In particular, SynTEX

Method	Translation		Rotation		Scaling		Mixed		Overall	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR	0.104	0.365	0.088	0.365	0.088	0.366	0.093	0.373	0.102	0.320
SSIM	0.194	0.388	0.199	0.390	0.196	0.394	0.207	0.393	0.211	0.232
MS-SSIM	0.183	0.381	0.191	0.373	0.202	0.387	0.213	0.389	0.191	0.202
VIF	0.239	0.417	0.224	0.409	0.209	0.402	0.214	0.399	0.369	0.373
FSIM	0.370	0.558	0.382	0.575	0.376	0.566	0.393	0.590	0.336	0.522
NLPD	0.153	0.180	0.170	0.196	0.189	0.223	0.172	0.202	0.265	0.324
Style	0.740	0.758	0.735	0.747	0.729	0.738	0.740	0.752	0.744	0.758
PieAPP	0.805	0.811	0.814	0.821	0.812	0.818	0.820	0.825	0.815	0.819
LPIPS	0.781	0.788	0.806	0.808	0.811	0.822	0.843	0.853	0.746	0.762
DISTS	**0.885**	**0.886**	**0.887**	**0.887**	**0.881**	**0.882**	**0.905**	**0.905**	**0.890**	**0.890**
DSD	0.497	0.670	0.500	0.682	0.501	0.686	0.487	0.688	0.496	0.638
DeepWSD	0.163	0.189	0.187	0.207	0.215	0.194	0.280	0.302	0.234	0.288
DID	**0.899**	**0.907**	**0.899**	**0.905**	**0.891**	**0.898**	**0.913**	**0.918**	**0.903**	**0.909**

Table 4: Performance comparison of DID against the state-of-the-art FR-IQA models on LIVE-GT dataset. The best two results are highlighted in boldface.

consists of 105 synthesized texture images, which were generated by five texture synthesis methods for 21 high-quality texture images. TQD contains ten reference texture images, and each of them is degraded by 15 distortion types, including seven synthetic distortions, four texture synthesis methods, and four randomly resampling versions. We show several sampled texture images in Fig. 1. For performance comparison, four representative knowledge-driven methods: SSIM Wang et al. (2004), VIF Sheikh & Bovik (2006), FSIM Zhang et al. (2011), and NLPD Laparra et al. (2016) and three data-driven methods: LPIPS Zhang et al. (2018), PieAPP Prashnani et al. (2018), and DISTS Ding et al. (2020) are selected. Furthermore, three models: STSIM Zujovic et al. (2013), NPTSM Alfarraj et al. (2016), and IGSTOA Golestaneh & Karam (2018), that designed especially for texture similarity are also included.

The SRCC and PLCC results are listed in Table 3. It is not surprising that DISTS achieves the best performance on two texture quality datasets since a great number of texture images were used to train the weights for measuring texture similarity. However, when the training session for the texture similarity term is absent (denoted as DISTS_s in Table 3), DID outperforms DISTS_s by a significant margin. Besides, ISGTQA exhibits noteworthy improvements for the texture similarity models but still falls behind DID. Furthermore, as shown in Fig. 3, we rank the sample texture images based on the DID score and MOS. We can observe that DID is consistent with human perception of texture quality. In particular, the images with visible artifacts have lower quality scores while the resampling images correspond to higher quality scores. Thus, we may draw the conclusion that the proposed dependency-based model provides a promising texture perception in the scenarios of texture synthesis, resampling, and transformation.

4.4 Performance on Geometric Transformations

Texture image quality usually presents an invariance to mild geometric transformations Ding et al. (2020). To study the performance of existing FR-IQA models on such a prior, we construct a new IQA dataset denoted as LIVE-GT with four geometric transformations (translation, rotation, scaling, and their mixed) involved. In particular, the four geometric transformations are implemented by randomly shifting 5% pixels in vertical or horizontal directions, randomly rotating 3° in clockwise or anticlockwise directions, scaling the image by a factor of 1.05, and mixing the above-mentioned transformations. We impose the four transformations to the reference images in the LIVE dataset and a total of $3,895$ distorted images are finally generated. Herein, we make a mild assumption that the modest geometric transformations will not change the MOS of each image. Then, the performance comparison on the LIVE-GT dataset can be conducted.

We list the SRCC and PLCC results for overall and individual geometric transformation types in Table 4. We can observe that the performance of knowledge-driven models drops dramatically on the LIVE-GT dataset while our DID model achieves the best performance on all four transformations. We

Backbone	Layer	Quality prediction				Texture similarity		Geo invariance
		LIVE	TID2013	KADID-10k	PIPAL	SynTEX	TQD	LIVE-GT
VGG16	ReLU1_2	0.825	0.637	0.709	0.573	0.699	0.733	0.679
	ReLU2_2	0.892	0.771	0.854	0.601	0.837	0.806	0.774
	ReLU3_3	0.938	0.845	**0.899**	0.652	0.869	0.860	0.892
	ReLU4_3	0.948	**0.858**	**0.905**	**0.677**	0.896	0.889	0.903
	ReLU5_3	0.946	0.855	0.881	0.656	0.893	0.867	0.893
ResNet50	Layer_3	**0.953**	**0.865**	0.877	0.612	0.903	0.897	0.889
	Layer_4	**0.950**	0.857	0.875	0.640	**0.925**	0.898	**0.905**
DenseNet121	Block_3	0.937	0.856	0.850	**0.658**	0.916	**0.920**	**0.919**
	Block_4	0.949	**0.865**	0.881	0.654	**0.918**	**0.921**	0.856

Table 5: Ablation experiments of DID with different CNN backbones and tailored at different layers in terms of SRCC results. Geo invariance is the abbreviation of geometric invariance. The best two results are highlighted in boldface, and the default setting of DID is highlighted with an underline.

believe the invariance property is brought by the dependency-based measure which avoids measuring the feature difference in a deterministic way. In addition, it is worth noting that the DeepWSD is also a joint distribution-based FR-IQA. However, it is vulnerable to transformations as the feature statistic comparison is performed locally. On the contrary, we construct the feature joint distribution in a global manner with both spatial and channel dimensions involved. In summary, we can conclude that the dependency derived from the global statistic of features contributes to the robustness of our model.

4.5 Ablation studies

In this subsection, we conduct ablation experiments to investigate the effect of the pre-trained CNN backbone and the tailored layer in the proposed model. Except for the default VGG16 Simonyan & Zisserman (2015) backbone, we applied the proposed BDC based FR-IQA method to other two widely used ImageNet pre-trained image classification networks - ResNet50 He et al. (2016) and DenseNet121 Huang et al. (2017). We tailored VGG16 at the last ReLU nonlinearity layer of each stage. We take the last two stages of ResNet50 (denoted as Layer_3 and Layer_4, respectively) and DenseNet121 (denote as Block_3 and Block_4, respectively) as the comparison variants. The results of three kinds of quality assessment tasks, *i.e.*, quality prediction, texture similarity, and geometric invariance, are shown in Table 5, from which we can have the following observations. First, the proposed dependency model performs better in the deeper layers, as the deeper layers are able to capture the semantic information. Second, the proposed BDC based model is quite robust to the CNN backbones according to the superior performance. Last, while the ResNet50 and DenseNe121 outperform the VGG16 in some small-scale datasets, we still choose VGG16 as the default backbone due to the satisfactory trade-off between model complexity and performance.

5 Conclusion

In this paper, we have presented the new design philosophy for FR-IQA method and shown that the feature-dependency is particularly effective for generative images. The paradigm shift brings a fresh new perspective regarding how image quality shall be alternatively defined, given the available reference image. We obtain the conclusion that instead of gauging the feature distance, the dependency which is characterized by BDC could well reflect the image quality. In addition, the proposed measure can be treated as a plug-and-play module and incorporated into different backbones, dynamically adapting the application scenarios. We also believe the paradigm shifted from distance to dependency will shed light on more generalized quality measures and inspire more works on the exploration of feature dependency.

REFERENCES

Motaz Alfarraj, Yazeed Alaudah, and Ghassan AlRegib. Content-adaptive non-parametric texture similarity measure. In *IEEE International Workshop on Multimedia Signal Processing*, pp. 1–6, 2016.

Shahrukh Athar and Zhou Wang. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access*, 7:140030–140070, 2019.

Alexander Berardino, Valero Laparra, Johannes Ballé, and Eero Simoncelli. Eigen-distortions of hierarchical representations. In *Neural Information Processing Systems*, pp. 3530–3539, 2017.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2017.

Yue Cao, Zhaolin Wan, Dongwei Ren, Zifei Yan, and Wangmeng Zuo. Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5851–5861, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. Locally adaptive structure and texture similarity for image quality assessment. *ACM International Conference on Multimedia*, pp. 2483–2491, 2021.

Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. Quantifying visual image quality: A bayesian view. *Annual Review of Vision Science*, 7(1):437–464, 2021.

Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Neural Information Processing Systems*, pp. 1–8, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

Alireza Golestaneh and Lina J Karam. Synthesized texture quality assessment via multi-scale spatial and statistical texture attributes of image and gradient magnitude coefficients. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 738–744, 2018.

S Alireza Golestaneh, Mahesh M Subedar, and Lina J Karam. The effect of texture granularity on texture synthesis quality. *Applications of Digital Image Processing XXXVIII*, 9599:356–361, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, pp. 1–9, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pp. 633–651, 2020.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711, 2016.

Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1676–1684, 2017.

Idan Kligvasser, Tamar Shaham, Yuval Bahat, and Tomer Michaeli. Deep self-dissimilarities as powerful visual fingerprints. In *Neural Information Processing Systems*, pp. 3939–3951, 2021.

Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016(16):1–6, 2016.

Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):1–21, 2010.

Xigran Liao, Baoliang Chen, Hanwei Zhu, Shiqi Wang, Mingliang Zhou, and Sam Kwong. DeepWSD: Projecting degradations in perceptual space to wasserstein distance in deep feature space. In *ACM International Conference on Multimedia*, 2022.

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

Weisi Lin, Dong Li, and Ping Xue. Discriminative analysis of pixel difference towards picture quality prediction. In *International Conference on Image Processing*, pp. 193–196, 2003.

Pablo Navarrete Michelini, Dan Zhu, and Hanwen Liu. Multi–scale recursive and perception–distortion controllable image super–resolution. In *European Conference on Computer Vision workshops*, pp. 1–14, 2018.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12): 2117–2128, 2005.

Hamid R Sheikh, Zhou Wang, Lawrence Cormack, and Alan C Bovik. Image and video quality assessment research at LIVE. 2006. URL `https://live.ece.utexas.edu/research/Quality/subjective.htm`.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pp. 1–6, 2015.

Xavier Snelgrove. High-resolution multi-scale neural texture synthesis. In *SIGGRAPH Asia Technical Briefs*, pp. 1–4, 2017.

Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, 2000. URL `http://www.vqeg.org`.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2010.

Zhou Wang and Eero P Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 573–576, 2005.

Zhou Wang and Eero P Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, 2008.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, pp. 1398–1402, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7972–7981, 2022.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Jana Zujovic, Thrasyvoulos N Pappas, and David L Neuhoff. Structural texture similarity metrics for image analysis and retrieval. *IEEE Transactions on Image Processing*, 22(7):2545–2558, 2013.