Image Quality Assessment: Unifying Structure and Texture Similarity

Keyan Ding, Kede Ma, *Member, IEEE,* Shiqi Wang, *Member, IEEE,* and Eero P. Simoncelli, *Fellow, IEEE*

Abstract—Objective measures of image quality generally operate by comparing pixels of a "degraded" image to those of the original. Relative to human observers, these measures are overly sensitive to resampling of texture regions (*e.g.*, replacing one patch of grass with another). Here, we develop the first full-reference image quality model with explicit tolerance to texture resampling. Using a convolutional neural network, we construct an injective and differentiable function that transforms images to multi-scale overcomplete representations. We demonstrate empirically that the spatial averages of the feature maps in this representation capture texture appearance, in that they provide a set of sufficient statistical constraints to synthesize a wide variety of texture patterns. We then describe an image quality method that combines correlations of these spatial averages ("texture similarity") with correlations of the feature maps ("structure similarity"). The parameters of the proposed measure are jointly optimized to match human ratings of image quality, while minimizing the reported distances between subimages cropped from the same texture images. Experiments show that the optimized method explains human perceptual scores, both on conventional image quality databases, as well as on texture databases. The measure also offers competitive performance on related tasks such as texture classification and retrieval. Finally, we show that our method is relatively insensitive to geometric transformations (*e.g.*, translation and dilation), without use of any specialized training or data augmentation. Code is available at https://github.com/dingkeyan93/DISTS.

Index Terms—Image quality assessment, structure similarity, texture similarity, perceptual optimization.

IMAGE quality assessment (IQA) – the quantification of human perception of image quality – is a fundamental problem in both human and computational vision, and is of paramount importance in a variety of real-world applications, such as image restoration, compression, and rendering. For more than 50 years, the mean squared error (MSE) was the standard full-reference method for assessing signal fidelity and quality, and it continues to play a fundamental role in the development of signal and image processing algorithms, despite its poor correlation with human perception [1], [2].

A variety of proposed full-reference IQA methods provide a better account of human perception than MSE [3]–[8], and the Structural Similarity (SSIM) index [3] has become a *de facto* standard in the field of image processing. But these methods rely on alignment of the images being compared, and are thus highly sensitive to differences between images of the same texture (e.g., two different cropped regions of the same bed of pebbles). Two samples of the same texture differ substantially in the precise arrangement of their features, while appearing nearly the same to a human observer (see Fig. 1). Since textured surfaces are ubiquitous in photographic images, it is important to develop objective IQA metrics that are consistent with this aspect of perceptual similarity. Such a metric would allow the development of a new generation of image processing solutions - for example, a compression engine that statistically synthesizes texture regions rather than trying to exactly re-create the pixels of the original image [9], [10].

We present the first full-reference IQA method that is insensitive to resampling of visual textures. Our method is constructed by first nonlinearly transforming images to a multi-scale overcomplete representation, using a variant of the VGG convolutional neural network (CNN) [14]. We show that the spatial averages of the feature maps provide a compact set of statistical constraints that is sufficient to capture the visual appearance of textures [15]. Specifically, we use the test originally proposed by Julesz [16], and demonstrate that synthesizing a new image by forcing it to match the channel averages computed from a given texture image results in an image of similar visual appearance. Although the number of statistics in the set is substantially smaller than that of pixels in the image, we find that the result holds for a wide variety of textures, regardless of the initialization, thus revealing the robustness of this model to adversarial examples [17].

After transforming the original and corrupted images, we construct our measure by combining two terms over all feature maps: one that compares the spatial averages (and thus, the texture properties) of the two images, and a second that compares the structural details. The final distortion score is computed as a weighted sum of these two terms, with the weights adjusted to match human perception of image quality and invariance to resampled texture patches. The first is achieved by comparing the responses of the model with a database of human image quality ratings. The second is achieved by minimizing the distance between pairs of patches sampled from the same texture images. We

- *Keyan Ding, Kede Ma, and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: keyan.ding@my.cityu.edu.hk, kede.ma@cityu.edu.hk, shiqwang@cityu.edu.hk).*
- *Eero P. Simoncelli is with the Flatiron Institute of the Simons Foundation, and the Center for Neural Science and the Courant Institute of Mathematical Sciences, New York University, New York, NY 10003, USA (e-mail: eero.simoncelli@nyu.edu).*

(a) (b) (c)

Fig. 1. Existing full-reference IQA models are overly sensitive to point-by-point deviations between images of the same texture. **(a)** A grass image and **(b)** the same image, distorted by JPEG compression. **(c)** Resampling of the same grass as in (a). Popular IQA measures, including PSNR, SSIM [3], FSIM [11], VIF [4], GMSD [12], DeepIQA [13], PieAPP [8], and LPIPS [7], predict that image (b) has a better perceived quality than image (c), which is in disagreement with human rating. In contrast, the proposed DISTS model makes the correct prediction. (Zoom in to improve visibility of details).

show that the resulting Deep Image Structure and Texture Similarity (DISTS) index can be transformed into a proper metric in the mathematical sense. Moreover, DISTS correlates well with human quality judgments in several independent datasets, and achieves a high degree of invariance to texture substitution. We also demonstrate competitive performance of DISTS on tasks of texture classification and retrieval. Last, we show that DISTS is insensitive to mild local and global geometric distortions [18], [19], which may be imperceptible to the human visual system (HVS).

1 Background

Pioneering work on perceptual full-reference IQA dated back to the 1970s, when Mannos and Sakrison [20] investigated a class of visual fidelity measures in the context of rate-distortion optimization. A number of alternative models were subsequently proposed [21], [22], each mimicking certain functionalities of the HVS and penalizing the errors between the reference and distorted images "perceptually". However, the HVS is a complex and highly nonlinear system [23], and most IQA measures within the error visibility framework rely on strong assumptions and simplifications (*e.g.*, linear or quasi-linear models for early vision characterized by restricted visual stimuli), and exhibit shortcomings regarding the definition of visual quality, quantification of suprathreshold distortions, and generalization to natural images [24]. The SSIM index [3] introduced the concept of comparing structure similarity (instead of measuring error visibility), opening the door to a new class of full-reference IQA measures [11], [12], [18], [25]. Other design methodologies for knowledge-driven IQA include information-theoretic criterion [4] and perception-based pooling [26]. Recently, there has been a surge of interest in leveraging advances in large-scale optimization to develop data-driven IQA measures [7], [8], [13], [19]. However, databases of human quality scores are often insufficiently rich to constrain the large number of model parameters. As a result, these learned methods are at risk of over-fitting [27].

Nearly all knowledge-driven full-reference IQA models base their quality measurements on point-by-point comparisons between pixels or convolution responses (*e.g.*, wavelets). As such, they are not capable of handling "visual textures", which are loosely defined as spatially homogeneous regions with repeated elements, often subject to some randomization in their location, size, color, and orientation [15]. Different images of the same texture can look nearly the same to the human eye, while differing substantially at the level of pixel intensities. Research on visual texture has a long history, and can be partitioned into four problems: texture classification, texture segmentation, texture synthesis, and shape from texture. At the core of texture analysis is an efficient description (*i.e.*, representation) that matches human perception of visual textures. In this paper, we aim to measure perceptual texture similarity, a goal first elucidated and explored in [28], [29].

The response amplitudes and variances of computational texture features (*e.g.*, Gabor basis functions [30], local binary patterns [31]) have achieved good performance for texture classification, but are not well correlated with human perceptual ratings of texture similarity [28], [29]. Texture representations that incorporate more sophisticated statistical features, such as correlations of complex wavelet coefficients [15], have shown significantly more power for texture synthesis, suggesting that they may provide a good substrate for similarity measures. In recent years, the use of such statistics extracted from CNN-based representations [32]–[34] has led to even richer texture description.

2 The DISTS Index

Our goal is to develop a new full-reference IQA model that combines sensitivity to structural distortions (*e.g.*, artifacts due to noise, blur, or compression) with a tolerance of texture resampling (exchanging the content of a texture region with a new sample of the same texture). As is common in many IQA methods, we first transform the reference and distorted images to a new representation, using a CNN. Within this representation, we develop a set of measurements that are sufficient to capture the appearance of a variety of different visual textures. Finally, we combine these texture parameters with global structural measurements to form an IQA measure.

2.1 Initial Transformation

Our model is built on an initial transformation, $f : \mathbb{R}^n \mapsto \mathbb{R}^r$, that maps the reference and distorted images (x and

y, respectively) to "perceptual" representations ($\tilde{x}$ and $\tilde{y}$, respectively). The primary motivation is that perceptual distances are non-uniform in the pixel space [35], [36], and this is the main reason that MSE is inadequate as a perceptual IQA model. The purpose of function f is to transform the pixel representation to a space that is more perceptually uniform. Previous IQA methods have used filter banks to capture the frequency-dependence of error visibility [5], [21]. Others have used transformations that mimic the early visual system [22], [37]–[39]. More recently, deep CNNs have shown surprising power in representing perceptual image distortions [7], [8], [13]. In particular, Zhang *et al*. [7] have demonstrated that pre-trained deep features from VGG can be used as a substrate for quantifying perceptual quality.

As such, we also chose to base our model on the VGG16 CNN [14], pre-trained for object recognition [40] on the ImageNet database [41]. The VGG transformation is constructed by a feedforward cascade of layers, each including spatial convolution, halfwave rectification, and downsampling. All operations are continuous and differentiable, both advantageous for an IQA method that is to be used in optimizing image processing systems. We modified the VGG architecture to achieve two additional desired properties. First, in order to provide a good substrate for the invariances needed for texture resampling, we wanted the initial transformation to be *aliasing-free*. The "max pooling" operation of the original VGG architecture has been shown to introduce visible aliasing artifacts when used to interpolate between images with geodesic sequences [42]. To avoid aliasing when subsampling by a factor of two, the Nyquist theorem requires blurring with a filter whose cutoff frequency is below $\frac{\pi}{2}$ radians/sample [43]. Following this principle, we replaced all max pooling layers in VGG with weighted ℓ_2 pooling [42]:

$$P(x) = \sqrt{g * (x \odot x)}, \tag{1}$$

where $\odot$ denotes pointwise product, and the blurring kernel $g(\cdot)$ was implemented by a Hanning window that approximately enforces the Nyquist criterion with a stride of 2. As additional motivation, we note that ℓ_2 pooling has been used to describe the behavior of complex cells in primary visual cortex [44], and is also closely related to the complex modulus used in the scattering transform [45].

A second desired property for our transformation is that it should be *injective*: distinct inputs should map to distinct outputs. This is necessary to ensure that the final quality measure is a proper metric (in the mathematical sense) - if the representation of an image is non-unique, then equality of the output representations will not imply equality of the input images. This property has proven useful in perceptual optimization, although it is not present in many recent methods. For example, the mapping function in GMSD [12] extracts image gradients, discarding local luminance information that is essential to human perception of image quality. Similarly, GTI-CNN [19], makes deliberate use of a surjective transformation, in an attempt to achieve invariance to mild geometric transformations, but throws away a substantial amount of structural information that is perceptually important.

Considerable effort has been made in developing invertible CNN-based transformations in the context of density modeling [46]–[49]. These methods place strict constraints on either network architectures [46], [48] or network parameters [49], which limit the expressiveness in learning quality-relevant representations. Ma *et al*. [50] proved that under Gaussian-distributed random weights and ReLU nonlinearity, a two-layer CNN is injective provided that it is sufficiently expansive (*i.e*., the output dimension of each layer should increase by at least a logarithmic factor). Although mathematically appealing, this result does not constrain parameter settings of CNNs of more than two layers. In addition, a Gaussian-weighted CNN is less likely to be perceptually relevant [19], [32].

Like most CNNs, VGG discards information at each stage of transformation. To ensure an injective mapping, we simply included the input image as an additional feature map (the "zeroth" layer of the network). The representation then consists of the input image x, concatenated with the convolution responses of five VGG layers (labelled conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3):

$$f(x) = \{\tilde{x}_j^{(i)}; i = 0, \ldots, m; j = 1, \ldots, n_i\}, \tag{2}$$

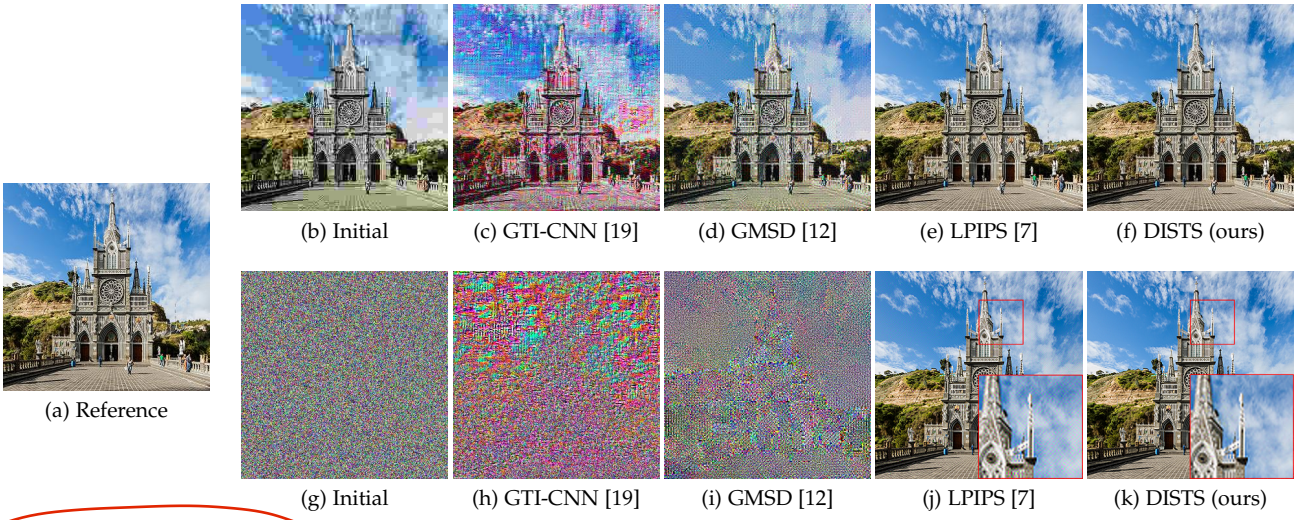
where $m = 5$ denotes the number of convolution layers chosen to construct f, n_i is the number of feature maps in the i-th convolution layer, and $\tilde{x}^{(0)} = x$. Similarly, we also computed the representation of the distorted image:

$$f(y) = \{\tilde{y}_j^{(i)}; i = 0, \ldots, m; j = 1, \ldots, n_i\}. \tag{3}$$

We used a naïve task – reference image recovery – to visually demonstrate the necessity of injective feature transformations. Specifically, given an original image x and an initial image y_0, we aim to recover x by numerically optimizing $y^\star = \arg\min_y D(x, y)$, where D denotes a full-reference IQA measure with a lower score indicating higher predicted quality, and $y^\star$ is the recovered image. For example, if D is the MSE, the (trivial) analytical solution is $y^\star = x$, indicating full recoverability. For the majority of existing IQA models, which are continuous and differentiable, solutions must be sought numerically using gradient-based iterative solvers. Fig. 2 shows the recovery results of our method from a JPEG-corrupted copy of the original image and a white Gaussian noise image, respectively, in comparison to three state-of-the-art models: GTI-CNN [19], GMSD [12], and LPIPS [7]. The first two, which are based on surjective mappings, fail dramatically on this simple task when initialized with purely white Gaussian noise. LPIPS, which is built on VGG but with no enforcement of the injective property, recovers most structures and details, but leaves some visible artifacts in the converged image (Fig. 2 (j)). In contrast, DISTS successfully recovers the reference image from any initialization.

2.2 Texture Representation

The visual appearance of textures is often characterized in terms of sets of local statistics [16] that are presumably measured by the HVS. Models consisting of various sets of features [15], [32], [51], [52] have been tested using synthesis: one generates an image with statistics that match those of a texture photograph. If the set of statistical measurements is



(a) Reference (b) Initial (c) GTI-CNN [19] (d) GMSD [12] (e) LPIPS [7] (f) DISTS (ours)

(g) Initial (h) GTI-CNN [19] (i) GMSD [12] (j) LPIPS [7] (k) DISTS (ours)

Fig. 2. Recovery of a reference image by optimization of IQA measures. Recovery is implemented by solving $y^\star = \arg\min_y D(x, y)$ with gradient descent, where D is an IQA distortion measure and x is a given reference image. **(a)** Reference image. **(b)** Corrupted initial image y_0, obtained by compressing the reference image using JPEG at a low bitrate. **(c)-(f)** Images recovered from (b) by optimizing different metrics (as indicated). **(g)** Corrupted initial image, obtained by adding white Gaussian noise. **(h)-(k)** Images recovered from (g) by optimizing indicated metrics. In all cases, the optimization converges, yielding a distortion score substantially lower than that of the initial.

a complete description of the appearance of the texture, then the synthesized image should be perceptually indistinguishable from the original [16], at least based on preattentive judgments [53].

Portilla and Simoncelli [15] found that the local correlations (and other pairwise statistics) of complex wavelet responses were sufficient to capture the visual appearance of a wide variety of textures, while at the same time being of low enough dimensionality ($\sim$ 700 dimensions). Gatys *et al*. [32] used correlations across channels of many layers in a VGG network, and were able to synthesize consistently better textures, albeit with a much larger set of statistics ($\sim$ 306K parameters). Since the number of statistics is typically larger than that of pixels in the input image, it is likely that this image was unique in matching these statistics. In this case, diversity in the synthesis results reflects local optima of the optimization procedure, rather than the entropy of the implicitly represented probability distribution. Ustyuzhaninov *et al*. [54] provided more direct evidence of this hypothesis: If the number of the statistical measurements is sufficiently large (on the order of millions), a single-layer CNN with random filters can always produce textures that are visually indiscernible to the human eye. Subsequent results suggest that a reduced set of statistics, containing only the mean and variance of CNN channels, is sufficient for texture classification or style transfer [55]–[57].

In our experiments, we found that an even more reduced set, containing only the spatial means of the feature maps (a total of $1,475$ statistics), provides an effective parametric model for visual textures. Specifically, we used this model to synthesize textures [15] by solving

$$y^\star = \arg\min_y D(x, y) = \arg\min_y \sum_{i,j} \left(\mu^{(i)}_{\tilde{x}_j} - \mu^{(i)}_{\tilde{y}_j}\right)^2, \quad (4)$$

where x is the target texture image, and $y^\star$ is the synthesized texture image, obtained by gradient descent optimization from a random initialization. $\mu^{(i)}_{\tilde{x}_j}$ and $\mu^{(i)}_{\tilde{y}_j}$ are the spatial averages of channels $\tilde{x}^{(i)}_j$ and $\tilde{y}^{(i)}_j$, respectively. Fig. 3 shows the synthesis results of our texture model using statistical constraints from individual and combined convolution layers of the pre-trained VGG. Similar to observations in Gatys *et al*. [32], we found that measurements from early layers appear to capture basic intensity and color information, and those from later layers summarize the shape and structure information. When matching statistics up to layer conv5_3, the synthesized texture appears visually similar to the reference.

Fig. 4 shows three synthesis results of our 1475-parameter texture model in comparison with the 710-parameter texture model of Portilla & Simoncelli [15] and the $\sim$ 306k-parameter model of Gatys *et al*. [32]. As one might expect, the visual quality of samples synthesized by our model lies between the other two.

2.3 Perceptual Distance Measure

Next, we specified quality measurements based on $f(x)$ and $f(y)$. Fig. 5 visualizes some feature maps of the six stages of the reference image "Buildings". As can been seen, spatial structures are present at all stages, indicating strong statistical dependencies between neighbouring coefficients. Therefore, use of an ℓ_p-norm, that assumes statistical independence of errors at different locations, is not appropriate. Inspired by the form of SSIM [3], we defined separate quality measurements for the texture (using the global means) and the structure (using the global correlations) of each pair of corresponding feature maps:

$$l(\tilde{x}^{(i)}_j, \tilde{y}^{(i)}_j) = \frac{2\mu^{(i)}_{\tilde{x}_j}\mu^{(i)}_{\tilde{y}_j} + c_1}{\left(\mu^{(i)}_{\tilde{x}_j}\right)^2 + \left(\mu^{(i)}_{\tilde{y}_j}\right)^2 + c_1}, \quad (5)$$

$$s(\tilde{x}^{(i)}_j, \tilde{y}^{(i)}_j) = \frac{2\sigma^{(i)}_{\tilde{x}_j\tilde{y}_j} + c_2}{\left(\sigma^{(i)}_{\tilde{x}_j}\right)^2 + \left(\sigma^{(i)}_{\tilde{y}_j}\right)^2 + c_2}, \quad (6)$$

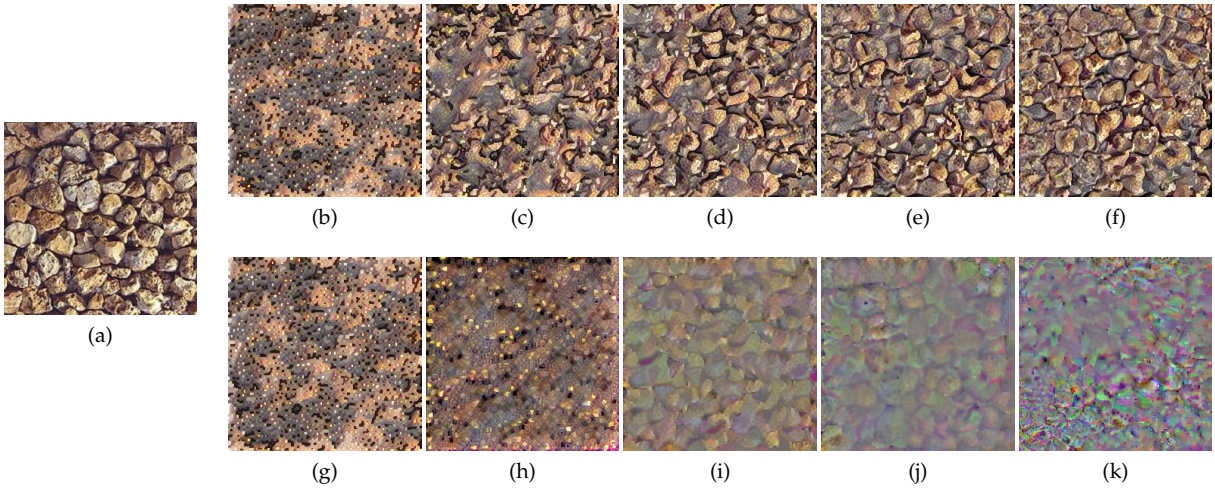


Fig. 3. Images synthesized to match the mean values of channels up to a given layer (top) or from individual layers (bottom) of the pre-trained VGG network. **(a)** Reference texture. **(b)** Up to conv1_2. **(c)** Up to conv2_2. **(d)** Up to conv3_3. **(e)** Up to conv4_3. **(f)** Up to conv5_3. **(g)** Only conv1_2. **(h)** Only conv2_2. **(i)** Only conv3_3. **(j)** Only conv4_3. **(k)** Only conv5_3.

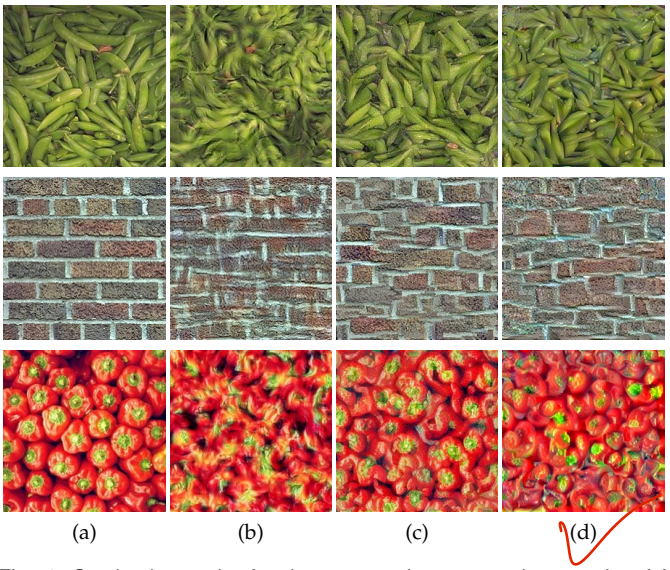


Fig. 4. Synthesis results for three example texture photographs. **(a)** Reference textures. **(b)** Images synthesized using the method of Portilla & Simoncelli [15]. **(c)** Images synthesized using Gatys *et al.* [32]. **(d)** Images synthesized using our texture model (Eq. (4)).

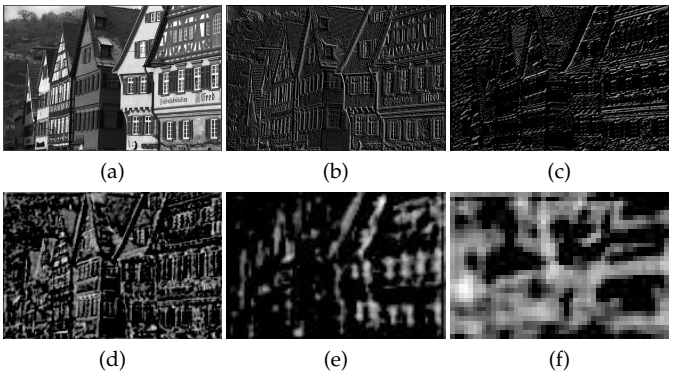


Fig. 5. Selected feature maps from the six layers of the VGG decomposition of the "buildings" image. **(a)** Zeroth stage (original image). **(b)** First stage. **(c)** Second stage. **(d)** Third stage. **(e)** Fourth stage. **(f)** Fifth stage. The feature map intensities are re-scaled for better visibility.

where $\mu_{\tilde{x}_j}^{(i)}$, $\mu_{\tilde{y}_j}^{(i)}$, $(\sigma_{\tilde{x}_j}^{(i)})^2$, $(\sigma_{\tilde{y}_j}^{(i)})^2$, and $\sigma_{\tilde{x}_j\tilde{y}_j}^{(i)}$ represent the global means and variances of $\tilde{x}_j^{(i)}$ and $\tilde{y}_j^{(i)}$, and the global covariance between $\tilde{x}_j^{(i)}$ and $\tilde{y}_j^{(i)}$, respectively. Two small positive constants, c_1 and c_2, are included to avoid numerical instability when the denominators are close to zero. The normalization mechanisms in Eq. (5) and Eq. (6) serve to equalize the magnitudes of feature maps at different stages.

Finally, the proposed DISTS model combines the quality measurements from different convolution layers using a weighted sum:

$$D(x,y;\alpha,\beta) = 1 - \sum_{i=0}^{m}\sum_{j=1}^{n_i}\left(\alpha_{ij}l(\tilde{x}_j^{(i)},\tilde{y}_j^{(i)}) + \beta_{ij}s(\tilde{x}_j^{(i)},\tilde{y}_j^{(i)})\right), \tag{7}$$

where $\{\alpha_{ij},\beta_{ij}\}$ are positive learnable weights, satisfying $\sum_{i=0}^{m}\sum_{j=1}^{n_i}(\alpha_{ij}+\beta_{ij}) = 1$. Note that the convolution kernels are fixed throughout the development of the method. Fig. 6 shows the full computation diagram of our quality assessment system.

Lemma 1. *For $\forall\ \tilde{x}_j^{(i)},\tilde{y}_j^{(i)} \in \mathbb{R}_+^n$ (as is the case for responses after ReLU nonlinearity), it can be shown that*

$$d(x,y) = \sqrt{D(x,y)} \tag{8}$$

is a proper metric, satisfying

- *non-negativity:* $d(x,y) \geq 0$;
- *symmetry:* $d(x,y) = d(y,x)$;
- *triangle inequality:* $d(x,z) \leq d(x,y) + d(y,z)$;
- *identity of indiscernibles (i.e., unique minimum):* $d(x,y) = 0 \Leftrightarrow x = y$.

Proof. The non-negative and symmetric properties are immediately apparent. The identity of indiscernibles is guaranteed due to the injective mapping function and the use

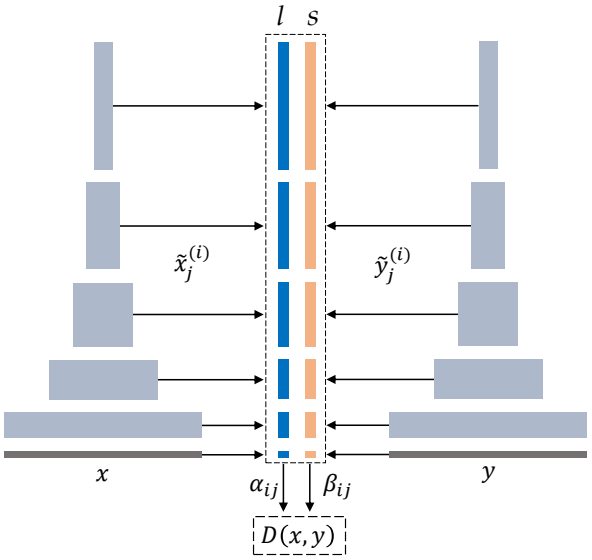


Fig. 6. VGG-based perceptual representation for the proposed DISTS model. It contains a total of six stages (including the zeroth stage of raw pixels), and the numbers of feature maps at each stage are 3, 64, 128, 256, 512 and 512, respectively. Global texture and structure similarity measurements are made at each stage, and combined with a weighted summation, giving rise to the final model defined in Eq. (7).

of SSIM-motivated quality measurements. To verify the triangle inequality, we first rewrite $d(x,y)$ as

$$d(x,y) = \sqrt{\sum_{i=0}^{m}\sum_{j=1}^{n_i} d_{ij}^2(x,y)}, \tag{9}$$

where

$$d_{ij}(x,y) = \sqrt{\alpha_{ij}(1-l(\tilde{x}_j^{(i)},\tilde{y}_j^{(i)})) + \beta_{ij}(1-s(\tilde{x}_j^{(i)},\tilde{y}_j^{(i)}))}. \tag{10}$$

Brunet *et al.* [58] have proved that $d_{ij}(x,y)$ is a metric for $\alpha_{ij} \geq 0$ and $\beta_{ij} \geq 0$. Then,

$$d(x,y) \leq \sqrt{\sum_{i,j}(d_{ij}(x,z)+d_{ij}(z,y))^2} \tag{11}$$

$$\leq \sqrt{\sum_{i,j} d_{ij}^2(x,z)} + \sqrt{\sum_{i,j} d_{ij}^2(y,z)} \tag{12}$$

$$= d(x,z) + d(z,y), \tag{13}$$

where Eq. (12) follows from the Cauchy–Schwarz inequality. □

2.4 Model Training

The perceptual weights $\{\alpha, \beta\}$ in Eq. (7) were jointly optimized for human perception of image quality and texture invariance. Specifically, for image quality, we minimized the absolute error between model predictions and human ratings:

$$E_1(x,y;\alpha,\beta) = |D(x,y;\alpha,\beta) - q(y))|, \tag{14}$$

where $q(y)$ denotes the normalized ground-truth quality score of y collected from psychophysical experiments. We chose the large-scale IQA dataset KADID-10k [59] as the training set, which contains 81 reference images, each of which is distorted by 25 distortion types at 5 distortion levels. In addition, we explicitly enforced the model to be invariant to texture substitution in a data-driven fashion. We minimized the distance (measured by Eq. (7)) between two patches (z_1, z_2) sampled from the same texture image z:

$$E_2(z;\alpha,\beta) = D(z_1,z_2;\alpha,\beta). \tag{15}$$

We selected texture images from the describable textures dataset (DTD) [60], consisting of 5,640 images (47 categories and 120 images for each category). In practice, we randomly sampled two minibatches $\mathcal{Q}$ and $\mathcal{T}$ from KADID-10k and DTD, respectively, and used a variant of stochastic gradient descent to adjust the parameters $\{\alpha, \beta\}$:

$$E(\mathcal{Q},\mathcal{T};\alpha,\beta) = \frac{1}{|\mathcal{Q}|}\sum_{x,y\in\mathcal{Q}} E_1(x,y;\alpha,\beta) + \lambda\frac{1}{|\mathcal{T}|}\sum_{z\in\mathcal{T}} E_2(z;\alpha,\beta) \tag{16}$$

where λ governs the trade-off between the two terms.

2.5 Connections to Other Full-Reference IQA Methods

The proposed DISTS model has a close relationship to a number of existing IQA methods.

- *SSIM and its variants* [3], [25], [63]: The multi-scale extension of SSIM [63] incorporates the variations of viewing conditions in IQA, and calibrates the cross-scale parameters via subjective testing on artificially synthesized images. Our model follows a similar approach, building on a multi-scale hierarchical representation and directly calibrating cross-scale parameters (*i.e.*, α, β) using subject-rated natural images with various distortions. The extension of SSIM into the complex wavelet domain [25] gains invariance to small geometric transformations by measuring relative phase patterns of the wavelet coefficients. As we show in Section 3.5, by optimizing for texture invariance, DISTS inherits insensitivity to mild geometric transformations. It is worth noting that unlike SSIM and its variants, DISTS is based on global spatial statistics, and thus does not provide a spatial map of quality.
- *The adaptive linear system framework* [18] decomposes the distortion between two images into a linear combination of components that are adapted to local image structures, separating structural and non-structural distortions. It generalizes many IQA models, including MSE, space/frequency weighting [20], [65], transform domain masking [22], and the tangent distance [66]. DISTS can be seen as an adaptive nonlinear system, where structure comparison captures structural distortions, and texture comparison measures non-structural distortions, with basis functions adapted to global image content.
- *Style and content separation* [55] based on the pre-trained VGG network has reignited the field of style transfer. Specifically, the style loss is built upon the correlations between convolution responses at the same stages (*i.e.*, the Gram matrix) while the content

TABLE 1
Performance comparison on three standard IQA databases. Larger PLCC, SRCC and KRCC values indicate better performance. CNN-based methods are highlighted in italics

Method	LIVE [61]			CSIQ [5]			TID2013 [62]		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	0.865	0.873	0.680	0.819	0.810	0.601	0.677	0.687	0.496
SSIM [3]	0.937	0.948	0.796	0.852	0.865	0.680	0.777	0.727	0.545
MS-SSIM [63]	0.940	0.951	0.805	0.889	0.906	0.730	0.830	0.786	0.605
VSI [64]	0.948	0.952	0.806	0.928	0.942	0.786	**0.900**	**0.897**	**0.718**
MAD [5]	**0.968**	**0.967**	**0.842**	**0.950**	**0.947**	**0.797**	0.827	0.781	0.604
VIF [4]	0.960	0.964	0.828	0.913	0.911	0.743	0.771	0.677	0.518
FSIM_c [11]	**0.961**	**0.965**	**0.836**	0.919	0.931	0.769	**0.877**	0.851	0.667
NLPD [39]	0.932	0.937	0.778	0.923	0.932	0.769	0.839	0.800	0.625
GMSD [12]	0.957	0.960	0.827	**0.945**	**0.950**	**0.804**	0.855	0.804	0.634
DeepIQA [13]	0.940	0.947	0.791	0.901	0.909	0.732	0.834	0.831	0.631
PieAPP [8]	0.908	0.919	0.750	0.877	0.892	0.715	0.859	**0.876**	**0.683**
LPIPS [7]	0.934	0.932	0.765	0.896	0.876	0.689	0.749	0.670	0.497
DISTS (ours)	0.954	0.954	0.811	0.928	0.929	0.767	0.855	0.830	0.639

loss is defined by the MSE between the two representations. These two components are redundant, and the combined loss does not have the desired property of unique minima we seek.

- *Image restoration losses* [67] in the era of deep learning are typically defined as a weighted sum of ℓ_p-norm distances computed on the raw pixels and several stages of VGG feature maps, where the weights are manually tuned for tasks at hand. Later stages of the VGG representation are often preferred so as to incorporate image semantics into low-level vision, encouraging perceptually meaningful details that are not necessarily aligned with the underlying image. This type of loss does not achieve the level of texture invariance we are looking for.

3 Experiments

In this section, we present the implementation details of the proposed DISTS. We then compare our method with a wide range of image similarity models in terms of quality prediction, texture similarity, texture classification/retrieval, and invariance of geometric transformations.

3.1 Implementation Details

We fixed the filter kernels of the pre-trained VGG, and learned the perceptual weights $\{\alpha, \beta\}$. The training was carried out by optimizing the objective function in Eq. (16), assuming a value of $\lambda = 1$, using Adam [68] with a batch size of 32 and an initial learning rate of 1×10^{-4}. After every 1K iterations, we reduced the learning rate by a factor of 2. We trained DISTS for 5K iterations, which takes approximately one hour on an NVIDIA GTX 2080 GPU. To ensure a unique minimum of our model, we projected the weights of the zeroth stage onto the interval $[0.02, 1]$ after each gradient step. We chose a 5×5 Hanning window to reduce subsampling-induced aliasing in the VGG representation. Both c_1 in Eq. (5) and c_2 in Eq. (6) were set to 10^{-6}. During training and testing, we followed the suggestions in [3], and re-scaled the input images such that the smaller dimension has 256 pixels. The size of texture patches as input to Eq. (15) was $256 \times 256 \times 3$, cropped from the same texture images.

3.2 Performance on Quality Prediction

After training on the entire KADID dataset [59], DISTS was tested on the other three standard IQA databases LIVE [61], CSIQ [5] and TID2013 [62]. We used the Pearson linear correlation coefficient (PLCC), the Spearman rank correlation coefficient (SRCC), and the Kendall rank correlation coefficient (KRCC) as evaluation criteria. Before computing PLCC, we fitted a four-parameter function to allow and compensate for a smooth nonlinear relationship:

$$\hat{D} = (\eta_1 - \eta_2) / (1 + \exp(-(D - \eta_3) / |\eta_4|)) + \eta_2, \quad (17)$$

where $\{\eta_i\}_{i=1}^4$ are parameters. We compared DISTS against a set of full-reference IQA methods, including nine knowledge-driven models and three data-driven CNN-based models. The implementations of all methods were obtained from the respective authors, except for DeepIQA [13], which was retrained on KADID for fair comparison. As LPIPS [7] has different configurations, we chose the default one (known as *LPIPS-VGG-lin*).

Results, reported in Table 1, demonstrate that DISTS performs favorably in comparison to both classic methods (*e.g.*, PSNR and SSIM [3]) and CNN-based models (*e.g.*, DeepIQA [13] and LPIPS [7]). Overall, the best performances across all three databases and all comparison metrics are obtained with MAD [5], FSIM_c [11] and GMSD [12]. It is worth noting that these three databases have been re-used for many years throughout the algorithm design processes, and recent full-reference IQA methods may be unintentionally over-adapting via extensive computational module selection, raising the risk of over-fitting (see Fig. 2). Fig. 7 shows scatter plots of raw model predictions of representative IQA methods versus subjective mean opinion scores (MOSs) on the TID2013 database. From the fitted functions (Eq. (17)), one can observe that DISTS is nearly linear in MOS.

We also tested DISTS on BAPPS [7], a large-scale and highly-varied patch similarity dataset. BAPPS contains traditional synthetic distortions, such as geometric and photometric manipulation, noise contamination, blurring and compression, CNN-based distortions (e.g., from denoising autoencoders and image restoration tasks), and distortions generated by real-world image processing systems. The human judgments are obtained from a two-alternative forced

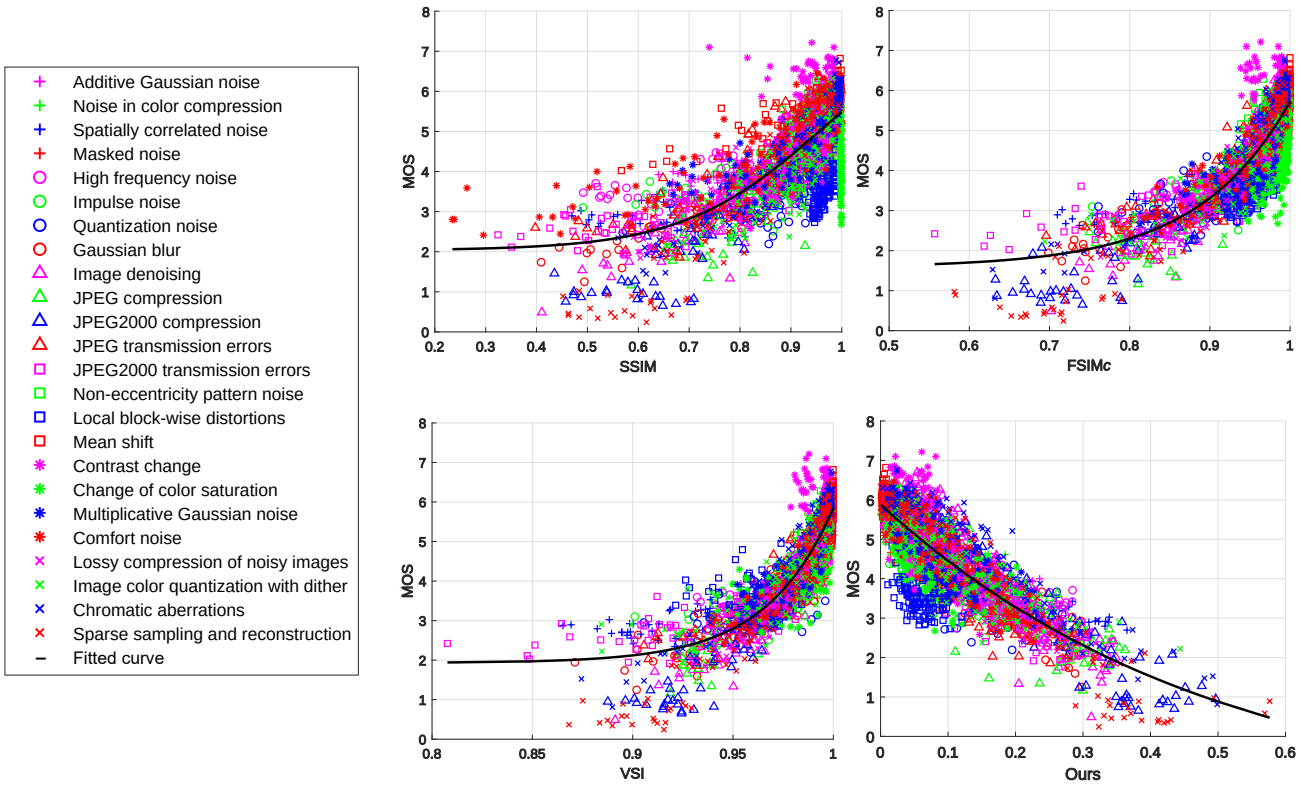


Fig. 7. Comparison of human mean opinion scores (MOSs) against SSIM, FSIM_c, VSI, and DISTS (ours) on the TID2013 database.

choice (2AFC) experiment. The evaluation criterion is the 2AFC score [7], which quantifies the proportion of human agreement with the IQA model, computed as $p\hat{p}+(1-p)(1-\hat{p})$, where p is the percentage of human choices in favor of a given image in each pair, and $\hat{p} \in \{0, 1\}$ is the preference of the IQA model. Larger values indicate better agreement between model predictions and human judgments. Results are compiled in Table 2, showing that DISTS (which was not trained on BAPPS, or any similar database) achieves a comparable performance to LPIPS [7] (which was trained on BAPPS). We conclude that DISTS predicts image quality well, and generalizes well to challenging unseen distortions, such as those caused by real-world algorithms.

3.3 Performance on Texture Similarity

We also tested the performance of DISTS on texture quality assessment. Since most knowledge-driven full-reference IQA models are not good at measuring texture similarity (see Fig 1), we only included a subset for reference. To these we added CW-SSIM [25] and three computational models specifically designed for texture similarity - STSIM [29], NPTSM [69] and IGSTQA [70]. STSIM is available in several configurations, and we chose local STSIM-2 that is publicly available[1].

We used a synthesized texture quality assessment database SynTEX [71], consisting of 21 reference textures with 105 synthesized versions generated by five texture synthesis algorithms. Table 3 shows the results of correlation coefficients, where we can see that texture similarity models generally perform better than IQA models. Focusing on texture similarity, IGSTQA [70] achieves a relatively high performance, but is still inferior to DISTS. This indicates that the VGG-based global measurements of DISTS capture the essential features and attributes of visual textures.

To further test the capabilities of DISTS in quantifying texture distortions, we constructed a texture quality database (TQD), based on 10 texture images selected from Pixabay[2]. Each texture image was corrupted with seven traditional synthetic distortions: additive white Gaussian noise, Gaussian blur, JPEG compression, JPEG2000 compression, pink noise, chromatic aberration, and image color quantization. For each distortion type, we randomly selected one distortion level from a set of three levels, and applied it to each texture image. We then created four copies of each texture using different texture synthesis algorithms, including two classical ones (a parametric model [15] and a non-parametric model [72]) and two CNN-based algorithms [32], [73]. Last, to produce "high-quality" images, we randomly cropped four subimages from each of the original textures. In total, TQD has 10×15 images. We gathered human data from 10 subjects, who had general knowledge of image processing but were unaware of the detailed purpose of the study. The viewing distance was fixed to enforce a visual resolution 32 pixels per degree of visual angle. Each subject was shown all ten sets of images, one set at a time, starting with the reference image, and was asked to rank the images according to their perceptual similarity to the reference. Rather than simply averaging

1. https://github.com/andreydung/Steerable-filter

2. https://pixabay.com/images/search/texture

TABLE 2
Performance comparison of various IQA methods on the BAPPS [7] dataset using the 2AFC score, which quantifies the agreement with human judgments. Values lie in the range $[0, 1]$, with a higher value indicating better agreement

Method	Synthetic distortions			Distortions by real-world algorithms					All
	Traditional	CNN-based	All	Super resolution	Video deblurring	Colorization	Frame interpolation	All	
Human	0.808	0.844	0.826	0.734	0.671	0.688	0.686	0.695	0.739
PSNR	0.573	0.801	0.687	0.642	0.590	0.624	0.543	0.614	0.633
SSIM [3]	0.605	0.806	0.705	0.647	0.589	0.624	0.573	0.617	0.640
MS-SSIM [63]	0.585	0.768	0.676	0.638	0.589	0.524	0.572	0.596	0.617
VSI [64]	0.630	0.818	0.724	0.668	0.592	0.597	0.568	0.622	0.648
MAD [5]	0.598	0.770	0.684	0.655	0.593	0.490	0.581	0.599	0.621
VIF [4]	0.556	0.744	0.650	0.651	0.594	0.515	0.597	0.603	0.615
FSIM_c [11]	0.627	0.794	0.710	0.660	0.590	0.573	0.581	0.615	0.640
NLPD [39]	0.550	0.764	0.657	0.655	0.584	0.528	0.552	0.600	0.615
GMSD [12]	0.609	0.772	0.690	0.677	0.594	0.517	0.575	0.613	0.633
DeepIQA [13]	0.703	0.794	0.748	0.660	0.582	0.585	0.598	0.615	0.650
PieAPP [8]	0.727	0.770	0.746	0.684	0.585	0.594	0.598	0.627	0.659
LPIPS [7]	**0.760**	**0.828**	**0.794**	**0.705**	**0.605**	**0.625**	**0.630**	**0.641**	**0.692**
DISTS (ours)	**0.772**	**0.822**	**0.797**	**0.710**	**0.600**	**0.627**	**0.625**	**0.651**	**0.689**

TABLE 3
Performance comparison on two texture quality databases. Texture similarity models are highlighted in italics

Method	SynTEX [71]			TQD (proposed)		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
SSIM [3]	0.619	0.620	0.446	0.330	0.307	0.185
CW-SSIM [25]	0.532	0.497	0.335	0.344	0.325	0.238
DeepIQA [13]	0.550	0.512	0.354	0.458	0.444	0.323
PieAPP [8]	0.719	0.715	0.532	0.721	0.718	0.556
LPIPS [7]	0.674	0.663	0.478	0.402	0.392	0.301
STSIM [29]	0.650	0.643	0.469	0.422	0.408	0.315
NPTSM [69]	0.505	0.496	0.361	0.678	0.679	0.547
IGSTQA [70]	**0.816**	**0.820**	**0.621**	**0.804**	**0.802**	**0.651**
DISTS (ours)	**0.901**	**0.923**	**0.759**	**0.903**	**0.910**	**0.785**

the human opinions, we used reciprocal rank fusion [74] to obtain the final ranking

$$r(x) = \sum_{k=1}^{K} \frac{1}{\gamma + r_k(x)}, \tag{18}$$

where $r_k(x)$ is the rank of x given by the k-th subject and γ is an additive constant that helps to mitigate the impact of outliers [74]. Table 3 lists the results, where we computed the correlations within each texture pattern and averaged them across textures. We found that nearly all existing models perform poorly on the new database, including those tailored for texture similarity. In contrast, DISTS significantly outperforms these methods by a large margin. Fig. 8 shows a set of texture examples, where we noticed that DISTS gives high rankings to resampled images and low rankings to images suffering from visible distortions. This demonstrates that DISTS is in close agreement with human perception of texture quality, and suggests potential uses in other texture analysis problems, such as high-quality texture retrieval.

3.4 Applications to Texture Classification and Retrieval

We also applied DISTS to texture classification and retrieval. We used the grayscale and color Brodatz texture databases [75] (denoted by GBT and CBT, respectively), each of which contains 112 different texture images. We resampled nine non-overlapping $256 \times 256 \times 3$ patches from each texture pattern. Fig. 9 shows a representative texture image from CBT, partitioned into nine patches.

The texture classification problem consists of assigning an unknown sample image to one of the known texture classes. For each texture, we randomly chose five patches for training, two for validation, and the remaining two for testing. A simple k-nearest neighbors (k-NN) classification algorithm was implemented, which allowed us to incorporate and compare different similarity models as distance measures. The predicted label of a test image was determined by a majority vote over its k nearest neighbors in the training set, where the value of k was chosen using the validation set. We implemented a baseline model - the bag-of-words of SIFT features [76] with k-NN. The classification accuracy results are listed in Table 4, where we can see that this baseline model beats most image similarity-based k-NN classifiers, except LPIPS (on CBT) and DISTS. This shows that our model is effective at discriminating and classifying textures that are visually different to the human eye.

The content-based texture retrieval problem consists of searching for images from a large database that are visually similar. In our experiment, for each texture, we set three patches as the queries, and aimed to retrieve the remaining six patches. Specifically, the distances between each query and the remaining images in the dataset were computed and ranked so as to retrieve the images with minimal distances. To evaluate the retrieval performance, we used mean average precision (mAP), which is defined by

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \left(\frac{1}{K} \sum_{k=1}^{K} P(k) \times \text{rel}(k) \right), \tag{19}$$

where Q is the number of queries, K is the number of similar images in the database, $P(k)$ is the precision at cut-off k in the ranked list, and $\text{rel}(k)$ is an indicator function equal to one if the item at rank k is a similar image and zero otherwise. As seen in Table 4, DISTS achieves the best performance on both CBT and GBT datasets. The classification/retrieval errors are primarily due to textures with noticeable inhomogeneities (*e.g.*, middle patch in Fig. 9).

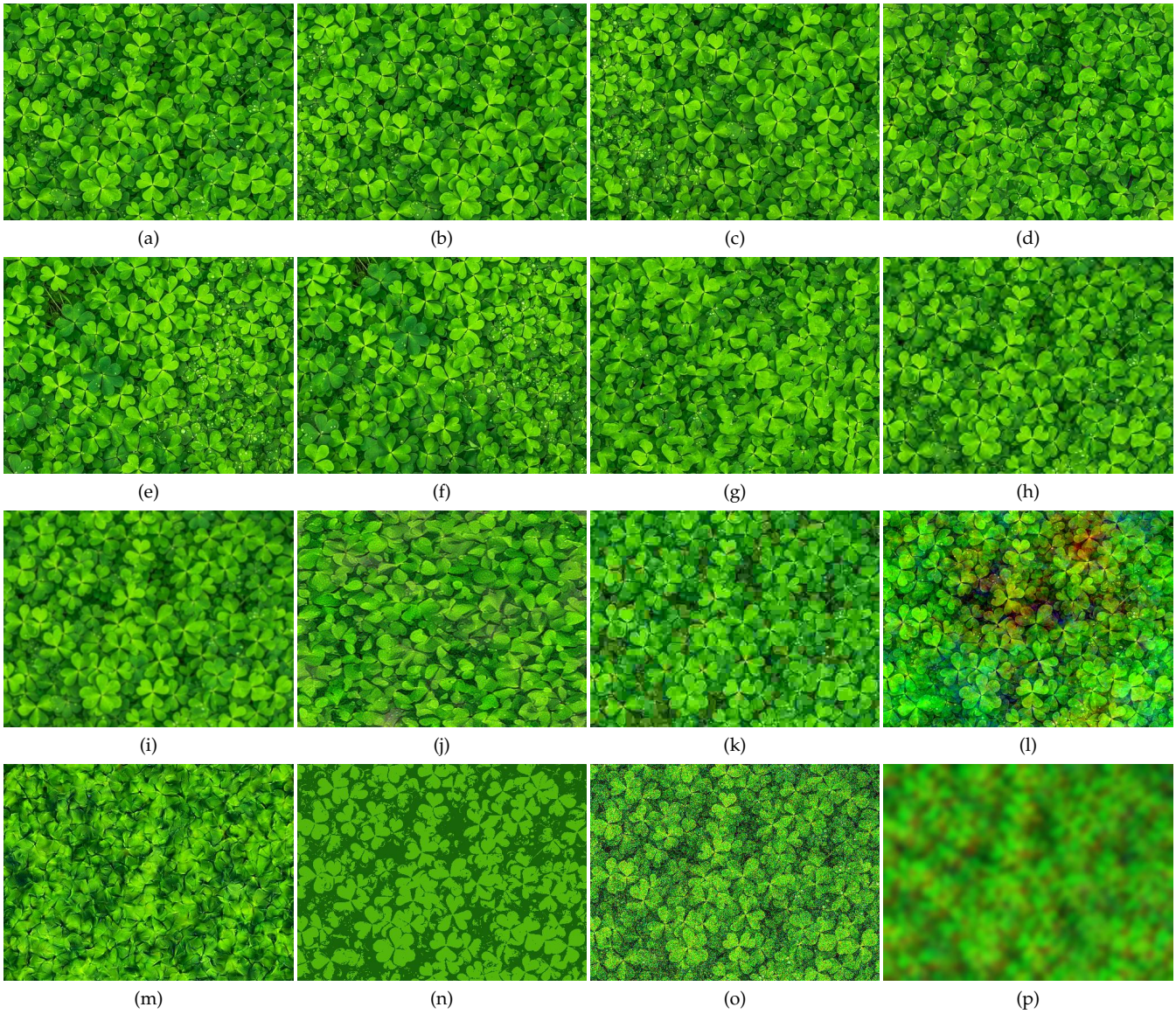


Fig. 8. One set of texture images from TQD, ordered according to their rankings by DISTS. **(a)** Reference image. **(b)-(p)** Corrupted images ranked by DISTS from high quality to low quality, respectively.

In addition, the performance on GBT is slightly reduced compared with that on CBT, indicating the importance of color information in these tasks.

Classification and retrieval of texture patches resampled from the same images are relatively easy tasks. We also tested DISTS on a more challenging large-scale texture database, the Amsterdam Library of Textures (ALOT) [77], containing photographs of 250 textured surfaces, from 100 different viewing angles and illumination conditions. Again, we adopted a naïve k-NN method ($k = 100$) using our model as the measure of distance, and tested it on 20% of the samples randomly selected from the database. Without training on ALOT, DISTS achieves a reasonable classification accuracy of 0.926, albeit lower than the value of 0.959 achieved by a knowledge-driven method [78] with hand-crafted features and support vector machines, and the value of 0.993 achieved by a data-driven CNN-based method [79]. The primary cause of errors when using DISTS in this task is that images from the same textured surface can appear quite different under different lighting or viewpoint conditions, as seen in the example in Fig. 10. DISTS, which is designed to capture visual appearance only, could likely be improved for this task by fine-tuning the perceptual weights (along with the VGG network parameters) on a small subset of human-labelled ALOT images.

3.5 Invariance to Geometric Transformations

Apart from texture similarity, most full-reference IQA measures fail dramatically when the original and distorted images are misregistered, either globally or locally. The underlying reason is again reliance on the assumption of pixel alignment. Although pre-registration can alleviate this issue, it comes with substantial computational complexity, and does not work well in the presence of severe distortions [19]. In this subsection, we investigated the degree of

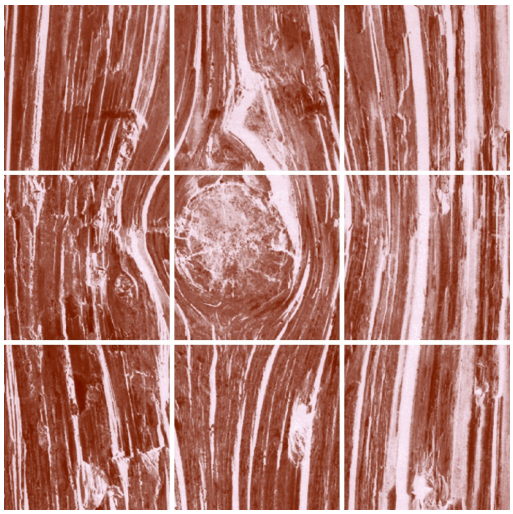


Fig. 9. Nine non-overlapping patches sampled from an example texture photograph in the Brodatz color texture dataset.

TABLE 4
Classification and retrieval performance comparison on the Brodatz texture dataset [75]

Method	Classification acc.		Retrieval mAP	
	CBT	GBT	CBT	GBT
SSIM [3]	0.397	0.210	0.371	0.145
CW-SSIM [25]	-	0.424	-	0.351
DeepIQA [13]	0.388	0.308	0.389	0.293
PieAPP [8]	0.173	0.115	0.257	0.153
LPIPS [7]	**0.960**	0.861	**0.951**	0.839
STSIM [29]	-	0.708	-	0.632
NPTSM [69]	-	0.895	-	0.837
IGSTQA [70]	-	0.862	-	0.798
SIFT [76]	0.924	**0.928**	0.859	**0.865**
DISTS (ours)	**0.995**	**0.968**	**0.988**	**0.951**

invariance of DISTS to geometric transformations that are imperceptible to the visual system.

As there are no subject-rated IQA databases designed for this specific purpose, we augmented the LIVE database [61] (LIVE_Aug) with geometric transformations. In real-world scenarios, an image should first undergo geometric transformations (*e.g.*, camera movement) and then distortions (*e.g.*, JPEG compression). We followed the suggestion in [19], and implemented an equivalent but much simpler approach - directly applying the transformations to the original image. Specifically, we augmented reference images using four geometric transformations: 1) shift by 5% pixels in horizontal direction, 2) clockwise rotation by a degree of 3°, 3) dilation by a factor of 1.05, and 4) their combination. This yields a set of $(4+1)\times 779$ reference-distortion pairs in the augmented LIVE database. Since the transformations are modest, the quality scores of distorted images with respect to the modified reference images are assumed to be the same as with respect to the original reference image.

The SRCC results of the augmented LIVE database are shown in Table 5. We found that data-driven methods based on CNNs significantly outperform traditional ones. Even so, their performance is often made worse by sensitivity to transformations that arises during downsampling without proper Nyquist band limiting. Trained on augmented data by geometric transformations, GTI-CNN [19] achieves desirable invariance at the cost of discarding perceptually important features (see Fig. 2). DISTS is seen to perform extremely well across all distortions and exhibit a high degree of robustness to geometric transformations, which we believe arises from 1) replacing max pooling with ℓ_2 pooling, 2) using global quality measurements, and 3) optimizing for invariance to texture resampling (see also Fig. 11).

TABLE 5
SRCC comparison of IQA models to human perception using the LIVE database augmented with geometric transformations

Method	Translation	Rotation	Dilation	Mixed	Total
PSNR	0.159	0.153	0.152	0.146	0.195
SSIM [3]	0.171	0.168	0.177	0.166	0.190
MS-SSIM [63]	0.165	0.174	0.198	0.174	0.177
CW-SSIM [25]	0.207	0.312	0.364	0.219	0.194
VSI [64]	0.282	0.360	0.372	0.297	0.309
MAD [5]	0.354	0.630	0.587	0.453	0.327
VIF [4]	0.296	0.433	0.522	0.387	0.294
FSIM_c [11]	0.380	0.396	0.408	0.365	0.339
NLPD [39]	0.062	0.074	0.083	0.066	0.112
GMSD [12]	0.252	0.299	0.303	0.247	0.288
DeepIQA [13]	0.822	**0.919**	**0.918**	0.881	0.859
PieAPP [8]	0.850	0.903	0.902	0.879	0.874
LPIPS [7]	0.811	0.908	0.893	0.861	0.779
GTI-CNN [19]	**0.864**	0.906	0.904	**0.890**	**0.875**
DISTS (ours)	**0.948**	**0.939**	**0.946**	**0.937**	**0.928**

3.6 Ablation Study

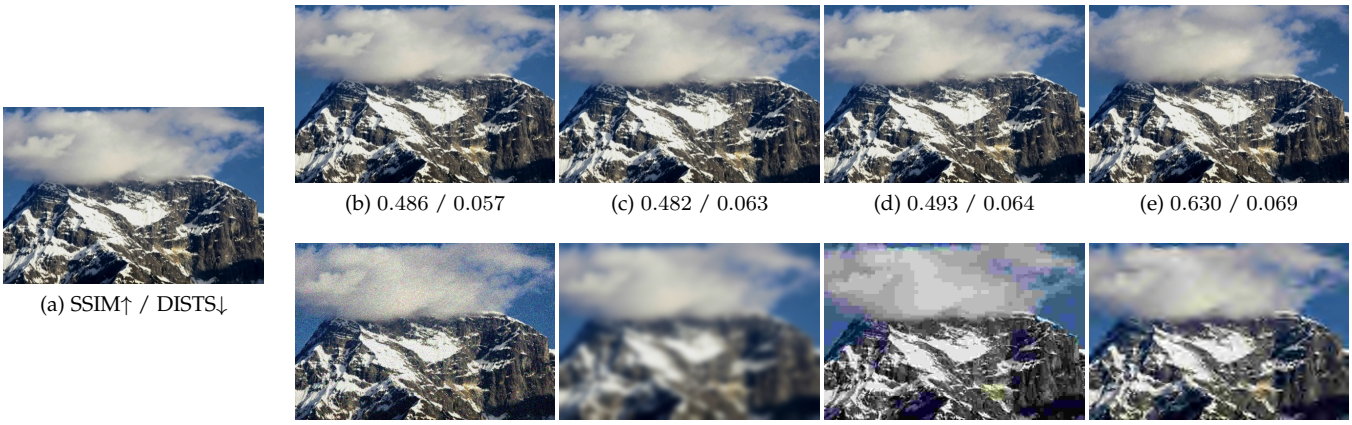
In this subsection, we conducted ablation experiments to single out the individual contributions of key modifications of DISTS, in comparison to the most closely related alternative - LPIPS. We trained a series of intermediate models between LPIPS and DISTS:

(a) Original LPIPS;
(b) Replace max pooling in LPIPS with ℓ_2 pooling;
(c) Add the input image on the top of (b);
(d) Replace the Euclidean distance in LPIPS with local SSIM measurements (within a sliding window of size 11×11) on top of (c);
(e) Replace the Euclidean distance in LPIPS with global SSIM measurements on top of (c);
(f) Train (c) by adding the E_2 term in Eq. (15);
(g) Train (d) by adding the E_2 term;
(h) Train (e) by adding the E_2 term, which is equivalent to DISTS.

Performance of these models is shown in Table 6, from which we draw several conclusions. First, ℓ_2 pooling is slightly better than max pooling. The main motivation of adopting ℓ_2 pooling is to de-alias the intermediate representations, as documented in [42]. Second, incorporating the input image in the representation has little impact on the performance, but it ensures a unique minimum of DISTS, which is beneficial in perceptual optimization [80]. Third, the global SSIM-like distance outperforms the Euclidean distance, especially in measuring similarity of visual textures and invariance to geometric transformations. We also tested local SSIM measurements within a sliding window size of 11×11 (d), which gives inferior performance. Last, training with the E_2 term is important for texture-related

(a) Reference (b) $D = 0.173$ (c) $D = 0.255$ (d) $D = 0.398$ (e) $D = 0.427$

Fig. 10. Five images of "soil", photographed under different lighting and viewpoint conditions, from the ALOT dataset. We computed the DISTS score for each of the images (b)-(e) with respect to the reference (a). Consistent with the significantly higher values, (d) and (e) are visually distinct from (a), although all of these images are drawn from the same category.



(a) SSIM↑ / DISTS↓ (b) 0.486 / 0.057 (c) 0.482 / 0.063 (d) 0.493 / 0.064 (e) 0.630 / 0.069

(f) 0.539 / 0.161 (g) 0.637/0.329 (h) 0.705 / 0.270 (i) 0.730 / 0.284

Fig. 11. A visual example to demonstrate robustness of DISTS to geometric transformations. **(a)** Reference image. **(b)** Translated rightward by 5% pixels. **(c)** Dilated by a factor 1.05. **(d)** Rotated by 3 degrees. **(e)** Cloud movement. **(f)** Corrupted with additive Gaussian noise. **(g)** Gaussian blur. **(h)** JPEG compression. **(i)** JPEG2000 compression. Below each image are the values of SSIM and DISTS, respectively. SSIM values are similar or better (larger) for the bottom row, whereas our model reports better (smaller) values for the top row, consistent with human perception.

tasks, improving invariance to geometric transformations, although it slightly hurts the performance on standard IQA databases. We concluded that the improved quality prediction and texture similarity performance of DISTS relative to LPIPS is due to the combination of these key modifications.

4 Conclusions

We have presented a new full-reference IQA method, DISTS, which is the first of its kind with built-in tolerance to texture resampling. Our model unifies structure and texture similarity, providing good predictions of human quality ratings on both textures and natural photographs, is robust to mild geometric distortions, and performs well in texture classification and retrieval.

DISTS is based on the pre-trained VGG network for object recognition. By computing the global means of convolution responses at each stage, we established a universal parametric texture model similar to that of Portilla & Simoncelli [15]. These statistical measurements provide a rich but relatively low-dimensional characterization of texture appearance, as verified using synthesis (Fig. 4). Despite the empirical success, we believe an important direction for future work is to analyze this "black box" to understand 1) what and how certain texture features and attributes are captured by the pre-trained network, and 2) the importance of cascaded convolution and subsampled pooling in summarizing useful texture information. It is also of interest to extend the current model to measure distortions locally, as is done in SSIM. In this case, the distance measure could be reformulated to adaptively select between structure and texture measures as appropriate, instead of linearly combining them with fixed weights.

The most direct use of IQA measures is for performance assessment and comparison of image processing systems. But perhaps more importantly, they may be used to optimize image processing methods, so as to improve the visual quality of their results. In this context, most existing IQA measures present major obstacles due to the fact that they lack desired mathematical properties that aid optimization (*e.g.*, injectivity, differentiability and convexity). In many cases, they rely on surjective mappings, and minima are non-unique (see Fig. 2). Although DISTS enjoys several advantageous mathematical properties, it is still highly non-convex (with abundant saddle points and plateaus), and recovery from random noise using stochastic gradient descent methods (see Fig. 2) requires many more iterations than for SSIM. In practice, the larger the weight of the structure term s at the zeroth stage (β_{0j} in Eq. (6)), the faster the optimization converges. However, to reach a reasonable level of texture invariance, the learned $\sum_{i,j} \alpha_{ij}$ should be larger than $\sum_{i,j} \beta_{ij}$, hindering optimization. We are currently analyzing DISTS in the context of perceptual optimization. Our initial results indicate that DISTS-based optimization of image processing applications, including denoising, deblurring, super-resolution, and compression,

TABLE 6
Ablation experiments: proposed DISTS model (last line) compared to LPIPS (first line), and intermediate variations. All models trained on KADID

Model	Quality prediction						Texture similarity						Geometric invariance		
	LIVE [61]			TID2013 [62]			SynTEX [71]			TQD (proposed)			LIVE_Aug		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
(a) LPIPS	0.934	0.936	0.769	0.850	0.824	0.626	0.591	0.589	0.452	0.403	0.401	0.302	0.801	0.793	0.629
(b) a + ℓ_2 pooling	0.937	0.938	0.770	0.851	0.824	0.626	0.594	0.592	0.459	0.410	0.406	0.305	0.807	0.802	0.633
(c) b + input image	0.935	0.935	0.768	0.851	0.825	0.627	0.582	0.581	0.449	0.410	0.409	0.303	0.795	0.789	0.625
(d) c + local SSIM	0.950	0.951	0.797	0.853	0.828	0.631	0.738	0.744	0.602	0.664	0.667	0.559	0.798	0.790	0.626
(e) c + global SSIM	**0.955**	**0.957**	**0.816**	**0.859**	**0.835**	**0.641**	**0.868**	**0.877**	**0.739**	**0.780**	**0.795**	**0.698**	**0.899**	**0.881**	**0.724**
(f) c + E_2 term	0.934	0.935	0.768	0.791	0.776	0.608	0.780	0.782	0.630	0.680	0.685	0.588	0.830	0.823	0.655
(g) d + E_2 term	0.929	0.931	0.766	0.801	0.783	0.615	0.774	0.778	0.625	0.672	0.678	0.579	0.820	0.816	0.649
(h) e + E_2 = DISTS	**0.954**	**0.954**	**0.811**	**0.855**	**0.830**	**0.639**	**0.901**	**0.923**	**0.759**	**0.903**	**0.910**	**0.785**	**0.931**	**0.928**	**0.762**

can lead to noticeable improvements in visual quality [80].

REFERENCES

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[2] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, A. B. Watson, Ed. The MIT Press, 1993, pp. 207–220.

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

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

[5] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1–21, 2010.

[6] V. Laparra, A. Berardino, J. Ballé, and E. P. Simoncelli, "Perceptually optimized image rendering," *Journal of the Optical Society of America A*, vol. 34, no. 9, pp. 1511–1525, 2017.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[8] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1808–1817.

[9] K. Popat and R. W. Picard, "Cluster-based probability model and its application to image and texture processing," *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 268–284, 1997.

[10] J. Balle, A. Stojanovic, and J.-R. Ohm, "Models for static and dynamic texture synthesis in image and video compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1353–1365, 2011.

[11] 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.

[12] 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.

[13] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015, pp. 1–14.

[15] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[16] B. Julesz, "Visual pattern discrimination," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 84–92, 1962.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *International Conference on Learning Representations*, pp. 1–10, 2013.

[18] Z. Wang and E. P. Simoncelli, "An adaptive linear system framework for image distortion analysis," in *IEEE International Conference on Image Processing*, 2005, pp. 1160–1163.

[19] K. Ma, Z. Duanmu, and Z. Wang, "Geometric transformation invariant image quality assessment using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 6732–6736.

[20] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974.

[21] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," in *Human Vision, Visual Processing, and Digital Display III*, 1992, pp. 2–15.

[22] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Human Vision, Visual Processing, and Digital Display V*, vol. 2179, 1994, pp. 127–141.

[23] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, 1995.

[24] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006.

[25] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. 573–576.

[26] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[27] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2019.

[28] A. D. Clarke, F. Halley, A. J. Newell, L. D. Griffin, and M. J. Chantler, "Perceptual similarity: A texture challenge," in *British Machine Vision Conference*, 2011, pp. 1–10.

[29] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2545–2558, 2013.

[30] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[31] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 7, pp. 971–987, 2002.

[32] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Conference on Neural Information Processing Systems*, 2015, pp. 262–270.

[33] H. Zhang, J. Xue, and K. Dana, "Deep TEN: Texture encoding network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 708–717.

[34] Y. Gao, Y. Gan, L. Qi, H. Zhou, X. Dong, and J. Dong, "A perception-inspired deep learning framework for predicting perceptual texture similarity," *IEEE Transactions on Circuits and Systems for Video Technology*, to appear, 2019.

[35] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational mod-

els of perceptual quantities," *Journal of Vision*, vol. 8, no. 12, pp. 1–13, Sep. 2008.

[36] A. Berardino, V. Laparra, J. Ballé, and E. Simoncelli, "Eigen-distortions of hierarchical representations," in *Conference on Neural Information Processing Systems*, 2017, pp. 3530–3539.

[37] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Nonlinear image representation for efficient perceptual coding," *IEEE Transactions on Image Processing*, vol. 15, no. 1, pp. 68–80, 2005.

[38] V. Laparra, J. Muñoz-Marí, and J. Malo, "Divisive normalization image quality metric revisited," *Journal of the Optical Society of America A*, vol. 27, no. 4, pp. 852–864, 2010.

[39] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[42] O. J. Hénaff and E. P. Simoncelli, "Geodesics of learned representations," *International Conference on Learning Representations*, pp. 1–10, 2016.

[43] A. V. Oppenheim, A. S. Willsky, and H. N. S., *Signals and Systems*. Pearson Education, 1998.

[44] B. Vintch, J. A. Movshon, and E. P. Simoncelli, "A convolutional subunit model for neuronal responses in macaque v1," *Journal of Neuroscience*, vol. 35, no. 44, pp. 14 829–14 841, 2015.

[45] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[46] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," in *International Conference on Learning Representations*, 2015, pp. 1–13.

[47] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017, pp. 1–27.

[48] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Conference on Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[49] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," in *International Conference on Machine Learning*, 2019, pp. 573–582.

[50] F. Ma, U. Ayaz, and S. Karaman, "Invertibility of convolutional generative networks from partial measurements," in *Conference on Neural Information Processing Systems*, 2018, pp. 9628–9637.

[51] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *22nd Annual Conference on Computer Graphics and Interactive Techniques*, 1995, pp. 229–238.

[52] S. C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, 1998.

[53] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *Journal of the Optical Society of America A*, vol. 7, no. 5, pp. 923–932, 1990.

[54] I. Ustyuzhaninov, W. Brendel, L. A. Gatys, and M. Bethge, "What does it take to generate natural textures?" in *International Conference on Learning Representations*, 2017, pp. 1–13.

[55] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[56] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *International Conference on Learning Representations*, 2017.

[57] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *International Joint Conferences on Artificial Intelligence*, 2017, pp. 2230–2236.

[58] 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, 2011.

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

[60] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.

[61] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. Cormack, "Image and video quality assessment research at LIVE," 2006, [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[62] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing Image Communication*, vol. 30, pp. 57–77, Jan. 2015.

[63] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, System and Computers*, 2003, pp. 1398–1402.

[64] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[65] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1164–1175, 1997.

[66] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition—tangent distance and tangent propagation," in *Neural networks: Tricks of the trade*, 1998, pp. 239–274.

[67] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015, pp. 1–15.

[69] M. Alfarraj, Y. Alaudah, and G. AlRegib, "Content-adaptive non-parametric texture similarity measure," in *IEEE International Workshop on Multimedia Signal Processing*, 2016, pp. 1–6.

[70] A. Golestaneh and L. 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 Workshop*, 2018, pp. 738–744.

[71] S. A. Golestaneh, M. M. Subedar, and L. J. Karam, "The effect of texture granularity on texture synthesis quality," in *Applications of Digital Image Processing XXXVIII*, vol. 9599, 2015, pp. 356 – 361.

[72] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1033–1038.

[73] X. Snelgrove, "High-resolution multi-scale neural texture synthesis," in *ACM SIGGRAPH Asia Technical Briefs*, 2017, pp. 13:1–13:4.

[74] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *ACM Special Interest Group on Information Retrieval*, vol. 9, 2009, pp. 758–759.

[75] S. Abdelmounaime and H. Dong-Chen, "New brodatz-based image databases for grayscale color and multiband texture analysis," *ISRN Machine Vision*, vol. 2013, 2013.

[76] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[77] G. J. Burghouts and J.-M. Geusebroek, "Material-specific adaptation of color invariant features," *Pattern Recognition Letters*, vol. 30, no. 3, pp. 306–313, 2009.

[78] M. Sulc and J. Matas, "Fast features invariant to rotation and scale of texture," in *European Conference on Computer Vision*, 2014, pp. 47–62.

[79] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 65–94, 2016.

[80] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of image quality models for optimization of image processing systems," *CoRR*, vol. abs/2005.01338, 2020. [Online]. Available: https://arxiv.org/abs/2005.01338