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 making local comparisons of 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 a multi-scale overcomplete representation. We empirically show 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 correlation of these spatial averages ("texture similarity") with correlation 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.

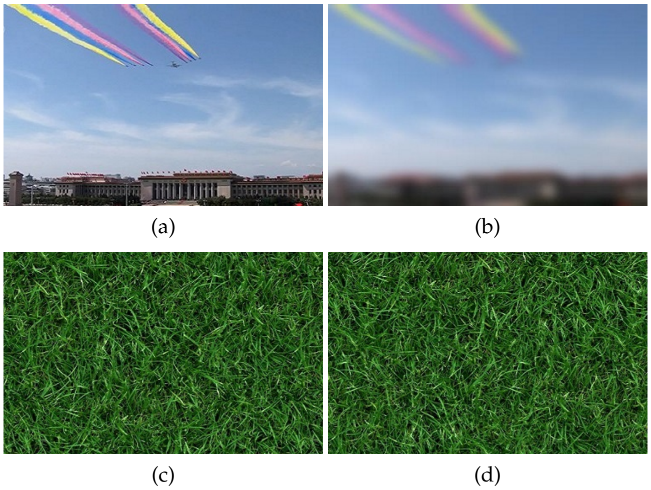
IMAGE quality assessment (IQA) aims to quantify human perception of image quality. The development of objective IQA measures is a fundamental problem in both human and computational vision, and is of paramount importance in a variety of real-world applications, such as image restoration, compression, 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].

A number of full-reference IQA methods have been proposed that provide a better account of human perception than MSE [2]–[7], and the *structural similarity* (SSIM) index [2] has become a *de facto* standard in the field of image processing. But these methods rely on perfect 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 will differ substantially in the precise arrangement of their features, but can appear nearly the same to a human observer (see Fig. 1). Since texture is 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 [8], [9].

Here 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) [10]. 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 [11]. Specifically, as originally proposed by Julesz [12] and used as a test of previous texture models [11], [13], [14], we 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 over a wide variety of textures, regardless of the initialization, thus revealing the robustness of this model to adversarial examples [15].

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 structure 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 optimized by comparing the responses of the model with a database of human image quality ratings. The second is optimized by minimizing the distance between pairs of

- *K. Ding, K. Ma, and S. Wang are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: keyanding2-c@my.cityu.edu.hk, kede.ma@cityu.edu.hk, shiqwang@cityu.edu.hk).*
- *E. P. Simoncelli is with the Howard Hughes Medical Institute, the Center for Neural Science, and the Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA (e-mail: eero.simoncelli@nyu.edu).*


(a)

(b)

(c)

(d)

Fig. 1. Existing full-reference IQA models are overly sensitive to point-by-point deviations between images of the same texture. **(a)** Original image and **(b)** same image, distorted by Gaussian blur. IQA scores: $\mathrm{MSE} = 23.15$, $\mathrm{SSIM} = 0.81$, and $\mathrm{DISTS} = 0.39$. (Note that, unlike MSE and DISTS, larger values of SSIM indicate better quality.) **(c)** Image of grass, cropped from a larger image and **(d)** a different crop from the same image. IQA scores: $\mathrm{MSE} = 376.32$, $\mathrm{SSIM} = 0.30$, and $\mathrm{DISTS} = 0.12$. The images (c) and (d) look very similar, while MSE and SSIM quantify them as differing more than the blurred and original images in (b) and (a). The proposed DISTS yields values consistent with human perception.

patches sampled from example texture images. We show that the resulting Deep Image Structure and Texture Similarity (DISTS) index can be transformed into a proper metric in the mathematical sense, and that it not only correlates well with human quality judgments in several independent datasets, but also 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 [16], [17], which may be imperceptible to the human visual system (HVS).

1 Background

Pioneering work on full-reference IQA dates back to the 1970s, when Mannos and Sakrison [18] investigated a class of visual fidelity measures in the context of rate-distortion optimization. A number of alternative measures were subsequently proposed [19], [20], trying to mimic certain functionalities of the HVS and penalize the errors between the reference and distorted images "perceptually". However, the HVS is a complex and highly nonlinear system [21], 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), leading to a number of problems regarding the definition of visual quality, quantification of suprathreshold distortions, and generalization to natural images [22]. The SSIM index [2] introduced the concept of comparing structure similarity (instead of measuring error visibility), opening the door to a new class of full-reference IQA measures [16], [23]–[25]. Other design methodologies for knowledge-driven IQA include information-theoretic criterion [3] 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 [6], [7], [17], [27]. However, databases of human quality scores are often insufficiently rich to constrain the large number of model parameters. As a result, the learned methods are at risk of over-fitting [28].

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 [11]. 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 texture. In this paper, we aim to measure the perceptual similarity of texture, a goal first elucidated and explored in [29], [30].

The response amplitudes and variances of computational texture features (*e.g.*, Gabor basis functions [31], local binary patterns [32]) have achieved good performance for texture classification, but do not correlate well with human perceptual observations as texture similarity measures [29], [30]. Texture representations that incorporate more sophisticated statistical features, such as correlations of complex wavelet coefficients [11], 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 within CNN-based representations [14], [33], [34] has led to even more powerful texture representations.

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 a texture region with a new sample that differs substantially in pixel values but looks essentially identical). 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, while exhibiting a high degree of tolerance to resampling. Finally, we combine these texture parameters with global structural measures 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 image x and the distorted image y 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

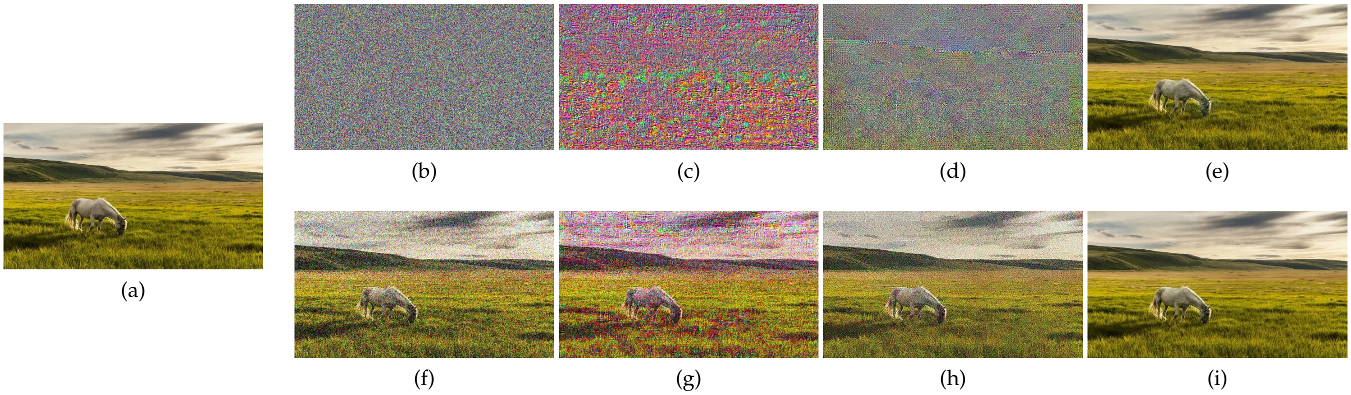


Fig. 2. Recovery of a reference image by optimization of IQA measures. Recovery is implemented by solving $y^\star = \arg\min_y D(x, y)$ with gradient descent, where D is an IQA distortion measure and x is a given reference image. **(a)** Original image. **(b)** Initial image y_0, containing purely white Gaussian noise. **(c)-(e)** Images recovered using GTI-CNN [17], GMSD [25], and the proposed DISTS, respectively. **(f)** Initial image by corrupting (a) with additive white Gaussian noise. **(g)-(i)** Images recovered by GTI-CNN, GMSD, and DISTS, respectively. In all cases, the optimization converges, yielding a distortion score substantially lower than that of the initial. GMSD and GTI-CNN both rely on surjective mappings, and as a result, fail on this simple task. In contrast, DISTS can successfully recover the image (a) from any initialization.

reason that MSE is inadequate as a perceptual IQA model. The function f should endeavor to map the pixel space to another space that is more perceptually uniform. Previous IQA methods have used filter banks for local frequency representation to capture the frequency-dependence of error visibility [4], [19]. Others have used transformations that mimic the early visual system [20], [37]–[39]. More recently, deep CNNs have shown surprising power in representing perceptual image distortions [6], [7], [27]. In particular, Zhang *et al*. [6] have demonstrated that pre-trained deep features from VGG have "reasonable" effectiveness in measuring perceptual quality.

As such, we chose to base our model on the VGG16 CNN [10], pre-trained for object recognition [40] on the ImageNet database [41]. The VGG transformation is constructed from 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 mathematical properties. First, in order to provide a good substrate for the invariances needed for texture resampling, we wanted the initial transformation to be *translation-invariant*. The "max pooling" operation of the original VGG architecture has been shown to disrupt translation-invariance [42], and leads to visible aliasing artifacts when used to interpolate between images with geodesic sequences [43]. 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 [44]. Following this principle, we replace all max pooling layers in VGG with weighted ℓ_2 pooling [43]:

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

where $\odot$ denotes pointwise product, and the blurring kernel $g(\cdot)$ is implemented by a Hanning window that approximately enforces the Nyquist criterion. As additional motivation, we note that ℓ_2 pooling has been used to describe the behavior of complex cells in primary visual cortex [45], and is also closely related to the complex modulus used in the scattering transform [46].

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 can be transferred into 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. Earlier IQA measures such as MSE and SSIM relied on an injective transformation (in fact, the identity mapping), but many more recent methods do not. For example, the mapping function in GMSD [25] extracts image gradients, discarding local luminance information that is essential to human perception of image quality. Similarly, GTI-CNN [17], uses a surjective CNN to construct the transformation, in an attempt to achieve invariance to mild geometric transformations.

Considerable effort has been made in developing invertible CNN-based transformations in the context of density modeling [47]–[50]. These methods place strict constraints on either network architectures [47], [49] or network parameters [50], which limit the expressiveness in learning quality-relevant representations (as empirically verified in our experiments). Ma *et al*. [51] 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 like to be perceptually relevant [14], [17].

Like most CNNs, VGG discards information at each stage of transformation. Given the difficulty of constraining parameters of VGG to ensure an injective mapping, we use a far simpler modification, incorporating the input image as an additional feature map (the "zeroth" layer of the network). The representation consists of the reference image x, concatenated with the convolution responses of five VGG

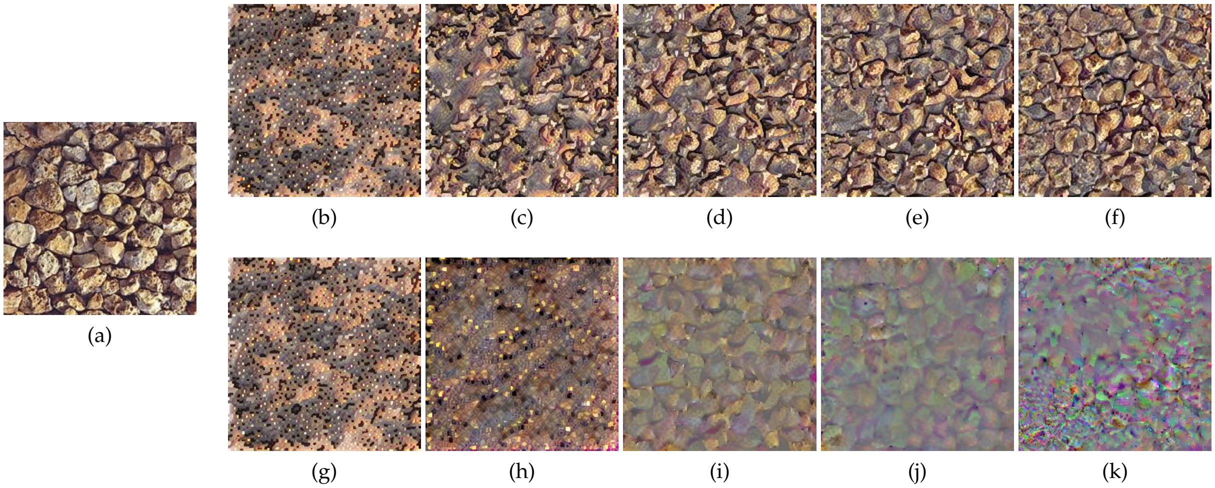


Fig. 3. Synthesis results of our texture model using statistical constraints up to a given layer (top) or from individual layers (bottom) of the pre-trained VGG. **(a)** Reference texture. **(b)** Up to conv1_2. **(c)** Up to conv2_2. **(d)** Up to conv3_3. **(e)** Up to conv4_3. **(f)** Up to conv5_3. **(g)** Only conv1_2. **(h)** Only conv2_2. **(i)** Only conv3_3. **(j)** Only conv4_3. **(k)** Only conv5_3.

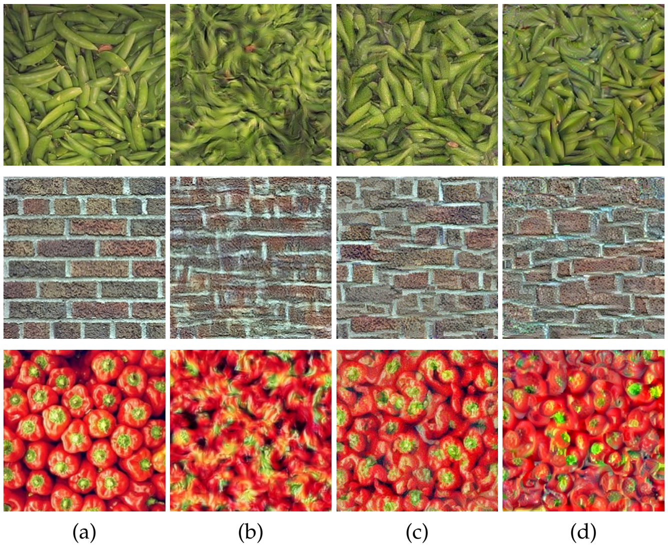


Fig. 4. Synthesis results for three example texture photographs. **(a)** Reference images. **(b)** Images synthesized using the method of Portilla & Simoncelli [11]. **(c)** Images synthesized using Gatys *et al*. [14]. **(d)** Images synthesized using our texture model (Eq. 4).

layers (labelled conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3):

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

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 compute the representaiton of the distorted image:

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

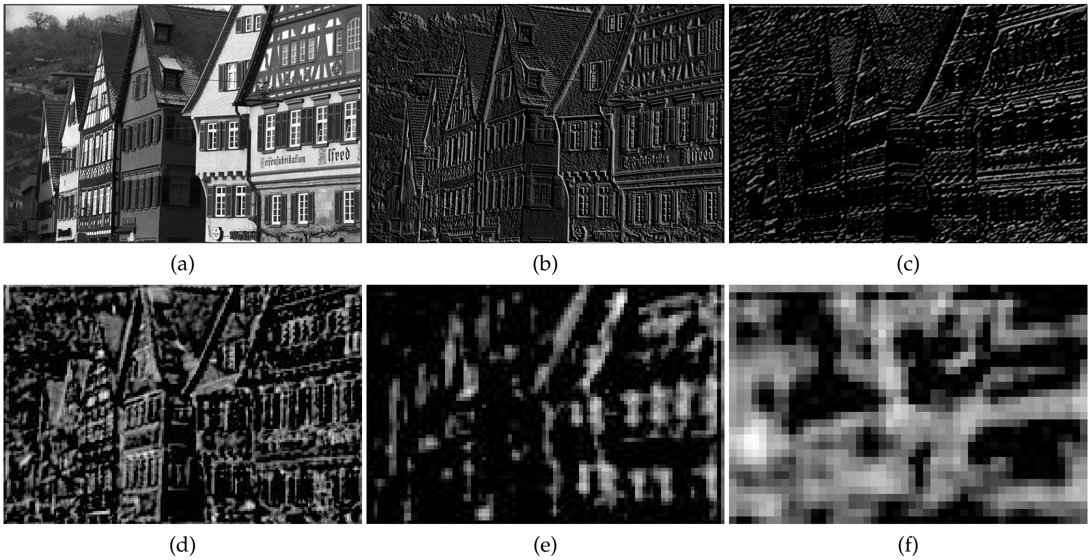
Fig. 2 demonstrates the injective property of the resulting transformation, in comparison to GMSD and GTI-CNN. For each IQA method, $D(x, y)$, we attempt to recover an original image x, by solving the optimization problem $y^\star = \arg\min_y D(x, y)$ with gradient descent. For initialization from white noise, or a noise-corrupted copy of the original image, both GMSD and GTI-CNN fail on this simple task.

2.2 Texture Representation

The visual appearance of texture is often characterized in terms of sets of local statistics [12] that are presumably measured by the HVS. Models consisting of various sets of features have been tested using synthesis [11], [13], [14], [52]: one generates an image with statistics that match those of a texture photograph. If the set of statistical measurements is a complete description of the appearance of the texture, then the synthesized image should be perceptually indistinguishable from the original [12], at least based on preattentive judgements [53].

Portilla & Simoncelli [11] found that the local correlations (and other pairwise statistics) of complex wavelet responses were sufficient to generate reasonable facsimiles of a wide variety of visual textures. Gatys *et al*. [14] used correlations across channels of several layers in a VGG network, and were able to synthesize consistently better textures, albeit with a much larger set of statistics (~ 306K parameters). Since this is typically larger than the number of pixels in the input image, it is likely that this image is unique in matching these statistics, and any diversity in the synthesis results may reflect local optima of the optimization procedure. Ustyuzhaninov *et al*. [54] provide 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 produce textures that are visually indiscernible to the human eye. Subsequent results suggest that matching only the mean and variance of CNN channels is sufficient for texture classification or style transfer [55]–[57].

In our experiments, we found that measuring only the spatial means of the feature maps (a total of $1,472$ statistics) provide an effective parametric model for visual texture.


(a) (b) (c)

(d) (e) (f)

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 scaled for better visibility.

Specifically, we used this model to synthesize textures [11] by solving

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

where x is the target texture image, $y^\star$ is the synthesized texture image, obtained by gradient descent optimization from a random initialization, and $\mu_{\tilde{x}_j}^{(i)}$ and $\mu_{\tilde{y}_j}^{(i)}$ are the global means of channels $\tilde{x}_j^{(i)}$ and $\tilde{y}_j^{(i)}$, 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. We find that measurements from early layers appear to capture basic intensity and color information, and those from later layers summarize the shape and structure information. By matching statistics up to layer conv5_3, the synthesized texture appears visually similar to the reference.

Fig. 4 shows three synthesis results of our texture model in comparison with the 710-parameter texture model of Portilla & Simoncelli [11] and the $\sim$ 306k-parameter model of Gatys *et al.* [14]. As one might expect, the synthesis quality of our model lies between the other two.

2.3 Perceptual Distance Measure

Next, we need to specify the 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 [2], we define separate quality measurements for the texture (using the global means) and the structure (using the global correlation) of each pair of corresponding feature maps:

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

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

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 is a weighted sum of global quality measurements at different convolution layers

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

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*

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

is a valid metric, satisfying

- *non-negativity:* $d(x,y) \geq 0$;

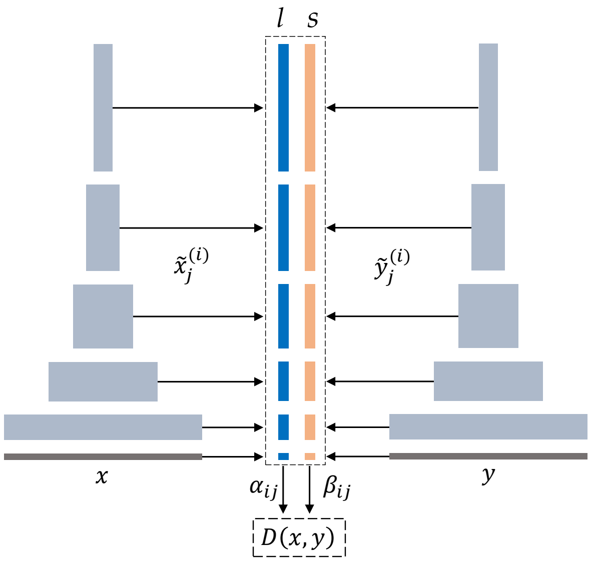


Fig. 6. VGG-based perceptual representation for the proposed DISTS model. It contains a total of six stages (including the zeroth stage of raw pixels), and the numbers of feature maps at each stage are 3, 64, 128, 256, 512 and 512, respectively. Global texture and structure similarity measurements are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

- *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 of SSIM-motivated quality measurements. It remains only to verify that d satisfies 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,

$$\begin{aligned} d(x,y) &\leq \sqrt{\sum_{i,j}(d_{ij}(x,z) + d_{ij}(z,y))^2} && (11)\\ &\leq \sqrt{\sum_{i,j} d_{ij}^2(x,z)} + \sqrt{\sum_{i,j} d_{ij}^2(y,z)} && (12)\\ &= d(x,z) + d(z,y), && (13) \end{aligned}$$

where Eq. (12) follows from the CauchySchwarz inequality. □

2.4 Model Training

The perceptual weights $\{\alpha, \beta\}$ in Eq. (7) are jointly optimized for human perception of image quality and texture invariance. Specifically, for image quality, we minimize 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 choose 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 enforce the model to be invariant to texture substitution in a data-driven fashion. We minimize 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 select 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 sample two minibatches $\mathcal{Q}$ and $\mathcal{T}$ from KADID-10k and DTD, respectively, and use 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* [2], [23], [63]: The multi-scale extension of SSIM [63] incorporates the variations of viewing conditions in IQA, and calibrates the cross-scale parameters via subject 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 in the complex wavelet domain [23] gains invariance to small geometric transformations by measuring relative phase patterns of the wavelet coefficients. As will be clear in Section 3.5, by optimizing for texture invariance, our method inherits insensitivity to mild geometric transformations. Nevertheless, DISTS does not offer a 2D map that indicates local quality variations across spatial locations as the SSIM family does.
- *The adaptive linear system framework* [16] decomposes the distortion between two images into a linear combination of adaptive components to local image structures, separating structural and non-structural distortions. It generalizes many IQA models, including MSE, space/frequency weighting [18], [65], transform domain masking [20], and the tangent distance [66]. DISTS can be seen as an adaptive nonlinear system, where structure comparison captures

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

Method	LIVE [61]				CSIQ [4]				TID2013 [62]			
	PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE
PSNR	0.865	0.873	0.680	13.716	0.819	0.810	0.601	0.154	0.677	0.687	0.496	0.912
SSIM [2]	0.937	0.948	0.796	9.575	0.852	0.865	0.680	0.138	0.677	0.687	0.496	0.912
MS-SSIM [63]	0.940	0.951	0.805	9.308	0.889	0.906	0.730	0.120	0.830	0.786	0.605	0.692
VSI [64]	0.948	0.952	0.806	8.682	0.928	0.942	0.786	0.098	**0.900**	**0.897**	**0.718**	**0.540**
MAD [4]	**0.968**	**0.967**	**0.842**	**6.907**	**0.950**	**0.947**	**0.797**	**0.082**	0.827	0.781	0.604	0.698
VIF [3]	0.960	0.964	0.828	7.679	0.913	0.911	0.743	0.107	0.771	0.677	0.518	0.789
FSIM$_c$ [24]	**0.961**	**0.965**	**0.836**	**7.530**	0.919	0.931	0.769	0.103	**0.877**	**0.851**	**0.667**	**0.596**
NLPD [39]	0.932	0.937	0.778	9.901	0.923	0.932	0.769	0.101	0.839	0.800	0.625	0.674
GMSD [25]	0.957	0.960	0.827	7.948	**0.945**	**0.950**	**0.804**	**0.086**	0.855	0.804	0.634	0.642
DeepIQA [27]	0.940	0.947	0.791	9.305	0.901	0.909	0.732	0.114	0.834	0.831	0.631	0.684
PieAPP [7]	0.909	0.918	0.749	11.417	0.873	0.890	0.705	0.128	0.829	0.844	0.657	0.694
LPIPS [6]	0.934	0.932	0.765	9.735	0.896	0.876	0.689	0.117	0.749	0.670	0.497	0.823
DISTS (ours)	0.954	0.954	0.811	8.214	0.928	0.929	0.767	0.098	0.855	0.830	0.639	0.643

structural distortions, and mean intensity comparison measures non-structural distortions, with basis functions adapted to global image content.

- *Style and content losses* [55] based on the pre-trained VGG network have reignited the field of style transfer. Specifically, the style loss is built upon the correlations between convolution responses at the same stages - the Gram matrix, while the content loss is defined by the MSE between two representations. The combined loss does not have the desired property of unique minimum we seek. By incorporating the input image as the zeroth stage feature representation of VGG and making SSIM-inspired quality measurements, the square root of DISTS is a valid metric.
- *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. Moreover, the weights of DISTS are jointly optimized for image quality and texture invariance, and can be used across multiple low-level vision tasks.

3 Experiments

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

3.1 Implementation details

We fix the filter kernels of the pre-trained VGG, and learn the perceptual weights $\{\alpha, \beta\}$. The training is 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 reduce the learning rate by a factor of 2. We train DISTS for 5K iterations, which takes approximately one hour on an NVIDIA GTX 2080 GPU. To ensure a unique minimum of our model, we project the weights of the zeroth stage onto the interval $[0.02, 1]$ after each gradient step. We choose a 5×5 Hanning window to anti-alias the VGG representation. Both c_1 in Eq. (5) and c_2 in Eq. (6) are set to 10^{-6}. During training and testing, we follow the suggestions in [2], and rescale the input images such that the smaller dimension is 256 pixels.

3.2 Performance on Quality Prediction

Trained on the entire KADID [59] dataset, DISTS is tested on the other three standard IQA databases LIVE [61], CSIQ [4] and TID2013 [62] to verify model generalizability. We use the Spearman rank correlation coefficient (SRCC), the Kendall rank correlation coefficient (KRCC), the Pearson linear correlation coefficient (PLCC), and the root mean square error (RMSE) as the evaluation criteria. Before computing PLCC and RMSE, we fit a four-parameter function to compensate the nonlinearity:

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

where $\{\eta_i\}_{i=1}^{4}$ are parameters to be fitted. We compare 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 are obtained from the respective authors, except for DeepIQA [27], which is retrained on KADID for fair comparison. As LPIPS [6] has different configurations, we choose the default one - *LPIPS-VGG-lin*.

Results, reported in Table 1, demonstrate that DISTS performs favorably in comparison to both classic methods (*e.g.*, PSNR and SSIM [2]) and CNN-based models (DeepIQA, PieAPP, and LPIPS). Overall, the best performances across all three databases and all comparison metrics are obtained with MAD [4], FSIM$_c$ [24] and GMSD [25]. It is worth noting that the three databases have been re-used for many years throughout the algorithm design processes, and recent full-reference IQA methods tend to adapt themselves to these databases, deliberately or unintentionally, via extensive computational module selection, raising the risk of overfitting (see Fig. 2). Fig. 7 shows scatter plots of model predictions of representative IQA methods versus the raw (i.e., before nonlinear mapping of Eq. 17) subjective mean opinion scores (MOSs) on the TID2013 database. From the fitted curves, one can observe that DISTS is nearly linear in MOS.

We also tested DISTS on BAPPS [6], a large-scale and highly-varied patch similarity dataset. BAPPS contains 1) traditional synthetic distortions, such as geometric and photometric manipulation, noise contamination, blurring, and compression, 2) CNN-based distortions, such as from denoising autoencoders and image restoration tasks, and distortions generated by real-world image processing systems. The human similarity judgments are obtained from a two-alternative forced choice (2AFC) experiment. From Table 2, we find that DISTS (which was not trained on BAPPS, or any similar database) achieves a comparable performance to LPIPS [6], which has been trained on BAPPS. We conclude that DISTS predicts image quality well, and generalizes 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 include SSIM [2] and FSIM$_c$ [24] for reference. We add CW-SSIM [23] and three computational models specifically designed for texture similarity - STSIM [30], NPTSM [69] and IGSTQA [70]. STSIM has several configurations, and we choose 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 SRCC and KRCC results, 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 investigate DISTS from texture similarity to texture quality, we construct a texture quality database (TQD), which contains 10 texture images selected from Pixabay[2]. For each texture image, we first add seven traditional synthetic distortions, including additive white Gaussian noise, Gaussian blur, JPEG compression, JPEG2000 compression, pink noise, chromatic aberration, and image color quantization. For each distortion type, we randomly select one distortion level from a total of three levels, and apply it to each texture image. We then create four copies for each texture using different texture synthesis algorithms, including two classical ones (a parametric model [11] and a non-parametric model [72]), and two CNN-based ones [14], [73]. Last, to produce "high-quality" images, we randomly crop four subimages from the original texture. In total, TQD has 10×15 images. We gather human data from 10 subjects, who have general knowledge of image processing but are unaware of the detailed purpose of the study. The viewing distance is fixed to 32 pixels per degree of visual angle. Each subject is shown all ten sets of images, one set at a time, and is asked to rank the images according to the perceptual similarity to the reference texture. Instead of simply averaging the human opinions, we use 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 a constant to mitigate the impact of high rankings by outlier systems [74]. Table 3 lists the SRCC and KRCC results, where we compute the correlations within each texture pattern and average them across textures. We find that nearly all existing models perform poorly on the new database, including those tailored to texture similarity. In contrast, DISTS significantly outperforms these methods by a large margin. Fig. 8 shows a set of texture examples, where we notice that DISTS gives high rankings to resampled images and low rankings to images suffering from visible distortions. This verifies that our model is in close agreement with human perception of texture quality, and has great potentials for use 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 representative texture samples extracted from CBT.

The texture classification problem consists of assigning an unknown sample image to one of the known texture classes. For each texture, we randomly choose five patches for training, two for validation, and the remaining two for testing. A simple k-nearest neighbors (k-NN) classification algorithm is implemented, which allows us to incorporate and compare different similarity models as distance measures. The predicted label of a test image is determined by a majority vote over its k nearest neighbors in the training set, where the value of k is chosen using the validation set. We implement 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 see that the 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 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 indistinguishable. In our experiment, for each texture, we set three patches as the queries and aim to retrieve the remaining six patches. Specifically, the distances between each query and the remaining images in the dataset are

[1] https://github.com/andreydung/Steerable-filter
[2] https://pixabay.com/images/search/texture

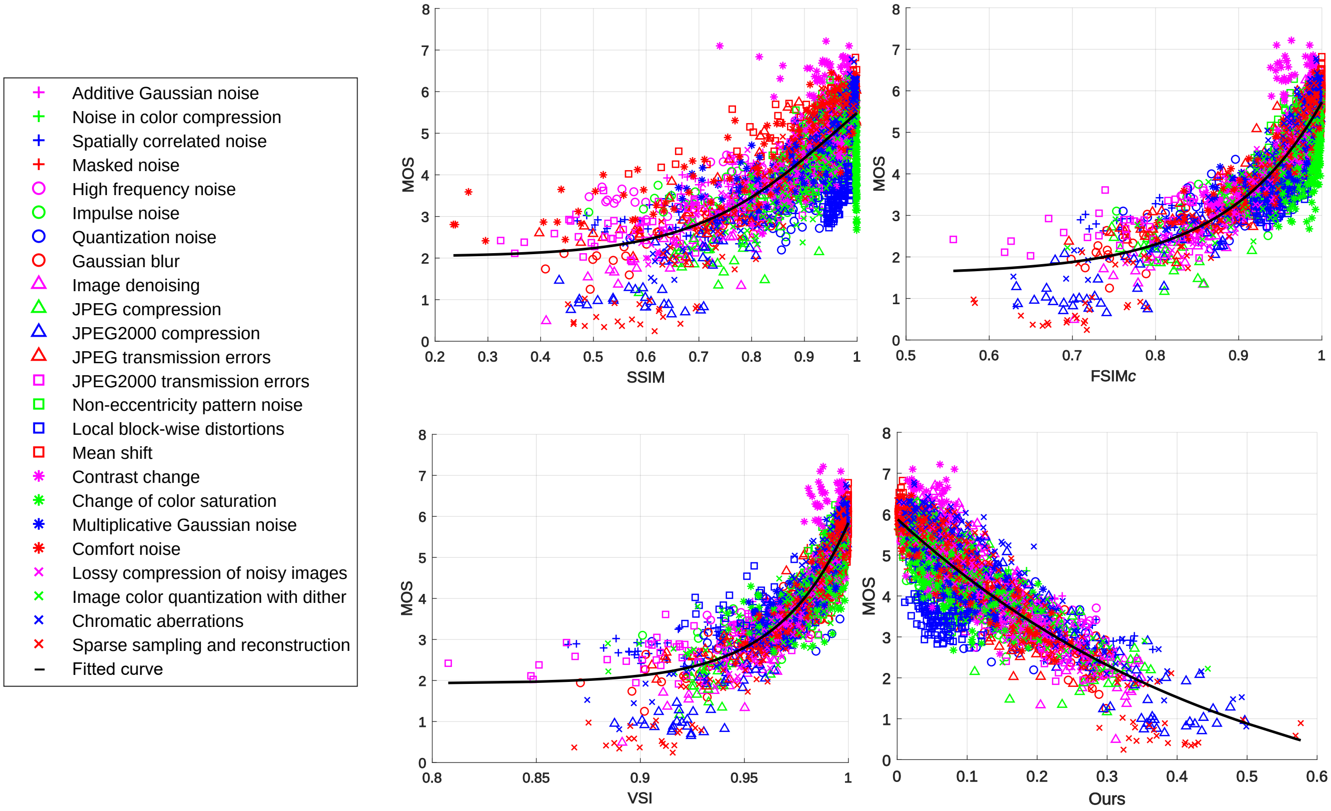


Fig. 7. Scatter plots of SSIM, FSIM$_c$, VSI, and DISTS (ours) on the TID2013 database. The fitted curve of DISTS is slightly more linear than the others.

TABLE 2
Performance comparison on the test sets of BAPPS [6] (higher is better)

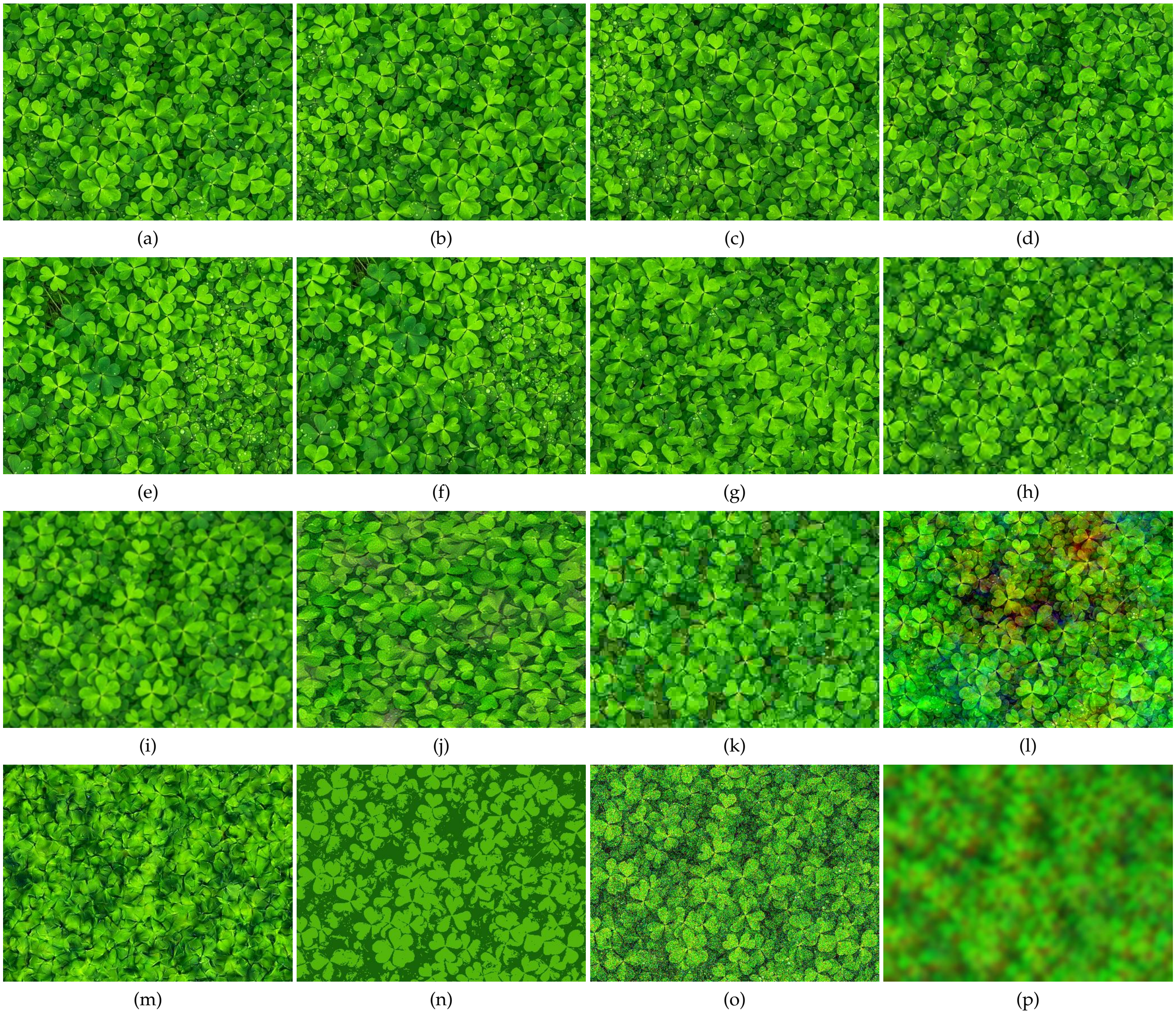
Method	Synthetic distortions			Distortions by real-world algorithms					All
	Traditional	CNN-based	All	Super resolution	Video deblurring	Colorization	Frame interpolation	All	
Human	0.808	0.844	0.826	0.734	0.671	0.688	0.686	0.695	0.739
PSNR	0.573	0.801	0.687	0.642	0.590	0.624	0.543	0.614	0.633
SSIM [2]	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 [4]	0.598	0.770	0.684	0.655	0.593	0.490	0.581	0.599	0.621
VIF [3]	0.556	0.744	0.650	0.651	0.594	0.515	0.597	0.603	0.615
FSIM$_c$ [24]	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 [25]	0.609	0.772	0.690	0.677	0.594	0.517	0.575	0.613	0.633
DeepIQA [27]	0.703	0.794	0.748	0.660	0.582	0.585	0.598	0.615	0.650
PieAPP [7]	0.725	0.769	0.747	0.685	0.582	0.594	0.598	0.626	0.658
LPIPS [6]	**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**

computed and ranked so as to retrieve the images with minimal distances. To evaluate the retrieval performance, we use mean average precision (mAP), which is defined by

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

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 $\mathrm{rel}(k)$ 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 (c)). 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


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

Fig. 8. A set of texture images ranked by DISTS on TQD. **(a)** Reference image. **(b)-(p)** are the distorted images ranked by DISTS from high quality to low quality, respectively.

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

Method	SynTEX [71]		TQD (proposed)	
	SRCC	KRCC	SRCC	KRCC
SSIM [2]	0.620	0.446	0.307	0.185
CW-SSIM [23]	0.497	0.335	0.325	0.238
DeepIQA [27]	0.512	0.354	0.444	0.323
PieAPP [7]	0.709	0.530	0.713	0.554
LPIPS [6]	0.663	0.478	0.392	0.301
STSIM [30]	0.643	0.469	0.408	0.315
NPTSM [69]	0.496	0.361	0.679	0.547
IGSTQA [70]	**0.820**	**0.621**	**0.802**	**0.651**
DISTS (Ours)	**0.923**	**0.759**	**0.910**	**0.785**

different viewing angles and illumination conditions. Again, we adopt a naïve k-NN method ($k = 100$) using our model as the measure of distance, and test 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, 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-labeled ALOT images.

3.5 Invariance to Geometric Transformations

Apart from texture similarity, most full-reference IQA measures fail dramatically when the original and distorted

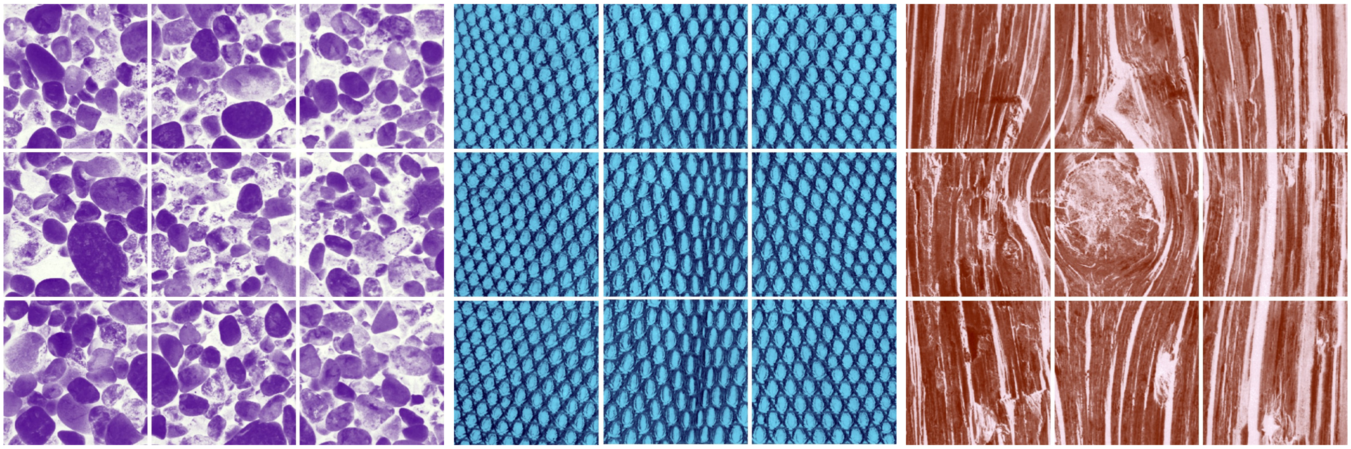

(a) (b) (c)

Fig. 9. Nine non-overlappin patches sampled from each of three example texture photographs in the Brodatz color texture dataset.

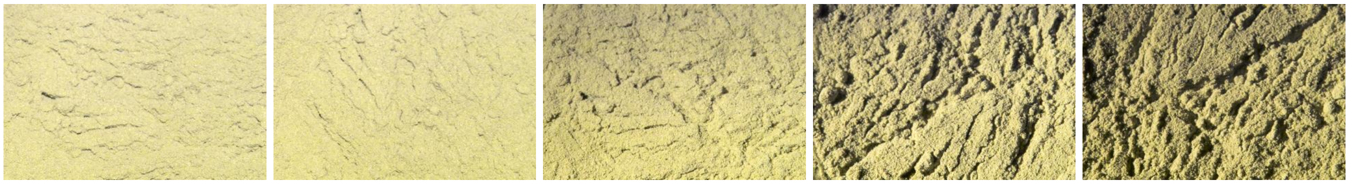

(a) Reference (b) $D = 0.173$ (c) $D = 0.255$ (d) $D = 0.398$ (e) $D = 0.427$

Fig. 10. Five images of "soil", photographed under different lighting and viewpoint conditions, from the ALOT dataset. We compute 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.

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

Method	Classification acc.		Retrieval mAP	
	CBT	GBT	CBT	GBT
SSIM [2]	0.397	0.210	0.371	0.145
CW-SSIM [23]	-	0.424	-	0.351
DeepIQA [27]	0.388	0.308	0.389	0.293
PieAPP [7]	0.178	0.117	0.260	0.157
LPIPS [6]	**0.960**	0.861	**0.951**	0.839
STSIM [30]	-	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**

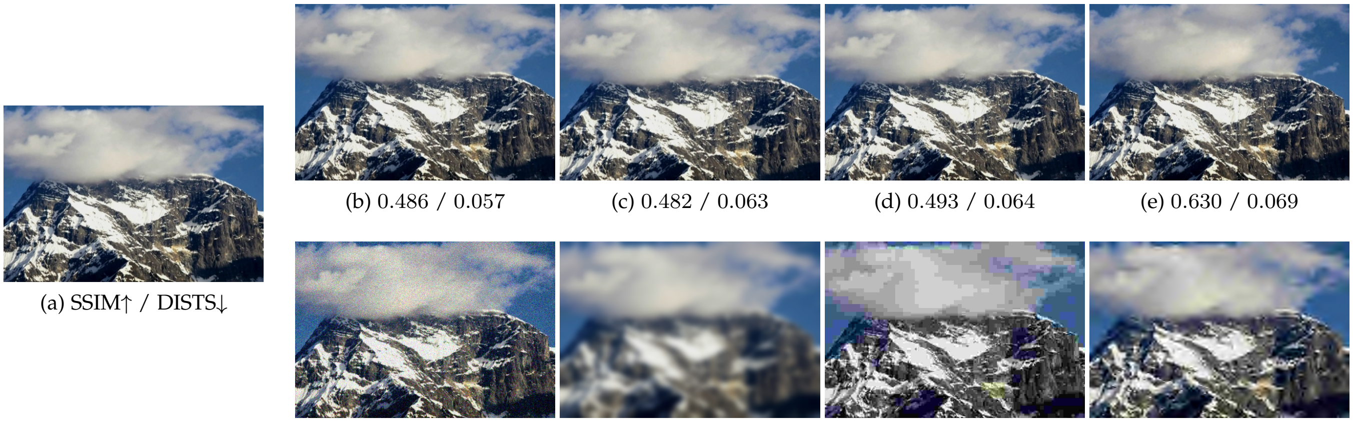
images are misregistered, either globally or locally. The underlying reason is again reliance on the assumption of pixel alignment. Although pre-registration alleviates this issue in certain occasions, it comes with substantial computational complexity, and does not work well in the presence of severe distortions [17]. Here we investigate the degree of invariance of DISTS to geometric transformations that are imperceptible to our visual system.

As there are no subject-rated IQA databases designed for this specific purpose, we augment 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 follow the suggestion in [17], and implement an equivalent but much simpler approach - directly applying the transformations to the original image. Specifically, we generate four augmented reference images using 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 find that data-driven methods based on CNNs outperform traditional ones by a large margin. Note that, even the simplest geometric transformation - translation - may hurt the performance of CNN-based methods, which indicates that this type of invariance does not come for free if CNNs ignore the Nyquist theorem when downsampling. Trained on augmented data by geometric transformations, GTI-CNN [17] 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 distortions

Method	Distortion type					Geometric transformation				Total
	JPEG2000	JPEG	Gauss. noise	Gauss. blur	Fast fading	Translation	Rotation	Dilation	Mixed	
PSNR	0.077	0.106	0.781	0.112	0.003	0.159	0.153	0.152	0.146	0.195
SSIM [2]	0.104	0.107	0.679	0.133	0.08	0.171	0.168	0.177	0.1660	0.190
MS-SSIM [63]	0.091	0.126	0.595	0.107	0.066	0.165	0.174	0.198	0.174	0.177
CW-SSIM [23]	0.062	0.182	0.579	0.065	0.054	0.207	0.312	0.364	0.219	0.194
VSI [64]	0.083	0.362	0.710	0.034	0.217	0.282	0.360	0.372	0.297	0.309
MAD [4]	0.195	0.418	0.542	0.149	0.274	0.354	0.630	0.587	0.453	0.327
VIF [3]	0.277	0.262	0.366	0.194	0.391	0.296	0.433	0.522	0.387	0.294
FSIM$_c$ [24]	0.104	0.432	0.634	0.106	0.283	0.380	0.396	0.408	0.365	0.339
NLPD [39]	0.060	0.069	0.501	0.166	0.047	0.062	0.074	0.083	0.066	0.112
GMSD [25]	0.048	0.470	0.477	0.106	0.235	0.252	0.299	0.303	0.247	0.288
DeepIQA [27]	0.813	0.873	0.948	0.827	0.813	0.822	**0.919**	**0.918**	0.881	0.859
PieAPP [7]	0.875	0.884	**0.952**	**0.912**	**0.908**	0.848	0.901	0.903	0.876	0.872
LPIPS [6]	0.730	0.872	0.919	0.592	0.743	0.811	0.908	0.893	0.861	0.779
GTI-CNN [17]	**0.879**	**0.910**	0.910	0.765	0.837	**0.864**	0.906	0.904	**0.890**	**0.875**
DISTS (ours)	**0.944**	**0.948**	**0.957**	**0.921**	**0.894**	**0.948**	**0.939**	**0.946**	**0.937**	**0.928**



(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.

4 Conclusions

We have presented a new full-reference IQA method, DISTS, which is the first of its kind with built-in invariance to texture resampling. Our model unifies structure and texture similarity, 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 [11]. Despite the empirical success, it is imperative to open this "black box" and to understand 1) what and how certain texture features and attributes are captured by the pre-trained network, 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 select between structure and texture measures as appropriate, instead of simply combining them linearly.

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}, 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, with the intention of learning a more suitable set of perceptual weights by adding the optimizability constraints. Initial results indicate that DISTS-based optimization of image processing applications, including denoising, deblurring, super-resolution, and compression can lead to noticeable improvements in visual quality.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[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] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[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 models 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] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *Journal of Machine Learning Research*, vol. 20, no. 184, pp. 1–25, 2019.

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

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

[45] 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.

[46] 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.

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

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

[49] 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.

[50] 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.

[51] 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.

[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 recognitiontangent 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.