Locally Adaptive Structure and Texture Similarity for Image Quality Assessment

Keyan Ding¹, Yi Liu², Xueyi Zou², Shiqi Wang¹, Kede Ma¹
¹City University of Hong Kong, Hong Kong
²Noah’s Ark Lab, Huawei Technologies, Shenzhen, China
keyan.ding@my.cityu.edu.hk, {liuyi113, zouxueyi}@huawei.com, {shiqwang, kede.ma}@cityu.edu.hk

ABSTRACT
The latest advances in full-reference image quality assessment (IQA) involve unifying structure and texture similarity based on deep representations. The resulting Deep Image Structure and Texture Similarity (DISTS) metric, however, makes rather global quality measurements, ignoring the fact that natural photographic images are locally structured and textured across space and scale. In this paper, we describe a locally adaptive structure and texture similarity index for full-reference IQA, which we term A-DISTS. Specifically, we rely on a single statistical feature, namely the dispersion index, to localize texture regions at different scales. The estimated probability (of one patch being texture) is in turn used to adaptively pool local structure and texture measurements. The resulting A-DISTS is adapted to local image content, and is free of expensive human perceptual scores for supervised training. We demonstrate the advantages of A-DISTS in terms of correlation with human data on ten IQA databases and optimization of single image super-resolution methods.

CCS CONCEPTS
• Computing methodologies → Image representations; Neural networks; • General and reference → Metrics.

KEYWORDS
Image quality assessment, structure similarity, texture similarity, perceptual optimization.

1 INTRODUCTION
Full-reference image quality assessment (IQA) aims to predict the perceived quality of a “distorted” image with reference to its original undistorted counterpart. It plays an indispensable role in the assessment and optimization of various image processing and computational photography algorithms. Humans appear to assess image quality quite easily and consistently, but the underlying mechanism is unclear, making bio-inspired IQA model development a challenging task.

For more than half a century, the field of full-reference IQA has been dominated by parsimonious knowledge-driven models with few hyperparameters. Representative examples include the mean squared error (MSE), the structural similarity (SSIM) index [34], the visual information fidelity (VIF) metric [26], the most apparent distortion (MAD) measure [16], and the normalized Laplacian pyramid distance (NLPD) [15]. Knowledge-driven IQA methods require statistical modeling of the natural image manifold and/or the human visual system (HVS) [8], which is highly nontrivial. Only crude computational approximations characterized by simplistic and restricted visual stimuli [36] have been developed.

Recently, there has been a trend shying away from knowledge-driven IQA models and toward data-driven ones, as evidenced by recent IQA methods [2, 6, 25, 41] based on deep neural networks (DNNs). Albeit with high correlation numbers on many image quality databases, these models have a list of theoretical and practical issues during deployment [32]. Arguably the most significant issue is with regard to gradient-based optimization. In [7], Ding et al. systematically evaluated more than 15 full-reference IQA models in the context of perceptual optimization, and found that a majority of methods fail in a naïve task of reference image recovery. This is not surprising because these methods rely on surjective mapping functions to transform the original and distorted images to a reduced “perceptual” space for quality computation [7]. Two DNN-based models that rely on (nearly) injective mappings are the exceptions: the learned perceptual image patch similarity (LPIPS) [41] and the deep image structure and texture similarity (DISTS) [6] metrics. According to the subjective user study in [7], LPIPS and DISTS are top-2 performers in optimization of three low-level vision tasks - blind image deblurring, single image super-resolution, and lossy image compression (see Fig. 8 in [7]).

LPIPS [41] makes quality measurements by point-by-point comparisons between deep features from the pre-trained VGG network [28]. As a result, it cannot properly handle “visual texture,” which is comprised of repeated patterns, subject to some randomization in their location, size, color, and orientation [24]. DISTS [6] provides a better account for texture similarity by identifying a compact set of statistical constraints as global spatial averages of VGG feature maps. When restricted to global texture images (see Fig. 1 (a)), the underlying texture model in DIISTS performs well in the analysis-by-synthesis test [24] originally advocated by Julesz [13]. However, it is well-known that natural photographic images are composed of
null
where $\{\alpha_{ij}, \beta_{ij}\}$ are positive learnable weights, optimized to match human perception of image quality and invariance to resampled texture patches [6]. Some key modifications of DISTS relative to SSIM and LPIPS are worth mentioning. First, $l_2$-pooling is adopted to replace the max pooling in the original VGG, which is conducive to de-alias and linearize the intermediate representations [11]. Second, the input image is incorporated as an additional feature map (i.e., $\tilde{X}^{(0)} = X$) to guarantee the injectivity of the feature transform $f$. Third, unlike Eq. (2), DISTS applies the "texture" similarity function $l(\cdot)$ and the structure similarity function $s(\cdot)$ globally to compare feature maps. It has been empirically proven sensitive to structural distortions and robust to texture substitutions.

2.4 DISTS

DISTS [6] is based on a variant of the VGG network, and makes global SSIM-like structure and texture similarity measurements:

$$\text{DISTS}(X, Y) = 1 - \sum_{i=0}^{M} \sum_{j=1}^{N_i} \left( \alpha_{ij} \left( \tilde{X}^{(i)}_j, \tilde{Y}^{(i)}_j \right) + \beta_{ij} s(\tilde{X}^{(i)}_j, \tilde{Y}^{(i)}_j) \right),$$

where $\{\alpha_{ij}, \beta_{ij}\}$ are positive learnable weights, optimized to match human perception of image quality and invariance to resampled texture patches [6]. Some key modifications of DISTS relative to SSIM and LPIPS are worth mentioning. First, $l_2$-pooling is adopted to replace the max pooling in the original VGG, which is conducive to de-alias and linearize the intermediate representations [11]. Second, the input image is incorporated as an additional feature map (i.e., $\tilde{X}^{(0)} = X$) to guarantee the injectivity of the feature transform $f$. Third, unlike Eq. (2), DISTS applies the "texture" similarity function $l(\cdot)$ and the structure similarity function $s(\cdot)$ globally to compare feature maps. It has been empirically proven sensitive to structural distortions and robust to texture substitutions.

3 A-DISTS

In this section, we present in detail the locally adaptive DISTS metric, namely A-DISTS. We first describe the use of the dispersion index to separate structure and texture at different locations and scales. We then compute the final quality score by adaptively weighting local structure and texture measurements.

Structure and Texture Separation. We want to identify robust statistics based on deep representations that are effective in separating structure and texture regions. However, the VGG network [28] used in DISTS suffers from the scale ambiguity. That is, we may re-scale a convolution filter by dividing the 3D tensor (and the associated bias term) by an arbitrary non-zero scalar. This can be compensated by re-scaling the next convolution filter connected to it by the same amount without changing the final softmax output. The scale ambiguity arises primarily from the adoption of half-wave rectification (i.e., ReLU) as the nonlinearity. As a consequence, the statistics computed from different convolution responses may be of arbitrary scale. To resolve this, we re-normalize the convolution filters (with size of height × width × in_channel) in VGG such that the $l_2$ norm of each filter is equal to one. With such re-normalization, all convolution filters have responses with similar ranges, making the computed statistics more comparable. Gatys et al. [9] noticed the same issue, and used a different form of re-scaling such that the average response of each filter over spatial locations and channels is equal to one.

We achieve the discrimination of structure and texture by exploiting two distinct characteristics. First, texture is spatially homogeneous, while structure is more precisely localized in space. Second, the perception of visual texture is scale-dependent. For small-scale visual texture (see Fig. 2 (a) and (b)), a small receptive field (e.g., a $16 \times 16$ window) is able to capture its intrinsic repetitiveness, while for large-scale visual texture that is a combination of small-scale textured surfaces and structural contours (see Fig. 2 (d) and (e)), a large receptive field (e.g., a $128 \times 128$ window) may be needed to sufficiently cover the repeated patterns. Computationally, we use the dispersion index [5] defined by the ratio of variance to mean as the structure/texture indicator. For each stage of VGG, we apply a sliding window approach to compute local dispersion indexes, followed by averaging across channels:

$$Y_s^{(i)} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left( \frac{\sigma_{ij}}{\mu_{ij}} \right)^2,$$
Figure 3: Sample (a) structure and (b) texture patches of size 128×128 in our image patch dataset manually cropped from the Waterloo Exploration Database [21] and the DIV2K dataset [30].

Figure 4: The conditional histograms (normalized to probabilities) of the dispersion index \( y^{(i)}_x \). One can observe that a clear separation between structure and texture at different stages is achieved.

where \( y^{(i)}_x \) and \( \sigma^{(i)}_x \) represent (respectively) the mean and standard deviation of the local feature patch \( x^{(i)}_j \) in \( X^{(i)}_j \). \( c \) is a small positive stabilizing constant. The average operation is legitimate due to the re-normalization of the convolution filters in VGG. Intuitively, texture is often under-dispersed compared with structure, leading to a smaller \( y^{(i)}_x \). As the receptive field of the VGG increases with the number of convolution and sub-sampling layers, we expect an early-stage \( f^{(i)}_x \) to be responsive to small-scale texture, while a late-stage \( \tilde{f}^{(i)}_x \) is responsible for large-scale texture. To verify this, we gather an image patch dataset, which contains 2, 500 structure patches and 2, 500 texture patches of five different sizes (i.e., \( 16 \times 16, 32 \times 32, 64 \times 64, 128 \times 128, \) and \( 256 \times 256 \)). All patches are cropped from the Waterloo Exploration Database [21] and the DIV2K dataset [30], and manually labeled. Fig. 3 shows sample patches of size \( 128 \times 128 \), where we see great variability in structure arrangements and texture appearances. We draw the conditional histograms in Fig. 4, where we find a clear separation between structure and texture at different scales.

We then feed the dispersion index \( y^{(i)}_x \) as a single statistical feature to logistic regression to compute the probability of the given patch being texture:

\[
p_x^{(i)} = p \left( \text{“} \times \text{” is texture} \mid y^{(i)}_x \right) = \frac{1}{1 + e^{-w^{(i)} y^{(i)}_x + b^{(i)}}}, \tag{7}
\]

where \( w^{(i)}, b^{(i)} \) are the weight and bias parameters to be fitted on our image patch dataset.

Fig. 5 shows the multi-scale texture probability maps of the “Farm” image, where a warmer color indicates a higher texture probability. Let us focus on the “hay” in the bottom right of the image. When we rely on \( y^{(i)}_x \) that uses a small receptive field, the hay is classified as rather isolated structure, as reflected in the probability map at the finest scale. When we increase the receptive field (e.g., using \( y^{(3)}_x \) or \( y^{(4)}_x \)), the hay is identified as texture, where the intrinsic repetitiveness is well captured. If we continue increasing the receptive field, the bottom right region containing the hay is classified towards structure again, which makes perfect sense because the receptive field is large enough to include surrounding structural contours (e.g., the boundaries of the hay and the farmhouse). Other small-scale texture such as the sky, the meadow, and
the roof has also been successfully captured by $y^{(i)}_j$ and well reflected in the corresponding probability maps. Similarly, when the receptive field is large enough to include object boundaries, the included texture is part of the structure patch.

**Perceptual Distance Metric.** With the multi-scale texture probability maps at hand, we are ready to design the spatial pooling strategy for combining local structure and texture similarity measurements. As the proposed quality model is full-reference, we are able to compute two set of probability maps, $\{p^x_k\}_{i=1}^{K_i}$ and $\{p^y_k\}_{i=1}^{K_i}$ from the VGG representations of the co-located reference and distorted patches, $x$ and $y$, respectively. For the purpose of quality assessment, we take the minimum of the two texture probabilities:

$$p^{(i)} = \min\left\{p^x_k(i), p^y_k(i)\right\},$$

which is conductive to penalizing the introduced structural artifacts (e.g., JPEG blocking). For the purpose of perceptual optimization (as described in Section 4.2), we may directly use $p^x_k(i)$ as weighting to full respect the reference information.

Finally, we define the A-DISTS index as

$$\text{A-DISTS}(X, Y) = 1 - \frac{1}{N} \sum_{i=0}^{M} \sum_{j=1}^{N_i} S(\tilde{X}^{(i)}_j, \tilde{Y}^{(i)}_j),$$

and

$$S(\tilde{X}^{(i)}_j, \tilde{Y}^{(i)}_j) = \frac{1}{K_i} \sum_{k=1}^{K_i} \left( \hat{p}_k(i) I(\tilde{x}^{(i)}_{j,k}, \tilde{y}^{(i)}_{j,k}) + \hat{q}_k(i) s(\tilde{x}^{(i)}_{j,k}, \tilde{y}^{(i)}_{j,k}) \right),$$

where $N = \sum_{i=0}^{M} N_i$, $\hat{q}_k(i) = 1 - \hat{p}_k(i)$, and $\hat{p}_k(i)$ is the texture probability of the $k$th patch viewed at the $i$th scale. $I(\cdot)$ and $s(\cdot)$ are defined in Eq. (2). A-DISTS ranges from zero to one, with a higher value indicating poorer predicted quality.

## 4 EXPERIMENTS

In this section, we first compare the proposed A-DISTS with a set of full-reference IQA models in term of quality assessment on traditional and novel algorithm-dependent distortions. We then compare A-DISTS against a smaller set of top-performing models in optimization of image super-resolution methods.
Table 1: Performance comparison of A-DISTS against twelve IQA models on four standard IQA databases. Larger PLCC, SRCC, and KRCC numbers represent better performance, with a maximum value of one. Top-2 results are highlighted in bold.

| Method      | LIVE [27] | CSIQ [16] | TID2013 [23] | KADID [18] |
|-------------|-----------|-----------|---------------|------------|
|             | 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 [34]   | 0.937     | 0.948     | 0.796         | 0.852      | 0.865     | 0.680         | 0.777      | 0.727     | 0.545         |
| MS-SSIM [35]| 0.940     | 0.951     | 0.805         | 0.889      | 0.906     | 0.730         | 0.830      | 0.786     | 0.605         |
| VIF [26]    | 0.960     | 0.964     | 0.828         | 0.913      | 0.911     | 0.743         | 0.771      | 0.677     | 0.518         |
| MAD [16]    | 0.968     | 0.967     | 0.842         | 0.950      | 0.947     | 0.797         | 0.827      | 0.781     | 0.604         |
| FSIMc [40]  | 0.961     | 0.965     | 0.836         | 0.919      | 0.931     | 0.769         | 0.877      | 0.851     | 0.667         |
| GMSD [37]   | 0.957     | 0.960     | 0.827         | 0.945      | 0.950     | 0.804         | 0.855      | 0.804     | 0.634         |
| VSI [39]    | 0.948     | 0.952     | 0.806         | 0.928      | 0.942     | 0.786         | 0.900      | 0.897     | 0.718         |
| NLPD [15]   | 0.932     | 0.937     | 0.778         | 0.923      | 0.932     | 0.769         | 0.839      | 0.800     | 0.625         |

A-DISTS (ours) 0.954 0.955 0.812 0.944 0.942 0.796 0.861 0.836 0.642

Table 2: 2AFC score comparison of IQA models on BAPPS and Ding20. It is computed by \( \hat{r} \cdot (1 - r) \cdot (1 - \hat{r}) \), where \( r \) is the ratio of human votes and \( \hat{r} \in \{0, 1\} \) is the preference of an IQA model. A higher score indicates better performance.

| IQA Model          | BAPPS [41] | Ding20 [7] |
|--------------------|------------|------------|
|                    | Colorization | Video  | Frame  | Super-resolution | All         | Denoising | Deblurring | Super-resolution | Compression | All         |
| Human              | 0.688      | 0.671     | 0.686  | 0.734      | 0.695       | 0.761     | 0.843     | 0.833      | 0.891       | 0.890       |
| PSNR               | 0.624      | 0.590     | 0.543  | 0.642      | 0.614       | 0.627     | 0.518     | 0.612      | 0.669       | 0.612       |
| SSIM [34]          | 0.522      | 0.583     | 0.548  | 0.613      | 0.617       | 0.636     | 0.575     | 0.599      | 0.649       | 0.615       |
| MS-SSIM [35]       | 0.522      | 0.589     | 0.572  | 0.638      | 0.596       | 0.623     | 0.568     | 0.665      | 0.665       | 0.628       |
| VIF [26]           | 0.515      | 0.594     | 0.597  | 0.651      | 0.603       | 0.589     | 0.607     | 0.655      | 0.540       | 0.598       |
| MAD [16]           | 0.490      | 0.593     | 0.581  | 0.655      | 0.599       | 0.624     | 0.671     | 0.681      | 0.651       | 0.657       |
| FSIMc [40]         | 0.573      | 0.590     | 0.581  | 0.660      | 0.615       | 0.522     | 0.490     | 0.525      | 0.563       | 0.525       |
| GMSD [37]          | 0.517      | 0.594     | 0.575  | 0.667      | 0.613       | 0.417     | 0.454     | 0.469      | 0.567       | 0.477       |
| VSI [39]           | 0.597      | 0.591     | 0.568  | 0.668      | 0.622       | 0.518     | 0.470     | 0.487      | 0.576       | 0.513       |
| NLPD [15]          | 0.528      | 0.584     | 0.552  | 0.655      | 0.600       | 0.622     | 0.514     | 0.629      | 0.652       | 0.604       |
| PieAPP [25]        | 0.594      | 0.582     | 0.598  | 0.685      | 0.626       | 0.625     | 0.734     | 0.744      | 0.822       | 0.732       |
| LPIPS [41]         | 0.625      | 0.605     | 0.630  | 0.705      | 0.641       | 0.657     | 0.788     | 0.768      | 0.834       | 0.761       |
| DISTS [6]          | 0.627      | 0.600     | 0.625  | 0.710      | 0.651       | 0.602     | 0.790     | 0.704      | 0.833       | 0.725       |
| A-DISTS (ours)     | 0.621      | 0.602     | 0.616  | 0.708      | 0.642       | 0.629     | 0.792     | 0.781      | 0.846       | 0.763       |

generated by image colorization, video deblurring, frame interpolation, and super-resolution algorithms. Ding20 [7] is a byproduct of a perceptual optimization experiment with 880 image pairs generated from four low-level vision tasks - image denoising, deblurring, super-resolution, and compression. Since the human opinions are collected in the two-alternative forced choice (2AFC) experiments, the 2AFC score [41], which quantifies the consistency of model predictions relative to human opinions, is employed as the evaluation criterion. Results in Table 2 show that A-DISTS without reliance on human perceptual scores achieves comparable performance to LPIPS, but is slightly inferior to DISTS on BAPPS. We attribute this to the small patch size (i.e., 64 x 64) of BAPPS, rendering local computation in A-DISTS less effective. For the images with relatively large size in Ding20, A-DISTS outperforms DISTS and the other models. We also test A-DISTS on another four publicly available image restoration databases with human judgements: Liu13 [19], Ma17 [20], Min19 [22], and Tian19 [29], including 1,200 motion-deblurred images, 1,620 super-resolved images, 600 dehazed images, and 140 rendered images based on depth information, respectively. Table 3 shows the correlation results, where one can observe that A-DISTS is best at explaining human data in these datasets. In summary, the proposed A-DISTS achieves better correlation performance than DISTS on all ten databases, except for the patch similarity dataset - BAPPS. This provides strong justifications of the key modifications in A-DISTS: structure and texture separation and locally adaptive weighting.
4.2 Performance on Perceptual Optimization

The application scope of objective IQA models is far beyond evaluating image processing algorithms; they can be used as objectives to guide the algorithm design and optimization. In this subsection, we test the gradient-based optimization performance of A-DISTS against four competing models - MAE, MS-SSIM [35], LPIPS [41], and DISTS [6] in the context of single image super-resolution. We exclude the rest IQA models in Table 1 because they have been empirically shown less competitive on this task [7].

Single image super-resolution aims to generate a high-resolution (HR) and high-quality image from a low-resolution (LR) one. In recent years, DNN-based methods [17, 31, 38, 42] have achieved dominant performance on this task. Here, we adopt the Residual Dense Block (RRDB) network proposed in [31] as the backbone to construct our super-resolution algorithms. Training is performed by optimizing a given IQA model:

$$\ell(\phi) = D(f(X_l; \phi), X_h),$$

(11)

where \( f(\cdot; \phi) \) denotes the RRDB network parameterized by a vector \( \phi \). \( X_l \in \mathbb{R}^K \) is the ground-truth HR image, \( X_h \in \mathbb{R}^{\frac{K}{4}} \) is the input LR image down-sampled by a factor of 4. \( D \) represents the IQA metric, with a lower value indicating higher predicted quality.

We use the DIV2K database [30] and the Waterloo Exploration Database [21] for training and testing, respectively. We generate LR images by downsampling HR images with bicubic interpolation. Following the practice of [7], the model parameters optimized for MAE are employed as the initializations for the networks to be optimized by other models. More training details (e.g., optimizer, learning rate, batch size, etc.) are inherited from [31]. We apply the trained networks to the test images, and conduct a subjective user study for quantitative evaluation. To ensure a fair comparison (i.e., to avoid potential cherry-picking test results), we adopt the debiased subjective assessment method in [4], which automatically samples a small set of adaptive and diverse test images by solving

$$X^* = \arg \max_{X \in \mathcal{X}} D(f_i(X_j), f_j(X_i)), \quad 1 \leq i \leq j \leq 5,$$

(12)

where \( \mathcal{X} \) denotes the set of LR images from the Waterloo Exploration Database [21], \( i \) and \( j \) are the algorithm indices. \( D \) is a measure to approximate the perceptual distance between the super-resolved images \( f_i(X_j) \) and \( f_j(X_i) \). We define \( D \) as the average of two IQA models \( D_1 \) and \( D_2 \) used to optimize \( f_i \) and \( f_j \), respectively. By adding a diversity term [4], we are able to automatically select a small subset of images in \( \mathcal{X} \) that best differentiate between two

![Figure 6: Sample image pairs in our debiased subjective quality assessment. First column: Reference images. Middle column (form top to bottom): Super-resolved images optimized for MAE, MS-SSIM, LPIPS and DISTS, respectively. Last column: Super-resolved images optimized for A-DISTS. See text for more details on image selection.](image)

1To compensate for the scale difference, the values of \( D_1 \) and \( D_2 \) are mapped to the same MOS scale (e.g., LIVE [27]) by fitting a logistic function.
Figure 7: Super-resolution results of three example images optimized for different IQA models.

Table 4: Global ranking of the five IQA models for use in optimizing single image super-resolution methods in the debiased subjective testing [4]. A higher ranking score indicates better performance.

| IQA model | MAE   | MS-SSIM | LPIPS | DISTS | A-DISTS |
|-----------|-------|---------|-------|-------|---------|
| Ranking score | -1.524 | -1.453  | 0.762 | 0.980 | 1.095   |

networks $f_i$ and $f_j$. The comparison is exhausted for all $\binom{5}{2}$ pairs of algorithms. Fig. 6 shows several sample image pairs in our debiased subjective quality assessment experiment.

We employ the 2AFC method for subjective rating. For each algorithm pair, we set 20 images according to Eq. (12). This leads to a total of $\binom{5}{2} \times 20 = 200$ paired comparisons for 5 IQA models. Subjects are required to choose the image with higher perceived quality with reference to the ground-truth image. Subjects are allowed to adjust the viewing distance and zoom in/out any part of the images for careful inspection. We gather data from 20 subjects with general background knowledge of multimedia signal processing. The Bradley–Terry model [3] is adopted to convert paired comparison results to a global ranking, as shown in Table 4. We find that the proposed A-DISTS achieves the best perceptual optimization results on average. The ranking of the remaining models is consistent with the conclusions in [7].

Fig. 7 shows three visual examples of super-resolution methods optimized for different IQA models. Like many other studies, we find MAE and MS-SSIM encourage blurry images. The results by DISTS are generally sharper, but appear distortions in structure regions and noise in texture regions. With locally adaptive structure and texture similarity measurements, A-DISTS generates better visual results with reduced structural artifacts and more plausible textures.

5 CONCLUSION AND DISCUSSION

We have developed a locally adaptive structure and texture similarity index for full-reference IQA. The keys to the success of our approach are 1) the separation of structure and texture across space and scale and 2) the adaptive weighting of quality measurements according to local image content. A-DISTS is free of expensive MOSs for supervised training, correlates well with human data in standard IQA and image restoration databases, and demonstrates competitive optimization performance for single image super-resolution.

One limitation of the proposed A-DISTS is that the performance on global texture-related tasks may be slightly compromised. For example, on the SynTex database [10] for texture similarity, A-DISTS obtains an SRCC of 0.760 compared to 0.923 by DISTS. Therefore, a generalized quality measure that translates in a content-dependent way from DISTS to A-DISTS is worth deeper investigation. Nevertheless, as most natural photographic images are made of “things and stuff”, we believe the proposed A-DISTS holds much promise for use in a wide range of real-world image processing applications.

ACKNOWLEDGEMENTS

This work was supported in part by Hong Kong RGC Early Career Scheme (No. 21213821 to KDM).

REFERENCES

[1] Edward H. Adelson. 2001. On seeing stuff: The perception of materials by humans and machines. In SPIE Human Vision and Electronic Imaging. 1–12.
[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2018. Deep neural networks for no-reference and full-reference
