International Journal of Computer Vision (2021) 129:1258–1281
https://doi.org/10.1007/s11263-020-01419-7

Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems

Keyan Ding[1] · Kede Ma[1] · Shiqi Wang[1] · Eero P. Simoncelli[2]

Received: 1 May 2020 / Accepted: 8 December 2020 / Published online: 21 January 2021

Abstract
The performance of objective image quality assessment (IQA) models has been evaluated primarily by comparing model predictions to human quality judgments. Perceptual datasets gathered for this purpose have provided useful benchmarks for improving IQA methods, but their heavy use creates a risk of overfitting. Here, we perform a large-scale comparison of IQA models in terms of their use as objectives for the optimization of image processing algorithms. Specifically, we use eleven full-reference IQA models to train deep neural networks for four low-level vision tasks: denoising, deblurring, super-resolution, and compression. Subjective testing on the optimized images allows us to rank the competing models in terms of their perceptual performance, elucidate their relative advantages and disadvantages in these tasks, and propose a set of desirable properties for incorporation into future IQA models.

Keywords Image quality assessment · Perceptual optimization · Performance evaluation

1 Introduction

The goal of objective image quality assessment (IQA) is the construction of computational models that predict the perceived quality of visual images. IQA models are generally classified according to their reliance on the availability of an original reference image. Full-reference methods compare a distorted image to the complete reference image, reduced-reference methods require only partial information about the reference image, and no-reference (or blind) methods operate solely on the distorted image. The standard paradigm for testing IQA models is to compare them to human quality ratings of distorted images, which have been made available in datasets such as LIVE (Sheikh et al. 2006) and TID2013 (Ponomarenko et al. 2015). However, excessive reuse of these test sets during IQA model development may lead to overfitting, and as a consequence, poor generalization to images corrupted by distortions that are not present in the test sets (see Table 4).

A highly promising but relatively under-studied application of IQA measures is to use them as objectives for the design and optimization of new image processing algorithms. The parameters of image processing methods are usually adjusted to minimize the mean squared error (MSE), the simplest of all fidelity metrics, despite the fact that it has been widely criticised for its poor correlation with human perception of image quality (Girod 1993). Early attempts at perceptual optimization using the *structural similarity* (SSIM) index (Wang et al. 2004) in place of MSE achieved perceptual gains in applications of image restoration (Channappayya et al. 2008), wireless video streaming (Vukadinovic and Karlsson 2009), video coding (Wang et al. 2011), and image synthesis (Snell et al. 2017). A recent publication used perceptual measures based on pre-trained deep neural networks (DNNs) for optimization of super-resolution results (Johnson et al. 2016), although these have not been tested against human judgments.

Communicated by Daniel Scharstein.

✉ Kede Ma
kede.ma@cityu.edu.hk

Keyan Ding
keyan.ding@my.cityu.edu.hk

Shiqi Wang
shiqwang@cityu.edu.hk

Eero P. Simoncelli
eero.simoncelli@nyu.edu

[1] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

[2] Howard Hughes Medical Institute, Center for Neural Science, and Courant Institute of Mathematical Sciences, New York University, New York, USA

In this paper, we systematically evaluate a large set of full-reference IQA models in the context of perceptual optimization. To determine their suitability for optimization, we first test the models on recovering a reference image from a given initialization by optimizing the model-reported distance to the reference. For many IQA methods, we find that the optimization does not converge to the reference image, and can generate severe distortions. These optima are either local, or global but non-unique. We select eleven optimization-suitable IQA models as perceptual objectives, and use them to optimize DNNs for four low-level vision tasks-image denoising, blind image deblurring, single image super-resolution, and lossy image compression. Extensive human perceptual tests on the optimized images reveal the relative performance of the competing models. Moreover, inspection of their visual failures indicates limitations in model design, providing guidance for the development of future IQA models.

2 Taxonomy of Full-Reference IQA Models

Full-reference IQA methods can be broadly classified into five categories:

- *Error visibility methods* apply a distance measure directly to pixels (e.g., MSE), or to transformed representations of the images. The MSE in particular possesses useful properties for optimization (e.g., differentiability and convexity), and when combined with linear-algebraic tools, analytical solutions can often be obtained. For example, the classical solution to the MSE-optimal denoising problem (assuming a translation-invariant Gaussian signal model) is the Wiener filter (Wiener 1950). Given that MSE in the pixel domain is poorly correlated with perceived image quality, many IQA models operate by first mapping images to more perceptually appropriate representations (Safranek and Johnston 1989; Daly 1992; Lubin 1993; Watson 1993; Teo and Heeger 1994; Watson et al. 1997; Larson and Chandler 2010; Laparra et al. 2016), and measuring MSE within that space.
- *Structural similarity (SSIM) methods* are constructed to measure the similarity of local image "structures", often using correlation measures. The prototype is the SSIM index (Wang et al. 2004), which combines similarity measures of three conceptually independent components—luminance, contrast and structure. It has become a *de facto* standard in the field of perceptual image processing, and provided a prototype for subsequent IQA models based on feature similarity (Zhang et al. 2011), gradient similarity (Liu et al. 2012a), edge strength similarity (Zhang et al. 2013), and saliency similarity (Zhang et al. 2014).
- *Information-theoretic methods* measure some approximation of the mutual information between the perceived reference and distorted images. Statistical modeling of the image source, the distortion process, and the human visual system (HVS) is critical in algorithm development. A prototypical example is the visual information fidelity (VIF) measure (Sheikh and Bovik 2006).
- *Learning-based methods* learn a metric from a training set of images and corresponding perceptual distances using supervised machine learning methods. By leveraging the power of DNNs, these methods have achieved state-of-the-art performance on existing image quality databases (Bosse et al. 2018; Prashnani et al. 2018). But given the high dimensionality of the input space (i.e., millions of pixels), these methods are prone to overfitting the limited available data. Strategies that compensate for the insufficiency of labeled training data include building on pre-trained networks (Zhang et al. 2018; Ding et al. 2020), training on local image patches (Bosse et al. 2018), and combining multiple IQA databases (Zhang et al. 2019b).
- *Fusion-based methods* combine existing IQA methods to build a "super-evaluator" that exploits the diversity and complementarity of their constituent methods (analogous to "boosting" methods in machine learning). Fusion combinations can be determined empirically (Ye et al. 2014) or learned from data (Liu et al. 2012b; Ma et al. 2019). Some methods incorporate deterministic or statistical image priors to regularize an IQA measure (Jordan 1881; Ulyanov et al. 2018). Since such regularizers can be seen as a form of no-reference IQA measures (Wang and Bovik 2011), we also view these as fusion solutions.

3 Screening of Full-Reference IQA Models for Perceptual Optimization

We used a naïve task to demonstrate the issues encountered when using IQA models in gradient-based perceptual optimization. This task also allows us to pre-screen existing models, and to motivate the design of experiments used in subsequent comparisons.

3.1 Reference Image Recovery

Given a reference (undistorted) image x and an initial image y_0, we aimed to recover x by numerically optimizing

$$y^\star = \arg\min_y D(x, y), \tag{1}$$

where D denotes a full-reference IQA measure with a lower score indicating higher predicted quality, and $y^\star$ is the recov-

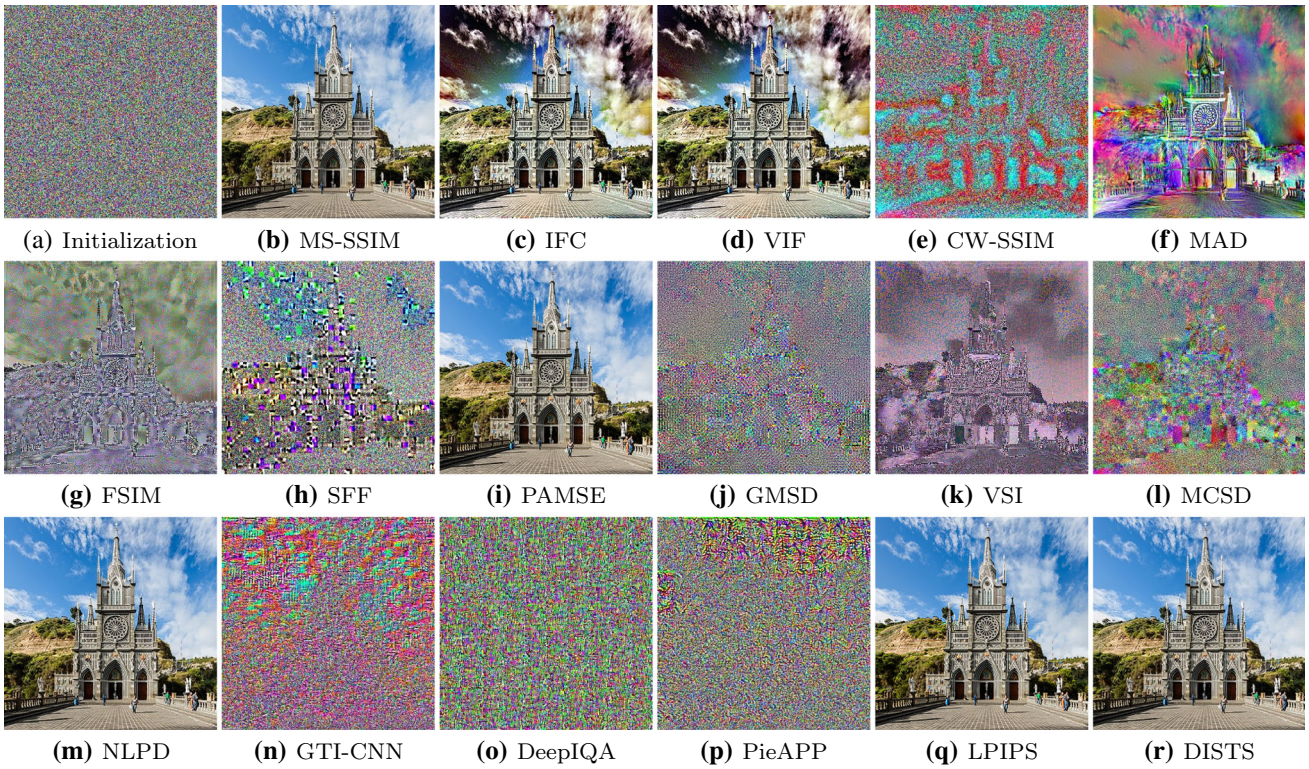


Fig. 1 Reference image recovery test. Starting from (**a**) a white Gaussian noise image, we recover images by optimizing the predicted quality relative to a reference image, using different IQA models (**b**)–(**r**)

ered image. For example, if D is MSE, the (trivial) analytical solution is $y^\star = x$, indicating full recoverability. The majority of current IQA models are continuous and differentiable, and solutions must be obtained numerically using gradient-based iterative solvers. We considered an initial set of 17 methods, which we believe cover the full spectrum of full-reference IQA methods. These include three error visibility methods—MAD (Larson and Chandler 2010), PAMSE (Xue et al. 2013) and NLPD (Laparra et al. 2016), seven structural similarity methods—MS-SSIM (Wang et al. 2003), CW-SSIM (Wang and Simoncelli 2005), FSIM (Zhang et al. 2011), SFF (Chang et al. 2013), GMSD (Xue et al. 2014) and VSI (Zhang et al. 2014), MCSD (Wang et al. 2016), two information-theoretical methods—IFC (Sheikh et al. 2005) and VIF (Sheikh and Bovik 2006), and five DNN methods—GTI-CNN (Ma et al. 2018), DeepIQA (Bosse et al. 2018), PieAPP (Prashnani et al. 2018), LPIPS (Zhang et al. 2018) and DISTS (Ding et al. 2020). As this paper focuses on the perceptual optimization performance of individual IQA measures, fusion-based methods are not included.

Figures 1 and 2 show recovery results from two different initializations—a white Gaussian noise image and a JPEG-compressed version of a reference image, respectively. For all IQA methods, the optimization converges to a final image with a substantially better score than that of the initial image. Models based on injective mappings such as MS-SSIM, PAMSE, NLPD and DISTS are able to recover the reference image (although the rate of convergence may depend on the choice of initial image). Many of the remaining IQA models generate a final image with worse visual quality than that of the initial image (e.g., compare Fig. 2(**a**) with (o) or (p)), often with noticeable model-dependent artifacts. This is because these methods rely on surjective mapping functions to transform the images to a reduced "perceptual" space for quality computation. For example, GTI-CNN (Ma et al. 2018) uses a surjective DNN with four stages of convolution, subsampling, and halfwave rectification. The resulting undercomplete representation is optimized for geometric transformation invariance, at the cost of significant information loss. The examples demonstrate that preservation of some aspects of this lost information is important for perceptual quality. Similar arguments can be applied to other surjective DNN-based IQA models, such as DeepIQA (Bosse et al. 2018) and PieAPP (Prashnani et al. 2018). Generally, optimization guided by the surjective models "recovers" more structures when initialized with the JPEG image (which provides roughly correct local luminances), as compared to initialization with purely white Gaussian noise.

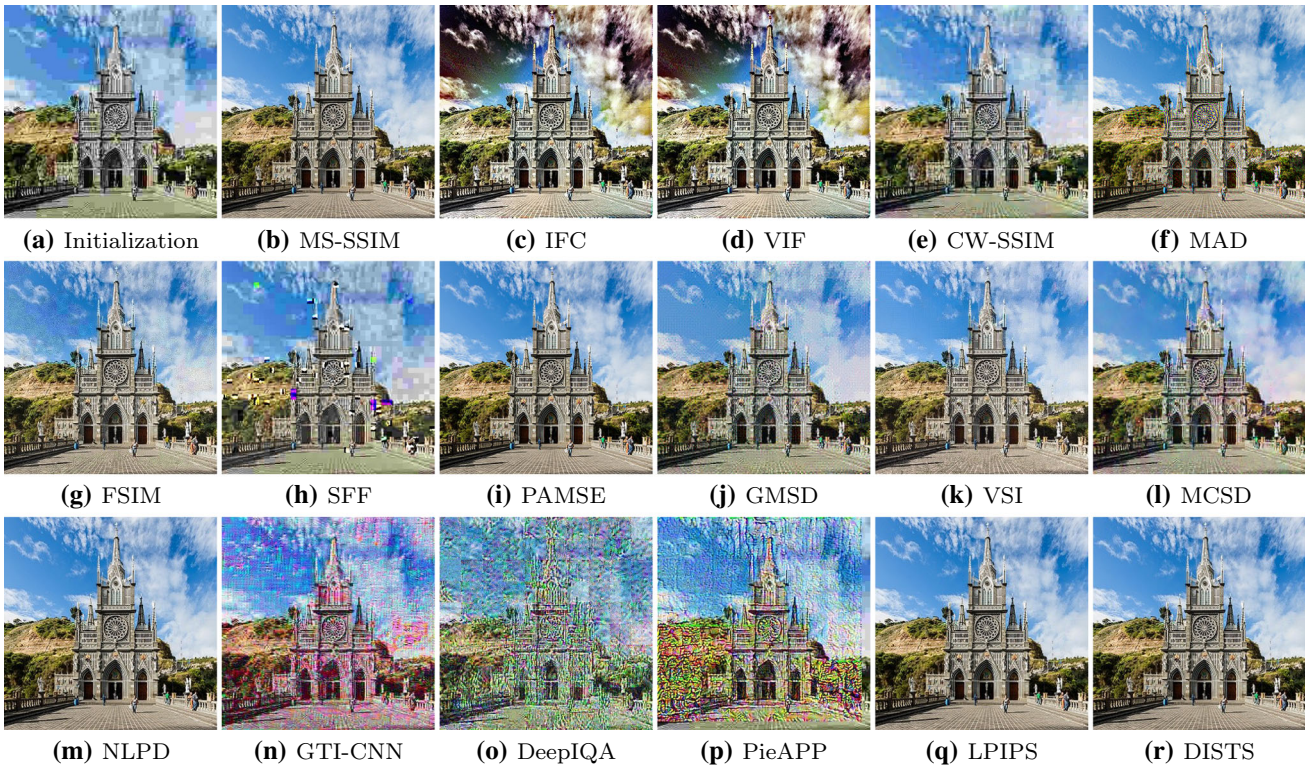


Fig. 2 Reference image recovery test. Staring from (**a**) a JPEG compressed version of a reference image, we recover images by optimizing the predicted quality relative to the reference image, using different IQA models (**b**)–(**r**)

3.2 IQA Model Selection

The reference image recovery test results were used to prescreen the initial set of IQA models, excluding those that perform poorly (due to surjectivity). In addition, we excluded models with similar designs. This process yielded 11 full-reference IQA models to be compared in our human subject evaluations:

1. MAE, the Mean Absolute Error (ℓ_1-norm) of pixel values, has been frequently adopted in optimization, despite its poor perceptual relevance. MAE has been shown to consistently outperform MSE (ℓ_2-norm) in image restoration tasks (Zhao et al. 2016).
2. MS-SSIM (Wang et al. 2003), the Multi-Scale extension of the SSIM index (Wang et al. 2004), provides more exibility than single-scale SSIM, allowing for a wider range of viewing distances. It decomposes the input images into Gaussian pyramids (Burt and Adelson 1983), and computes contrast and structure similarities at each scale and luminance similarity at the coarsest scale only. MS-SSIM has become a standard "perceptual" quality measure, and has been used to guide the design of DNN-based image super-resolution (Zhao et al. 2016; Snell et al. 2017) and compression (Ballé et al. 2018) algorithms.
3. VIF (Sheikh and Bovik 2006), the Visual Information Fidelity measure, quantifies how much information from the reference image is preserved in the distorted image. A Gaussian scale mixture (Portilla et al. 2003) is used as a source model to summarize natural image statistics, and mutual information is estimated assuming only signal attenuation and additive noise perturbations. A distinct property of VIF relative to other IQA models is that it can handle cases in which the "distorted" image is visually superior to the reference (Wang et al. 2015).
4. CW-SSIM (Wang and Simoncelli 2005), the Complex Wavelet SSIM index, is designed to be robust to small geometric distortions such as translation and rotation. The construction allows for consistent local phase shifts of wavelet coefficients, which preserves image features. CW-SSIM addresses a common limitation of IQA methods that require precise spatial registration of the reference and distorted images.
5. MAD (Larson and Chandler 2010), the Most Apparent Distortion measure, explicitly models adaptive strategies of the HVS. Specifically, a detection-based strategy considering local luminance and contrast masking is employed for near-threshold distortions, and an appearance-based strategy involving local spatial-frequency statistics is activated for supra-threshold distortions. The two strategies are combined by a weighted geomet-

ric mean, where the weight is determined based on the amount of distortion.

6. FSIM (Zhang et al. 2011), the Feature SIMilarity index, assumes that HVS understands an image mainly according to its low-level features. It computes quality estimates based on phase congruency (Kovesi 1999) as the primary feature, and incorporates the gradient magnitude as the complementary feature. Moreover, the phase congruency component serves as a local weighting factor to derive an overall quality score. FSIM also supplies a color version by making quality measurements from chromatic components.
7. GMSD (Xue et al. 2014), the Gradient Magnitude Similarity Deviation, focuses on computational efficiency of quality prediction, by simply computing pixel-wise gradient magnitude similarity followed by standard deviation (std) pooling. This pooling strategy is, however, problematic because an image with large but constant local distortion yields an std of zero (indicating the best predicted quality).
8. VSI (Zhang et al. 2014), the Visual Saliency Induced quality index, assumes that the change of salient regions due to image degradation is closely related to the change of visual quality. The saliency map is used not only as a quality feature, but also as a weighting function to characterize the importance of a local region. By combining saliency magnitude, gradient magnitude and chromatic features, VSI demonstrates good quality prediction performance, especially for localized distortions, such as local patch substitution (Ponomarenko et al. 2015).
9. NLPD (Laparra et al. 2016), the Normalized Laplacian Pyramid Distance, mimics the nonlinear transformations of the early visual system: local luminance subtraction and local gain control, and combines these values using weighted ℓ_p-norms. The parameters are optimized to minimize the representation redundancies, instead of matching human judgments. NLPD has been successfully employed to optimize image rendering algorithms (Ma et al. 2015; Laparra et al. 2017), where the input reference image has a much higher dynamic range than that of the display. It has also been used to optimize a compression system (Ballé et al. 2016).
10. LPIPS (Zhang et al. 2018), the Learned Perceptual Image Patch Similarity model, computes the Euclidean distance between deep representations of two images. The authors showed that feature maps of different DNN architectures have "reasonable" effectiveness in accounting for human perception of image quality. As LPIPS has many different configurations, we chose the default one based on the VGG network (Simonyan and Zisserman 2015) with the weights learned from the BAPPS dataset (Zhang et al. 2018). VGG-based LPIPS can be seen as a generalization of the "perceptual loss" (Johnson et al. 2016), which computes the Euclidean distance on convolution responses from one stage of VGG.
11. DISTS (Ding et al. 2020), the Deep Image Structure and Texture Similarity metric, is explicitly designed to tolerate texture resampling (e.g., replacing one patch of grass with another). DISTS is based on an injective mapping function built from a variant of the VGG network, and combines SSIM-like structure and texture similarity measurements between corresponding feature maps of the two images. It is sensitive to structural distortions but at the same time robust to texture resampling and modest geometric transformations.

We re-implemented all 11 of these models using PyTorch,[1] and verified that our code could reproduce the published performance results for each model on the LIVE (Sheikh et al. 2006), CSIQ (Larson and Chandler 2010), and TID2013 (Ponomarenko et al. 2015) databases (see Table 2 in "Appendix 1"). We modified grayscale-only models to accept color images, by computing scores on RGB channels separately and averaging them to obtain an overall quality estimate.

4 Perceptual Optimization of Standard Image Processing Tasks

We used each of the 11 full-reference IQA models to guide the learning of DNNs to solve four low-level vision tasks:

- image denoising,
- blind image deblurring,
- single image super-resolution,
- lossy image compression.

The parameters of each network are optimized to minimize an IQA measure over a database of corrupted and original image pairs via stochastic gradient descent. Implementations of all IQA models, as well as the DNNs for the four tasks, are available at https://github.com/dingkeyan93/IQA-optimization.

4.1 Image Denoising

Image denoising is a core application of classical image processing, and also plays an essential role in testing prior models of natural images. In its simplest form, one aims to recover an unknown clean image $x \in \mathbb{R}^N$ from an observed image y that has been corrupted by additive white Gaussian noise n of known variance σ^2, i.e., $y = x+n$. Denoising algorithms can be roughly classified into spatial domain methods

[1] https://pytorch.org.

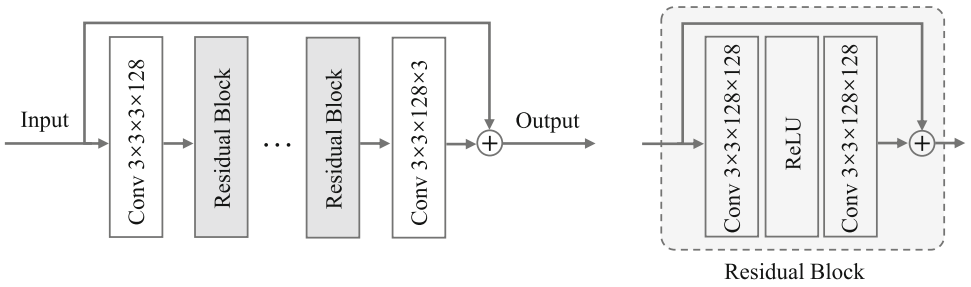
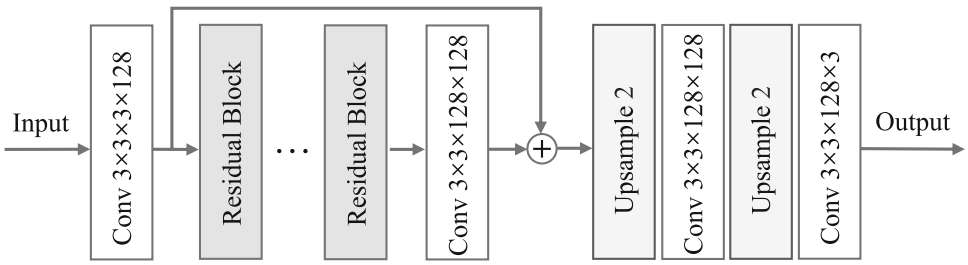


Fig. 3 Network architecture used for denoising and deblurring. In addition to initial and final convolutional blocks, it contains 16 residual blocks, each consisting of two convolutions and a halfwave rectifier (ReLU). Conv $h \times w \times c_{\text{in}} \times c_{\text{out}}$ indicates affine convolution with filter size $h \times w$, over c_{in} input channels, producing c_{out} output channels

Fig. 4 Network architecture used for super-resolution, containing 16 residual blocks followed by two upsampling modules, each composed of an upsampler (factor of 2, using nearest-neighbor interpolation) and a convolution



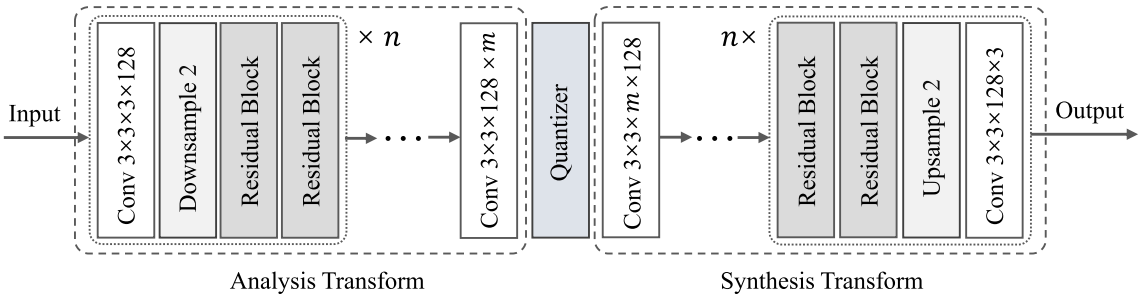


Fig. 5 Network architecture used for lossy image compression, which includes an analysis transformation f_a, a quantizer Q, and a synthesis transformation f_s. f_a is comprised of n blocks, each with a convolution and downsampling (stride) by 2 followed by two residual blocks. After the last block, another convolution layer with m filters is added to produce the internal code representation, the values of which are then quantized by Q. f_s consists of a cascade that is mirror-symmetric to f_a, with nearest-neighbor interpolation used to upsample the feature maps

[e.g., Wiener filter (Wiener 1950), bilateral filter (Tomasi and Manduchi 1998) and collaborative filtering (Dabov et al. 2007)], and wavelet transform methods (Donoho and Johnstone 1995; Simoncelli and Adelson 1996; Portilla et al. 2003). Adaptive sparsifying transforms (Elad and Aharon 2006) and variants of nonlinear shrinkage functions have also been directly learned from natural image data (Hel-Or and Shaked 2008; Raphan and Simoncelli 2008). In recent years, purely data-driven models based on DNNs have achieved state-of-the-art levels of performance (Zhang et al. 2017).

Here, we constructed a simplified DNN, shown in Fig. 3, inspired by the EDSR network (Lim et al. 2017). The network was trained to estimate the noise (which is then subtracted from the observation to yield a denoised image), by minimizing a loss function defined as

$$\ell(\phi) = D\left(y - f_\phi(y), x\right), \tag{2}$$

where D is an IQA measure and $f_\phi : \mathbb{R}^N \mapsto \mathbb{R}^N$ is the mapping of the DNN, parameterized by vector ϕ.

4.2 Blind Image Deblurring

The goal of image deblurring is to restore a sharp image x from a blurry observation y, which can occur due to defocus and motion of the camera, and motion of objects in a scene. The observation process is usually described by

$$y = Kx + n, \tag{3}$$

where $K \in \mathbb{R}^{N\times N}$ denotes a spatially-varying linear kernel. Blind deblurring refers to the problem in which the blur kernel is unknown. Most early methods, e.g., the classical Lucy–Richardson algorithm (Richardson 1972; Lucy 1974), focused on non-blind deblurring where the blur kernel is assumed known. Successful blind deblurring methods, such as (Fergus et al. 2006; Pan et al. 2016), rely heavily on statistical priors of natural images and geometric priors of blur kernels. With the success of deep learning, many DNN-based approaches (Tao et al. 2018; Kupyn et al. 2018) attempt to directly learn the mapping function for blind deblurring without explicitly estimating the blur kernel. Here we also adopted this "kernel-free" approach to train a DNN for image deblurring in an end-to-end fashion. We employed the same network architecture used in denoising (see Fig. 3) with the same loss function (Eq. (2)).

4.3 Single Image Super-Resolution

Single image super-resolution aims to enhance the resolution and quality of a low-resolution image, which can be modelled by

$$y = PKx + n, \tag{4}$$

where P denotes downsampling by a factor of β. This is an ill-posed problem, as downsampling is a projection onto a lower-dimensional subspace, and its solution must rely on some form of regularization or prior model. Early attempts exploited sampling theory (Li and Orchard 2001) or natural image statistics (Sun et al. 2008). Later methods focused on learning mapping functions between the low-resolution and high-resolution images through sparse coding (Yang et al. 2010), locally linear regression (Timofte et al. 2013), self-exemplars (Huang et al. 2015), etc. Since 2014, DNN-based methods have come to dominate this field as well (Dong et al. 2014). An efficient method of constructing a DNN-based mapping is to first extract features from the low-resolution input and then upscale them with sub-pixel convolution (Shi et al. 2016; Lim et al. 2017). Here, we followed this method in constructing a DNN-based function $f : \mathbb{R}^{\left\lfloor \frac{N}{\beta^2} \right\rfloor} \mapsto \mathbb{R}^N$, with architecture specified in Fig. 4. The loss is specified by

$$\ell(\phi) = D\left(f_\phi(y), x\right). \tag{5}$$

4.4 Lossy Image Compression

Data compression involves finding a more compact data representation from which the original image can be reconstructed. Compression can be either lossless or lossy. Here we followed a prevailing scheme in lossy image compression—transform coding, which consists of transformation, quantization, and entropy coding. Traditional image compression methods (e.g., the most widely used standard—JPEG) used a fixed linear transform for all bit rates. More recently, many researchers have demonstrated the visual benefits of nonlinear transforms, especially DNN-based learnable ones that are capable of adapting their parameters to different bitrate budgets. In this paper, we constructed two DNNs for analysis and synthesis transforms, respectively, as shown in Fig. 5. The analysis transform f_a maps the image to a latent feature vector z, whose values are then quantized to L levels with the centers being $\{c_1, \ldots, c_L\}$, where $c_i \in \mathbb{R}$ for $i = 1, \ldots, L$. This quantized representation $\bar{z} = Q(f_a(x))$, is fed to the synthesis transform f_s to reconstruct the compressed image: $y = f_s(\bar{z})$. The quantizer has zero gradients almost everywhere (and infinite gradients at the transitions), which prevents training via gradient descent (Ballé et al. 2017). Hence, we used a soft differentiable approximation (Mentzer et al. 2018)

$$\bar{z}_i = Q(z_i) = \sum_{j=1}^{L} \frac{\exp\left(-s(z_i - c_j)^2\right)}{\sum_{k=1}^{L} \exp\left(-s(z_i - c_k)^2\right)} c_j \tag{6}$$

to backpropagate gradients during training, where the scale parameter s controls the degree to which $Q(\cdot)$ approximates quantization.

Fig. 6 Test images (from the validation set of DIV2K) used in the subjective experiment

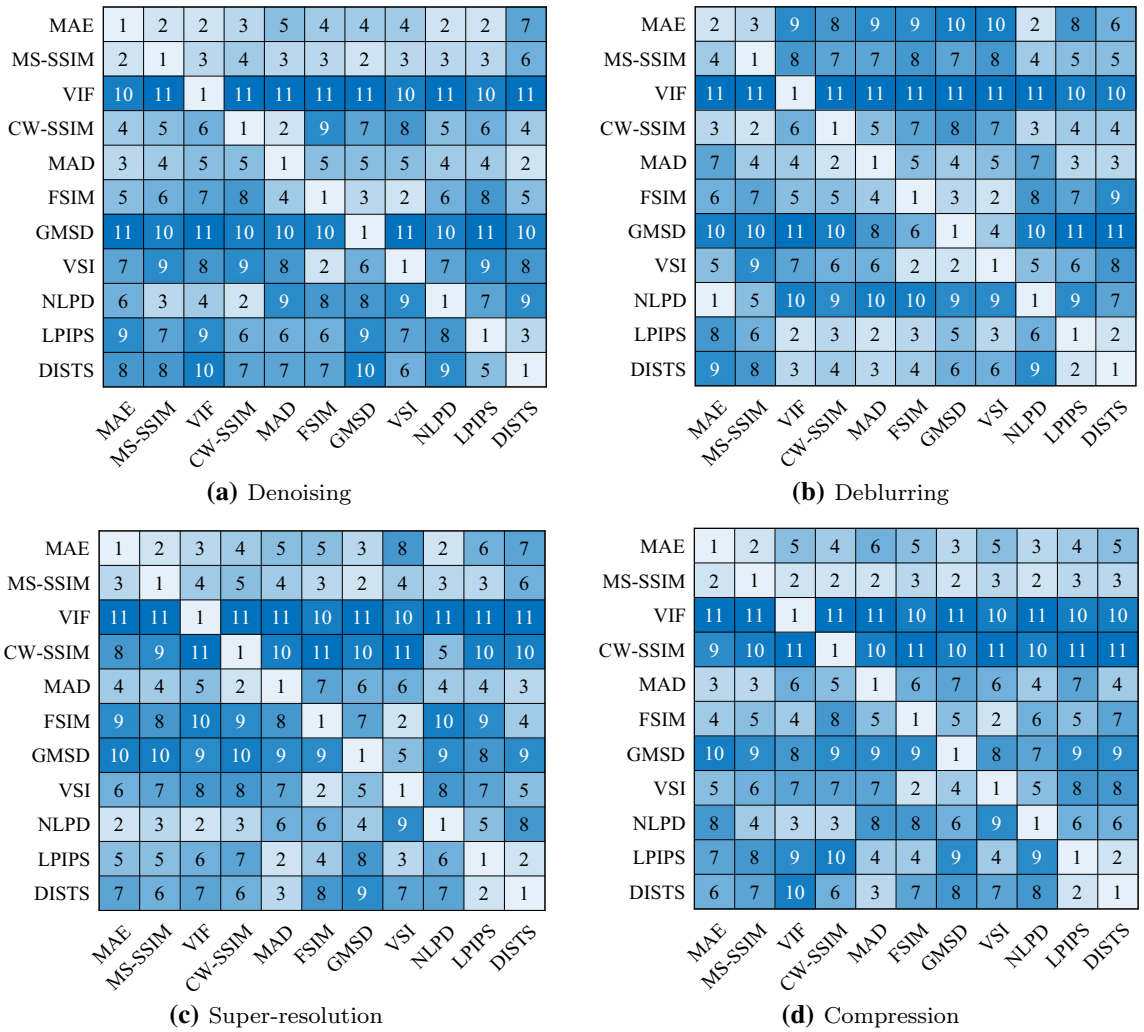


Fig. 7 Objective ranking of the final results in the four tasks. Vertical axis indicates IQA models used to train the networks, and horizontal axis indicates IQA models used to evaluate performance. The numbers of 1–11 indicate the rank order from the best to the worst

In lossy image compression, the objective function is a weighted sum of two terms that quantify the coding cost and the reconstruction error, respectively:

$$\ell = \lambda H[\bar{z}] + \mathbb{E}[D(y, x)]. \tag{7}$$

The first term is typically the entropy of the discrete codes $\bar{z}$, which provides a lower bound on the bitrate for transmitting the quantized coefficients (Shannon 1948). The second term is the distortion between the reconstructed image y and the original image x, as quantified by the full-reference IQA model D. The Lagrange multiplier λ controls the rate-distortion trade-off. Due to substantially different scales of IQA model values, λ should be adjusted for each model in order to enable fair comparison at similar bitrates, an extremely time-intensive process. To avoid this, following Agustsson et al. (2019), we set $\lambda = 0$ in Eq. (7), and controlled an upper bound on bitrate

$$H(\bar{z}) \leq \dim(\bar{z}) \log_2(L) \tag{8}$$

by adjusting the architecture of f_s (i.e., the dimension of $\bar{z}$) and the number of quantization levels L in Q. This elimination of the entropy from the objective also means that we did not need to continually re-estimate the probability mass function $P(\bar{z})$, which varies with changes in the network parameters. The optimization objective in Eq. (7) is reduced to

$$\ell(\phi, \psi) = \mathbb{E}\left[D\left(f_{s,\psi}\left(Q\left(f_{a,\phi}(x)\right)\right), x\right)\right], \tag{9}$$

where ϕ and ψ are the parameters of f_a and f_s, respectively. The expectation is approximated by averaging over mini-batches of training images.

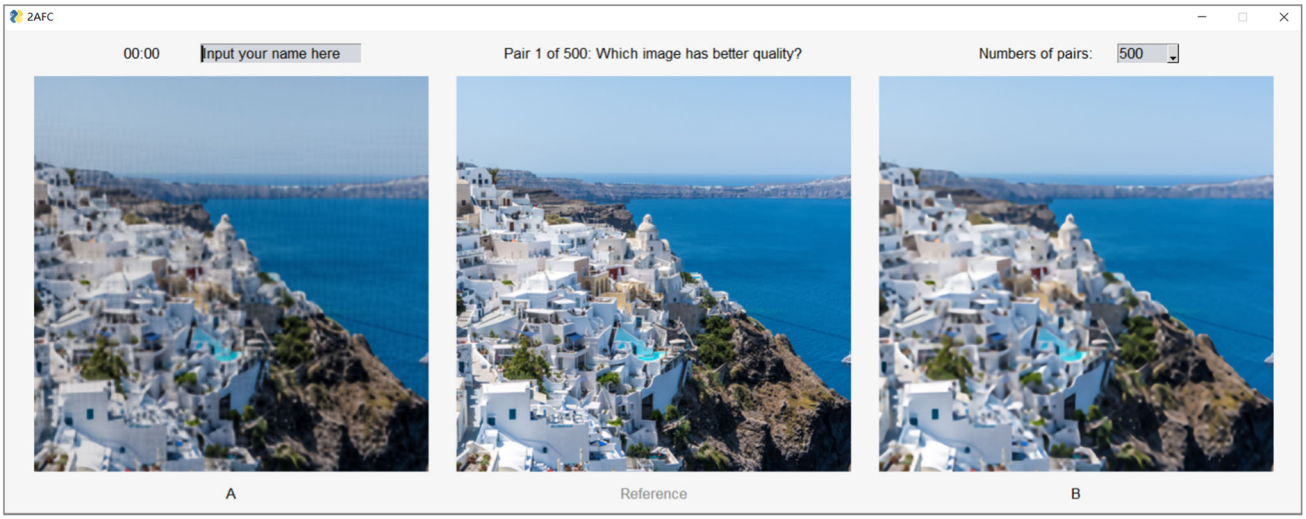


Fig. 8 Customized graphical user interface for subjective testing

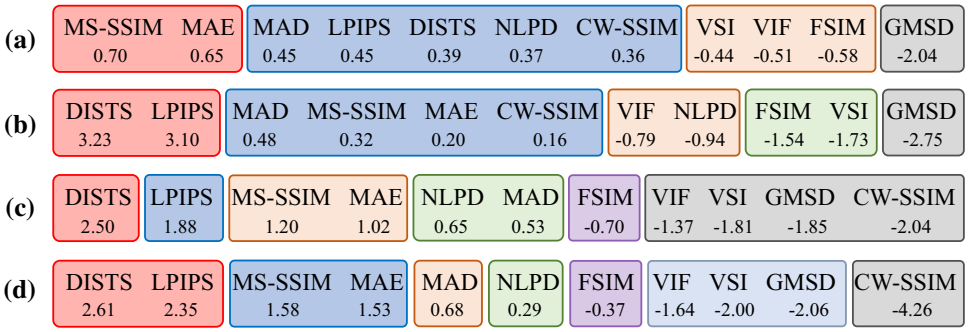


Fig. 9 Subjective ranking of the final results in the four tasks, based on human opinion scores. (**a**) Denoising, (**b**) deblurring, (**c**) super-resolution, (**d**) compression. The optimization performance of IQA models is ranked in the descending order from left to right. Below each model is the global ranking score (larger is better). Models with the same colored box have statistically insignificant performance

5 Implementation Issues

In this section, we describe in detail the training of our DNN-based computational models for the four low-level vision tasks, and the subjective testing procedure used to collect human ratings of the optimized images.

5.1 Model Training

For denoising, we fixed the noise std to $\sigma = 50$ (relative to pixel values in the range $[0, 255]$). For deblurring, we simulated various kernels with different motion patterns and blur levels as in Kupyn et al. (2018). For super-resolution, we generated low-resolution images by downsampling high-resolution images by a factor of $\beta = 4$ using bicubic interpolation. For compression, we set the number of quantization levels to $L = 2$ with centers $\{-1, 1\}$, the quantization scale parameter to $s = 1$, the number of downsampling stages to $n = 4$, and the number of output channels of f_a to $m = 64$. This leads to a maximum of $\frac{H(\bar{z})}{W \times H} \leq \frac{W \times H}{2^4 \cdot 2^4} \cdot 64 \cdot \log_2(2)/(W \times H) = 0.25$ bits per pixel (bpp).

We chose the 4744 high-quality images in the Waterloo Exploration Database (Ma et al. 2017b) as reference images. Training was performed in two stages. In the first stage, we pre-trained a network using MAE as the loss function for all four tasks (Wang et al. 2018). In the second stage, we fine-tuned the network parameters by optimizing the desired IQA model. Pre-training brings several advantages. First, some IQA models are sensitive to initializations (e.g., CW-SSIM, MAD, FSIM, GMSD, and VSI) and pre-training yields more reasonable optimization results (also validated in the task of reference image recovery). Second, models that require backpropagating gradients through multiple stages of computation (e.g., LPIPS and DISTS) converge much faster. Third, it helps us to test whether the recently proposed IQA models lead to consistent perceptual gains on top of MAE, a special case of the simple ℓ_p-norm distance.

For each training stage of the four tasks, we used the Adam optimization package (Kingma and Ba 2015) with a mini-

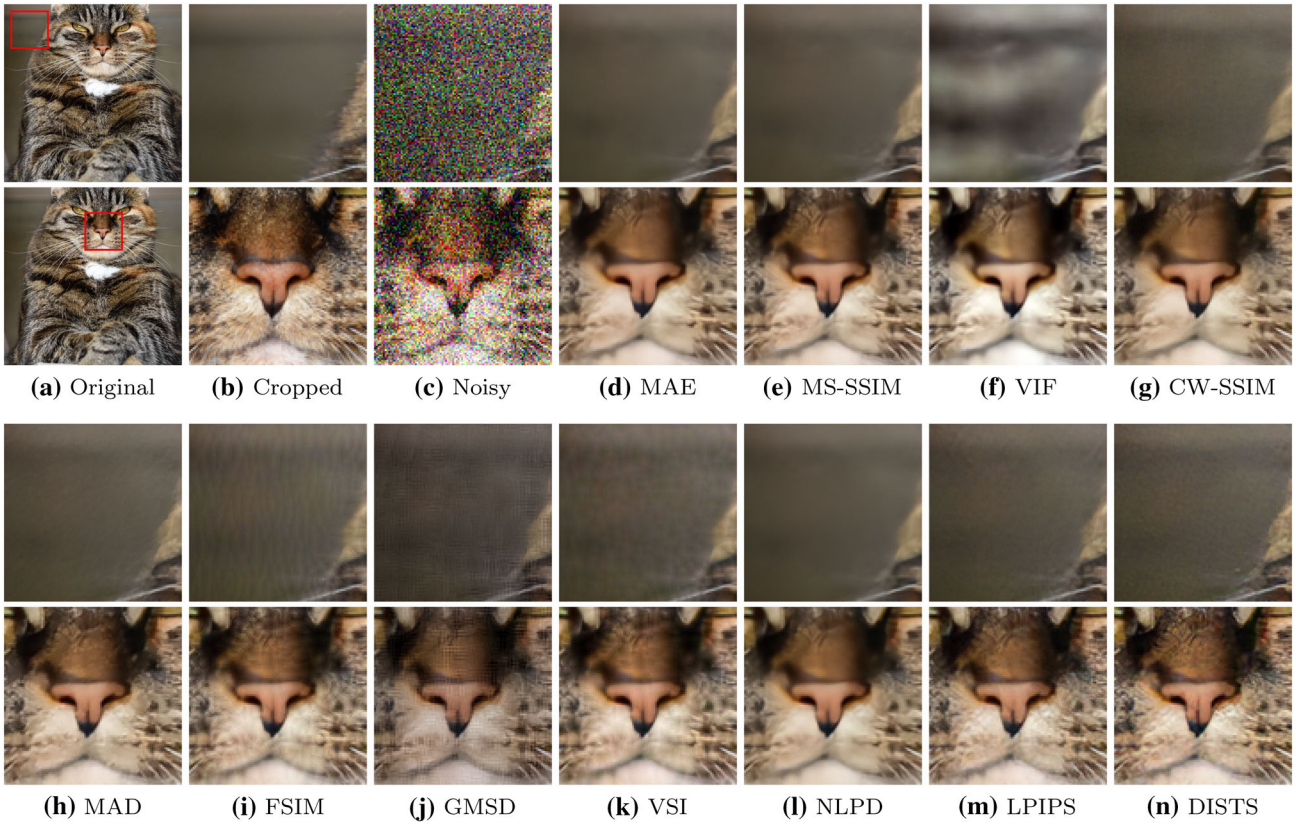


Fig. 10 Denoising results on two regions cropped from an example image, using a DNN optimized for different IQA models

batch size of 16 and an initial learning rate of 10^{-4}, which decays linearly by a factor of 2 for every 100K iterations, and we set the maximum number of iterations to 500K. We randomly extracted patches with the size of $192 \times 192 \times 3$ during training, and tested on 20 independent images selected from the DIV2K validation set (see Fig. 6). Training took roughly 1000 GPU hours (measured using an NVIDIA GTX 2080 device) for a total of $4 \times 11 = 44$ models. Special treatments (i.e., gradient clipping and a smaller learning rate) were given to FSIM and VSI, otherwise their losses are difficult to converge according to our trials.

Generally, it can be difficult to stabilize the training of DNNs to convergence, especially given that the gradients of different IQA models exhibit idiosyncratic behaviors. Fortunately, a simple criterion exists to test the validity of the optimization results: for a given low-level vision task, the DNN optimized for the IQA measure D_i should produce the best result (averaged over an independent set of images) in terms of D_i itself, when comparing to DNNs optimized for $\{D_j\}_{j \neq i}$. Figure 7 shows the ranking of results generated by networks optimized for each of the 11 IQA models (corresponding to one column in one subfigure) on the DIV2K validation set (Timofte et al. 2017), where 1 and 11 indicate the best and worst rankings, respectively. By inspecting the diagonal elements of the four matrices, we conclude that 43 out of 44 models satisfy the criterion, verifying the rationality of our training procedures. The only exception is when MAE is the optimization goal and NLPD (Laparra et al. 2016) is the evaluation measure for the deblurring task. Nevertheless, MAE ranks its own results the second place. As shown in Sect. 6.2, the resulting images from MAE and NLPD look visually similar.

5.2 Subjective Testing

We conducted an experiment to acquire human perceptual comparisons of the IQA optimized results. A two-alternative forced choice (2AFC) method was employed, allowing differentiation of fine-grained quality variations. On each trial, subjects were shown two images optimized according to two different IQA methods, presented on the left and right side of the corresponding reference image (see Fig. 8). Subjects were asked to choose which of the two images had better quality. Subjects were allowed unlimited viewing time, and were free to adjust their viewing distance. A customized graphical user interface (GUI) was used to display the images at resolution matched to the screen (i.e., 512×512 pixels), and subjects were able to zoom in to any portion of the images for more careful comparison. The screen had the resolution of 1920×1080 pixels, and was calibrated in accordance with the

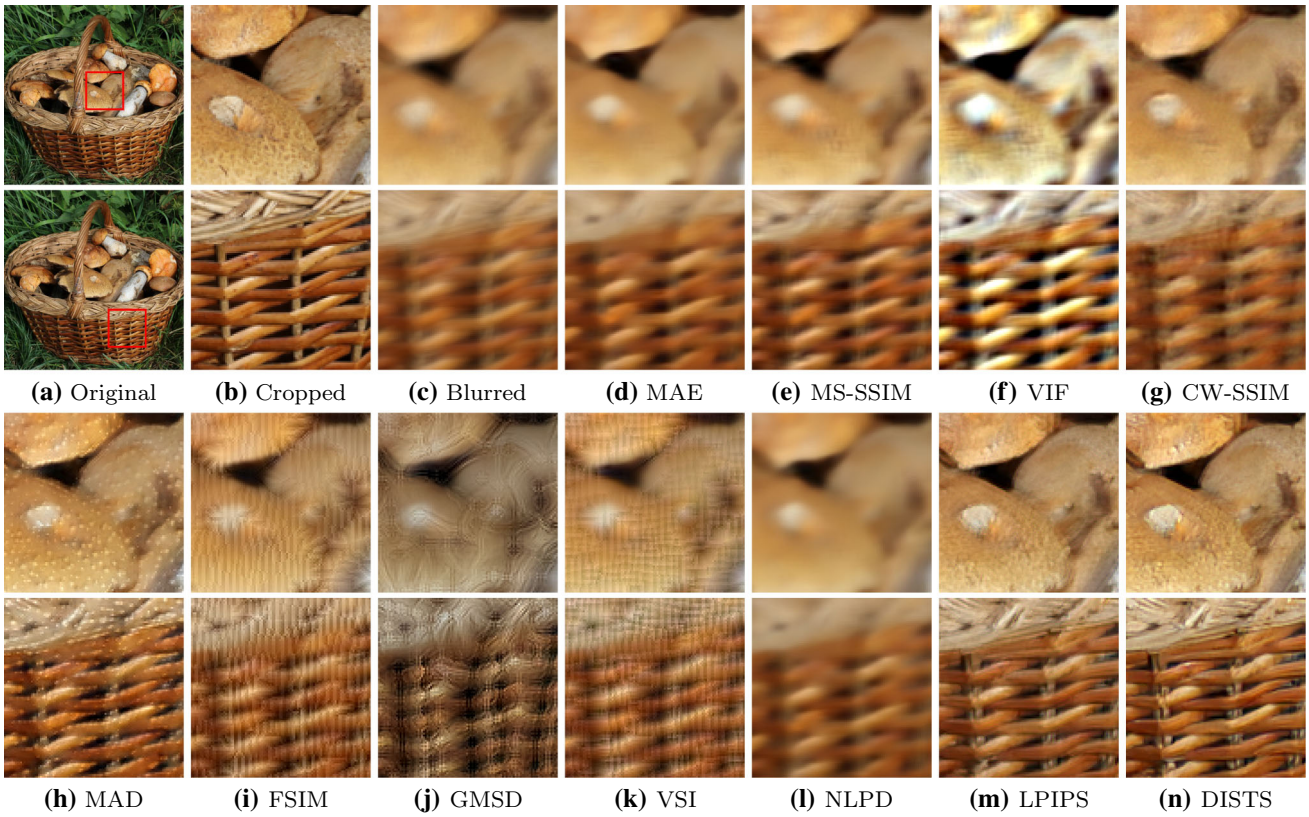


Fig. 11 Deblurring results for two regions cropped from an example image, using a DNN optimized for different IQA models

recommendations of ITU-R BT.500-11 ITU-R (2002). Tests were performed in indoor spaces with ordinary illumination levels.

We generated a total of $\binom{11}{2} \times 4 \times 20 = 4400$ paired comparisons for 11 IQA models, 4 tasks, and 20 test images. We gathered data from 25 subjects (13 males and 12 females) aged between 18 and 22, with normal or corrected-to-normal visual acuity. Subjects had general background knowledge of image processing and computer vision, but were otherwise naïve to the purpose of this study. To reduce fatigue, we performed the experiment in multiple sessions, each consisting of 500 randomly selected comparisons, with the randomized left–right presentation, and allowed subjects to take a break at any time during the session. Subjects were encouraged, but not required, to participate in multiple sessions. In order to detect subjects that were not properly performing the task, we included 5 pairs where one image was of unambiguously better quality (e.g., the original and a noisy image). Our intention was to discard the results of subjects who failed in more than one of these pairs, but the results of all subjects turned out to be valid. In total, each image pair was evaluated by at least 5 subjects, and each IQA model was ranked over 1000 times for each vision task.

6 Experimental Results

Based on the subjective data, we conducted a quantitative comparison of the IQA models through the lens of perceptual optimization. We also qualitatively compared the visual results associated with the IQA models. Last, we combined a top-performing IQA model with adversarial loss (Goodfellow et al. 2014) to test whether additional perceptual gains could be obtained in blind image deblurring.

6.1 Quantitative Results

We employed the Bradley-Terry model (Bradley and Terry 1952) to convert paired comparison results to global rankings. This probabilistic model assumes that the visual quality of the k-th test image optimized for the i-th IQA model, q_i^k, follows a Gumbel distribution with location μ_i^k and scale s. Assuming independence between q_i^k and q_j^k, the difference $q_i^k - q_j^k$ is a logistic random variable, and therefore $p_{ij}^k = P(q_i^k \geq q_j^k)$ can be computed using the logistic cumulative distribution function:

$$p_{ij}^k = P(q_i^k - q_j^k \geq 0) = \frac{\exp(\mu_i^k/s)}{\exp(\mu_i^k/s) + \exp(\mu_j^k/s)}, \quad (10)$$

Table 1 SRCC of objective ranking scores from the IQA models against subjective ranking scores

IQA model	Denoising	Deblurring	Superresolution	Compression
MAE	0.527	0.164	0.309	0.455
MS-SSIM	**0.564**	0.127	0.455	0.346
VIF	0.273	0.600	0.418	0.018
CW-SSIM	0.382	0.418	0.091	0.018
MAD	0.418	0.455	0.346	0.382
FSIM	0.236	0.054	0.091	0.127
GMSD	0.091	0.018	0.127	0.127
VSI	0.164	0.018	0.018	0.091
NLPD	0.491	0.127	0.200	0.309
LPIPS	**0.709**	**0.855**	**0.782**	**0.782**
DISTS	0.346	**0.891**	**0.782**	**0.855**

Top two results are marked in bold

Table 2 Verification of results obtained by our PyTorch re-implementations of the tested IQA models, on three IQA databases

IQA model	Grayscale			Color		
	LIVE	CSIQ	TID2013	LIVE	CSIQ	TID2013
MS-SSIM	0.951/0.951	0.906/0.886	0.786/0.782	0.931/0.932	0.902/0.886	0.801/0.816
CW-SSIM	0.786/0.781	0.745/0.738	0.673/0.680	0.741/0.747	0.744/0.744	0.709/0.725
VIF	0.964/0.963	0.911/0.911	0.677/0.676	0.957/0.957	0.894/0.894	0.654/0.654
NLPD	0.937/0.938	0.932/0.937	0.800/0.800	0.917/0.914	0.913/0.913	0.812/0.808
GMSD	0.960/0.960	0.950/0.950	0.804/0.804	0.949/0.948	0.937/0.934	0.830/0.823
MAD	0.967/0.960	0.947/0.941	0.781/0.773	0.954/0.951	0.937/0.935	0.758/0.740
FSIM	0.963/0.963	0.924/0.916	0.802/0.802	0.965/0.965	0.931/0.923	0.851/0.851
VSI	0.953/0.950	0.930/0.923	0.805/0.793	0.952/0.956	0.942/0.937	0.897/0.889
LPIPS	0.932/0.932	0.837/0.837	0.616/0.616	0.932/0.932	0.876/0.876	0.670/0.670
DISTS	0.942/0.942	0.905/0.905	0.764/0.764	0.954/0.954	0.929/0.929	0.830/0.830

Numbers indicate SRCC values (reported in original publication/produced by our re-implementation). Bold indicates methods that are computed only on grayscale images in their original versions; we have extended them to evaluate RGB images by averaging the values across all channels

Table 3 Summary of datasets for evaluating full-reference IQA models

Dataset	# of Reference images	# of Distorted images	Distortion types
LIVE (Sheikh et al. 2006)	29	779	Traditional
CSIQ (Larson and Chandler 2010)	30	866	Traditional
TID2013 (Ponomarenko et al. 2015)	25	3000	Traditional
FLT (Egiazarian et al. 2018)	75	300	Denoising
Liu13 (Liu et al. 2013)	40	1200	Deblurring
Lai16 (Lai et al. 2016)	25	2800	Deblurring
Ma17 (Ma et al. 2017a)	30	1620	Super-resolution
QADS (Zhou et al. 2019)	20	980	Super-resolution
SHRQ (Min et al. 2019)	80	600	Dehazing
Tian19 (Tian et al. 2018)	10	140	Rendering
SynTex (Golestaneh et al. 2015)	21	105	Texture synthesis
TQD (Ding et al. 2020)	10	150	Texture synthesis
BAPPS (Zhang et al. 2018)	–	26,904	Multiple
Proposed	20	880	Multiple

Traditional distortion types include artificial Gaussian noise, Gaussian blur, JPEG compression, etc. As in the dataset we describe in Sect. 5, BAPPS contains multiple distortion types, produced by computational methods for different vision tasks

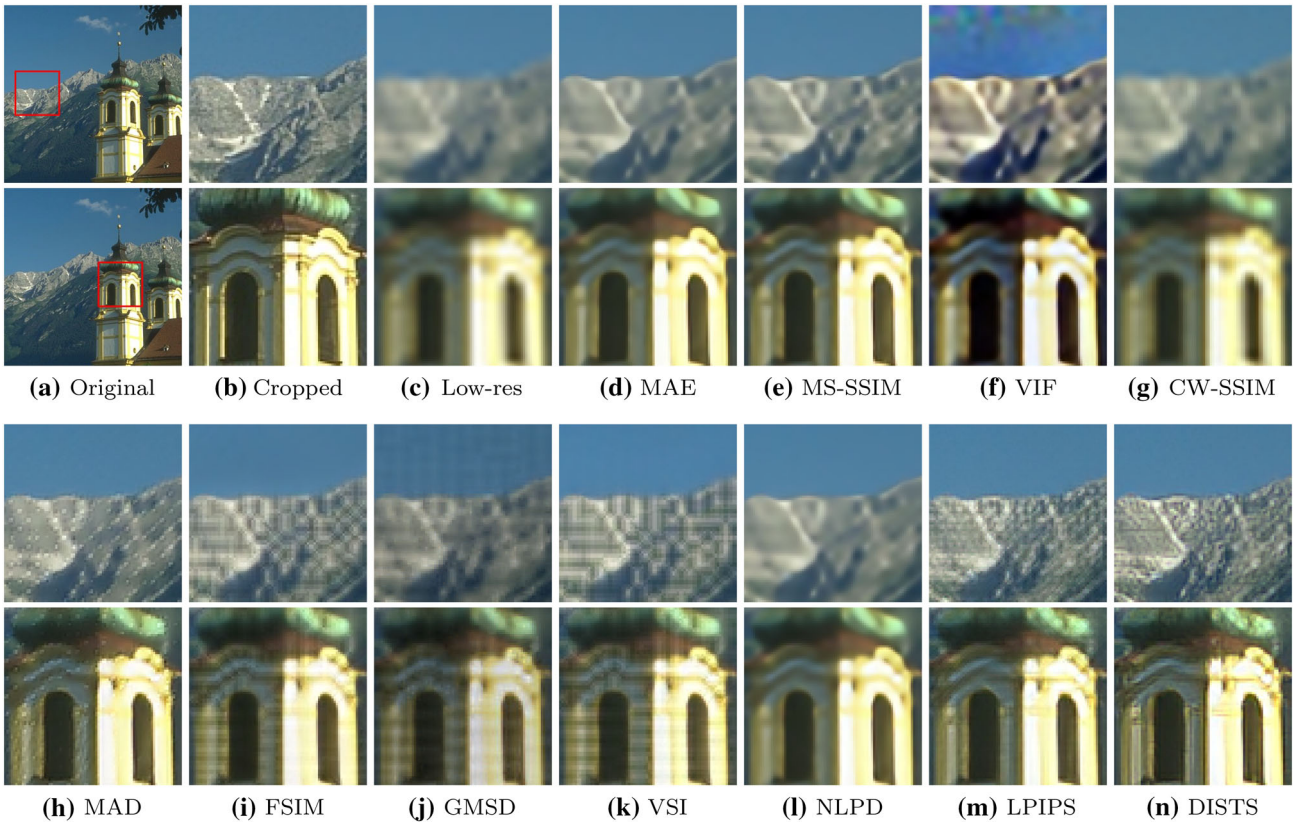


Fig. 12 Super-resolution results for two cropped regions from an example image, using a DNN optimized for different IQA models

where s is usually set to 1, leading to a simplified expression:

$$p_{ij}^k = \frac{e^{\mu_i^k}}{e^{\mu_i^k} + e^{\mu_j^k}}. \tag{11}$$

As such, we may obtain the negative log-likelihood of our pairwise count matrix W^k:

$$\ell(\mu^k | W^k) = \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \left(w_{ij}^k \log \left(e^{\mu_i^k} + e^{\mu_j^k} \right) - w_{ij}^k \mu_i^k \right), \tag{12}$$

where w_{ij}^k represents the number of times that D_i is preferred over D_j for the k-th test image. For each of the four low-level vision tasks, we minimized Eq. (12) iteratively using gradient descent to obtain the optimal estimate $\hat{\mu}^k$. We averaged $\hat{\mu}^k$ over the 20 test images, resulting in four global rankings of perceptual optimization performance, as shown in Fig. 9. It is clear that MS-SSIM (Wang et al. 2003) and MAE are superior to the other IQA models in the task of denoising, whereas DNN-based measures DISTS (Ding et al. 2020) and LPIPS (Zhang et al. 2018), outperform the others in all other tasks. Thus, there is no single IQA model that performs best across all tasks. We ascribe this to differences in the nature of the tasks: denoising requires distinguishing signal and noise, deblurring, super-resolution, and compression all require recovery of discarded information from partial deterministic measurements (for the first two, via linear projection, and for compression via quantization). MS-SSIM and MAE are both known to prefer smooth appearances, and are seen to excel at denoising. Both DISTS and LPIPS explicitly represent aspects of fine textures, and are superior for the remaining three tasks. Finally, it is important to note that many of the models, despite their impressive abilities to explain existing IQA databases, are outperformed by MAE, the simplest metric in our set.

To determine whether the optimization results of the IQA models are statistically significant, we conducted an independent paired-sample t-test. The null hypothesis is that the ranking scores $\{\mu_i^k\}_{k=1}^{20}$ for D_i and $\{\mu_j^k\}_{k=1}^{20}$ for D_j come from the same normal distribution with unknown variance. When the test cannot reject the null hypothesis at the $\alpha = 5\%$ significance level, the two IQA models have statistically indistinguishable performance, and we considered them to belong to the same group. Grouping results are shown in Fig. 9. Surprisingly, we find that the perceptual gains of MS-SSIM over MAE are statistically insignificant on all four tasks, despite the fact that MS-SSIM is far better than MAE in explaining existing IQA databases. Relying on similar sets

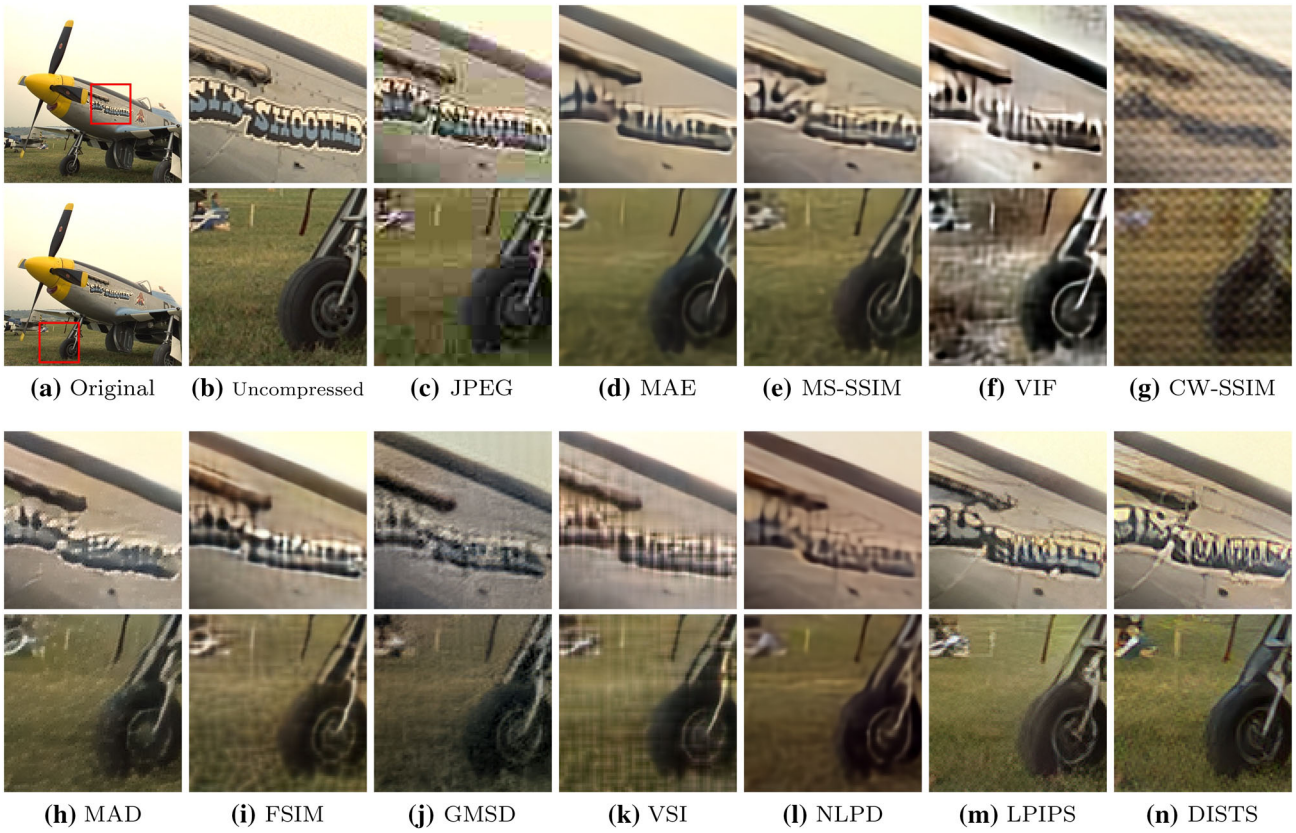


Fig. 13 Compression results for two cropped regions from an example image, using a DNN optimized for different IQA models

of VGG features (Simonyan and Zisserman 2015), DISTS and LPIPS also achieve similar performance, except for the super-resolution task where the former is statistically better.

By computing the Spearman's rank correlation coefficient (SRCC) between objective model rankings (in Fig. 7) and subjective human rankings (in Fig. 9), we are able to compare the algorithm-level performance of the 11 IQA models on the new dataset. We find from the Table 1 that there is a lack of correlation between model predictions and human judgments for the majority of IQA methods. DISTS and LPIPS tend to rank the images with complex model-dependent distortions in a more perceptually consistent way. We refer interested readers to "Appendix 1" for more comparisons on several IQA databases dedicated to low-level vision problems.

6.2 Qualitative Results

In this subsection, we show example images produced by each IQA-optimized method, qualitatively summarize the types of visual distortion, and use them to diagnose the shortcomings of the corresponding IQA models.

Figure 10 shows *denoising* results for the "cat" image. We observe that MAE, MS-SSIM, and NLPD do a good job in denoising flat regions, but tend to over-smooth texture regions. VIF encourages detail enhancement, leading to artificial local contrast, while GMSD produces a relatively dark appearance presumably because it discards local luminance information. Moreover, the results of FSIM and VSI exhibit noticeable artifacts. LPIPS and DISTS preserve fine details, but may not fully remove noise in smooth regions, mistaking the remaining noise as visually plausible texture. Overall, traditional IQA models MAE and MS-SSIM denoise images with various content variations robustly, keeping high-frequency information loss within the acceptable range. This may explain why they are the dominant objective functions for this task.

Figure 11 shows *deblurring* results for the "basket" image. We see that most of the IQA methods fail, but in different ways. Specifically, the results of MAE, MS-SSIM, CW-SSIM, and NLPD are quite blurred. FSIM, GMSD, and VSI generate severe ringing artifacts. VIF again fails to adjust the local contrast. MAD exhibits undesirable white dot artifacts, although the main structures are sharp. LPIPS succeeds in deblurring this example, while DISTS produces a result that

(a) Clean **(b)** Blurry **(c)** DeblurGAN-v2 **(d)** Fine-tuned

Fig. 14 Deblurring examples obtained by the original DebluGAN-v2 and the fine-tuned DeblurGAN-v2 (with the loss in Eq. (14))

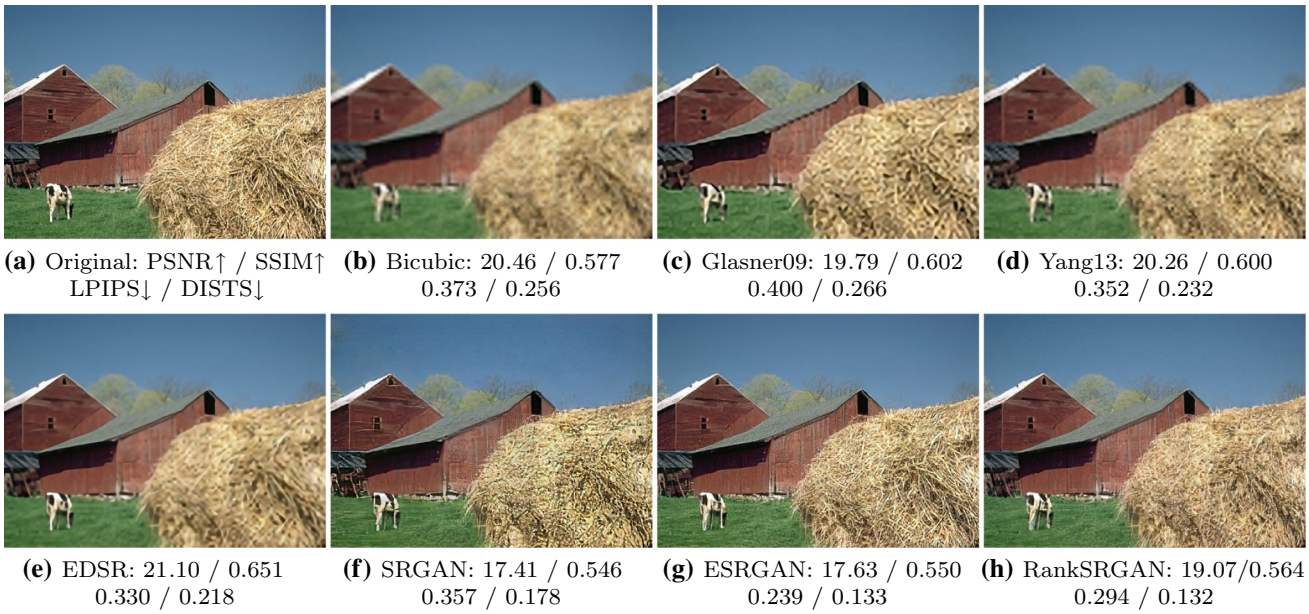
is closest to the original. This is consistent with current state-of-the-art deblurring results (Kupyn et al. 2019), generated by incorporating comparison of the VGG features into the loss.

Figure 12 shows *super-resolution* results for the "corner tower" image. Again, MAE, MS-SSIM, NLPD, and especially CW-SSIM produce somewhat blurred images, without recovering fine details. MAD, FSIM, GMSD, and VSI are able to generate some "structures", but these are perceived as unpleasant model-dependent artifacts. Benefiting from its texture synthesis capability, DISTS has the potential to super-resolve perceptually plausible fine details, although they differ from those of the original image.

Figure 13 shows *compression* results for the "airplane" image at 0.24 ± 0.01 bpp. A JPEG image, compressed to 0.25 bpp, suffers from block and blur artifacts. Overall, the main structures of the original image are well preserved for most IQA models, but the fine details (e.g., the grass) have to be discarded at this low bitrate, or are synthesized with other forms of distortion. VIF reconstitutes a desaturated image with over-enhanced global contrast, and CW-SSIM superimposes periodic artifacts on the underlying image. White dots and ringing artifacts are again apparent in the results of MAD and VSI, respectively. The image by NLPD is blurred and red-shifted. Both LPIPS and DISTS succeed in synthesizing textures that are visually similar to the original.

We can summarize the artifacts created during perceptual optimization, some of which are not found in traditional image databases for the purpose of quality assessment:

– *Blurring* is a frequently seen distortion type in all four of the tasks, and is mainly caused by error visibility methods (e.g., MAE and NLPD) and structural similarity methods (e.g., MS-SSIM), which rely on simple injective mappings. Specifically, MAE and SSIM work directly with pixels, and NLPD transforms the input image to a multi-scale overcomplete representation using a single stage of local mean subtraction and divisive normalization.



(a) Original: PSNR↑ / SSIM↑ LPIPS↓ / DISTS↓ **(b)** Bicubic: 20.46 / 0.577 0.373 / 0.256 **(c)** Glasner09: 19.79 / 0.602 0.400 / 0.266 **(d)** Yang13: 20.26 / 0.600 0.352 / 0.232

(e) EDSR: 21.10 / 0.651 0.330 / 0.218 **(f)** SRGAN: 17.41 / 0.546 0.357 / 0.178 **(g)** ESRGAN: 17.63 / 0.550 0.239 / 0.133 **(h)** RankSRGAN: 19.07/0.564 0.294 / 0.132

Fig. 15 A visual quality assessment example of super-resolution. (**a**) High-resolution image, (**b-h**) are the super-resolution results computed using bicubic interpolation, Glasner09 (Glasner et al. 2009), Yang13 (Yang andYang 2013),EDSR(Lim et al. 2017),SRGAN(Ledig et al. 2017), ESRGAN (Wang et al. 2018), and RankSRGAN (Zhang et al. 2019a), respectively. One can see that the GAN-based results (**f-h**) are visually superior to the others, contrary to the predictions of PSNR and SSIM. LPIPS indicates that the result (**f**) is worse than (**d**) and (**e**), in disagreement with visual inspection. DISTS is correlated well with human perception in this example

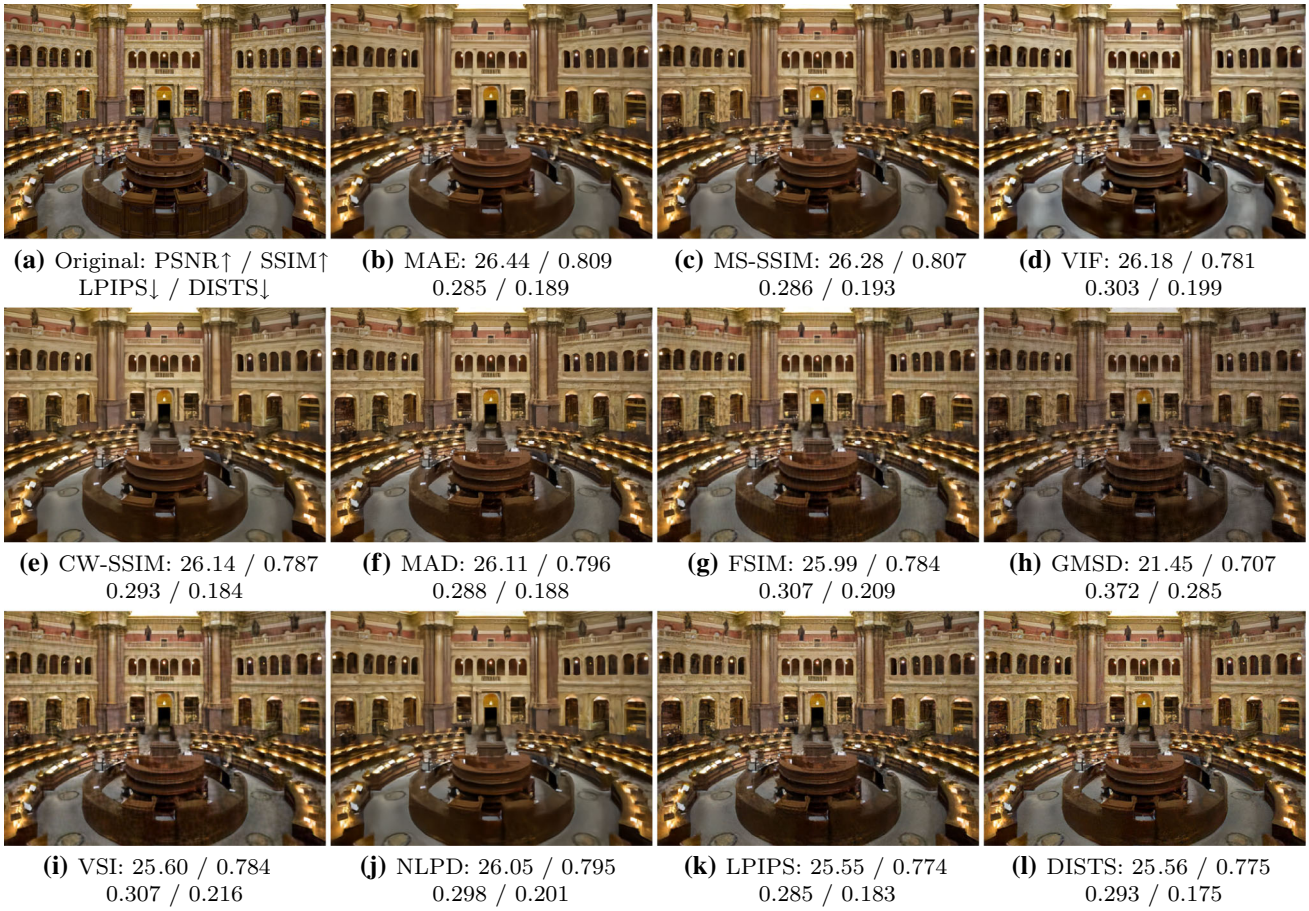
Table 4 SRCC comparison of IQA models on existing image generation/restoration databases

IQA model	Denoising	Deblurring		Super-resolution		Dehazing	Rendering	Texture synthesis	
	FLT	Lai16	Liu13	Ma17	QADS	SHRQ	Tian19	SynTEX	TQD
PSNR	0.183	0.301	0.803	0.592	0.360	0.740	0.536	0.114	0.233
SSIM	0.355	0.298	0.777	0.624	0.529	0.692	0.230	0.620	0.307
MS-SSIM	0.246	0.320	0.898	0.795	0.717	0.687	0.396	0.469	0.288
VIF	0.169	0.261	0.864	0.831	**0.815**	0.667	0.259	0.448	0.305
CW-SSIM	0.101	0.600	0.742	0.706	0.474	0.698	0.522	0.496	0.325
MAD	0.182	0.446	0.897	**0.864**	0.723	0.605	**0.622**	0.134	0.302
FSIM	0.555	0.297	**0.921**	0.747	0.687	0.695	0.476	0.093	0.176
GMSD	0.389	0.174	0.918	0.851	0.765	0.663	0.479	0.006	0.256
VSI	0.528	0.295	0.920	0.710	0.584	0.696	0.531	0.123	0.179
NLPD	0.151	0.323	0.853	0.732	0.591	0.608	0.463	0.483	0.271
PieAPP	**0.629**	**0.601**	0.786	0.771	**0.849**	0.725	0.298	**0.709**	**0.713**
LPIPS	0.457	0.347	0.867	0.788	0.669	**0.777**	0.311	0.663	0.392
DISTS	**0.636**	**0.754**	**0.941**	**0.878**	0.809	**0.789**	**0.671**	**0.923**	**0.910**

Top two results are marked in bold

Under strict constraints imposed by the tasks, they prefer to make a more conservative estimate, producing something akin to an average of all possible outcomes with sharp structures, as would occur when optimizing MSE.

– *Ringing* is a high-frequency distortion type that often occurs in the images optimized for FSIM, VSI and GMSD (see Fig. 11(i)–(k)). One common characteristic of the three models is that they rely heavily (in some cases, solely) on local gradient magnitude for feature similarity comparison, underweighting (or abandoning) other perceptually important features (such as local luminance and local phase). This creates "shortcuts" that the DNNs can exploit, generating distortions with similar local gradient statistics.
– *White dot* artifacts are typical in the optimization results of MAD, which extracts lower-order image statistics



(a) Original: PSNR↑ / SSIM↑ LPIPS↓ / DISTS↓ **(b)** MAE: 26.44 / 0.809 0.285 / 0.189 **(c)** MS-SSIM: 26.28 / 0.807 0.286 / 0.193 **(d)** VIF: 26.18 / 0.781 0.303 / 0.199

(e) CW-SSIM: 26.14 / 0.787 0.293 / 0.184 **(f)** MAD: 26.11 / 0.796 0.288 / 0.188 **(g)** FSIM: 25.99 / 0.784 0.307 / 0.209 **(h)** GMSD: 21.45 / 0.707 0.372 / 0.285

(i) VSI: 25.60 / 0.784 0.307 / 0.216 **(j)** NLPD: 26.05 / 0.795 0.298 / 0.201 **(k)** LPIPS: 25.55 / 0.774 0.285 / 0.183 **(l)** DISTS: 25.56 / 0.775 0.293 / 0.175

Fig. 16 Another set of denoising results optimized for different IQA models

Table 5 2AFC score comparison of IQA models on the BAPPS dataset and the proposed dataset

IQA model	BAPPS				Proposed			
	Colorization	Video deblurring	Frame interpolation	Superresolution	Denoising	Deblurring	Superresolution	Compression
PSNR	0.624	0.590	0.543	0.642	0.627	0.518	0.612	0.689
SSIM	0.522	0.583	0.548	0.613	**0.636**	0.575	0.599	0.649
MS-SSIM	0.522	0.589	0.572	0.638	0.623	0.568	0.655	0.665
VIF	0.515	0.594	0.597	0.651	0.589	0.607	0.655	0.540
CW-SSIM	0.512	**0.601**	0.604	0.665	0.623	0.651	0.584	0.496
MAD	0.490	0.593	0.581	0.655	0.624	0.671	0.681	0.651
FSIM	0.573	0.590	0.581	0.660	0.522	0.490	0.525	0.563
GMSD	0.517	0.594	0.575	0.676	0.417	0.454	0.469	0.567
VSI	0.597	0.591	0.568	0.668	0.518	0.470	0.487	0.576
NLPD	0.528	0.584	0.552	0.655	0.622	0.514	0.629	0.652
PieAPP	0.594	0.582	0.598	0.685	0.625	0.734	**0.744**	0.822
LPIPS	**0.625**	**0.605**	**0.630**	**0.705**	**0.657**	**0.788**	**0.768**	**0.834**
DISTS	**0.627**	0.600	**0.625**	**0.710**	0.602	**0.790**	0.704	**0.833**

Top two results are marked in bold

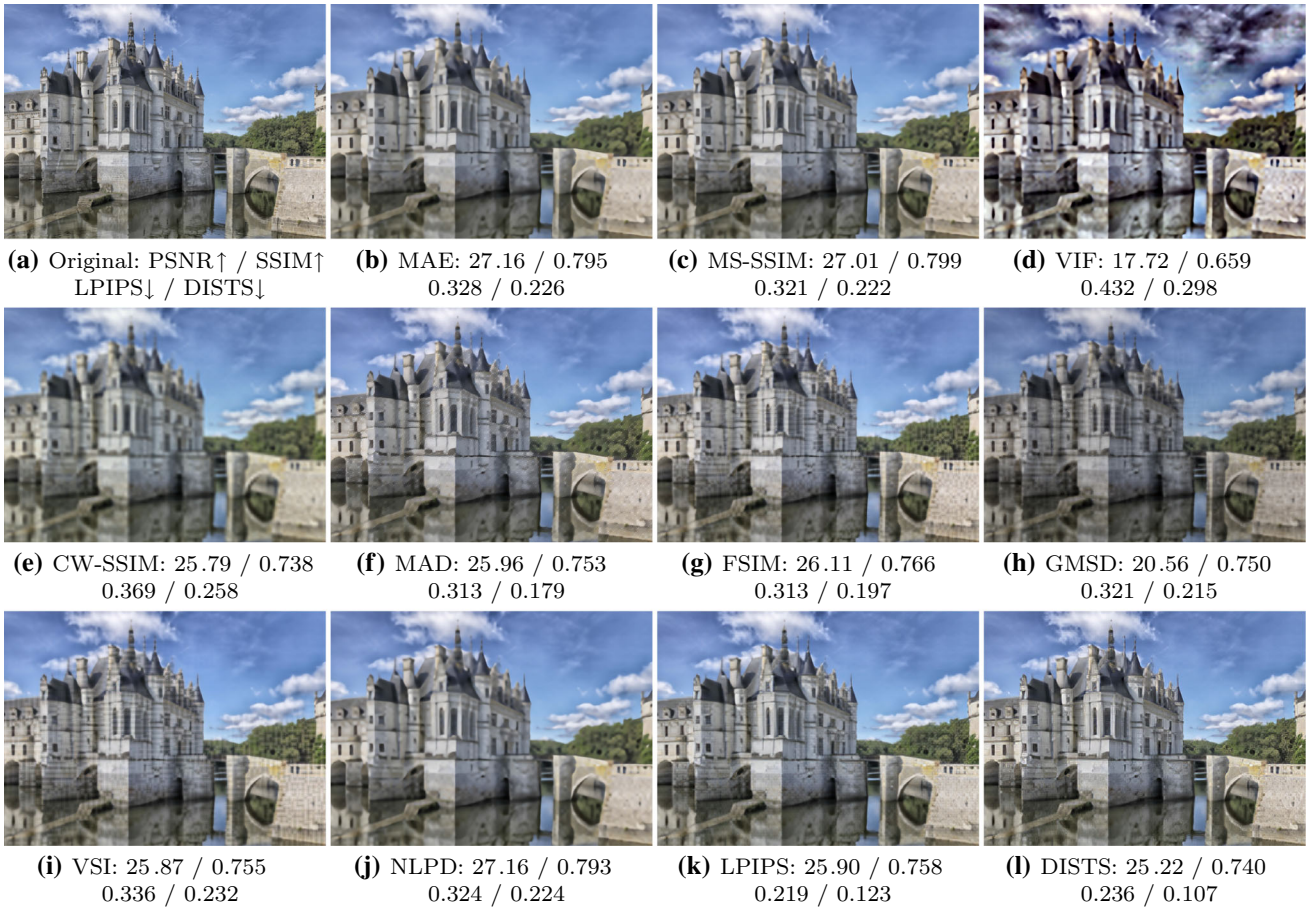
Fig. 17 Another set of deblurring results optimized for different IQA models

from responses of Gabor filters at multiple scales and orientations. The resulting set of statistical measurements seems insufficient to summarize natural image structures that exhibit higher-order dependencies. Therefore, MAD is "blind" to distortions that satisfy the same set of statistical constraints, and gives the optimized distorted image a high-quality score.

– *Over-enhancement* of local image contrast is encouraged by VIF, which, in most of our experiments, causes significant quality degradation. We believe this arises because VIF does not fully respect reference information when normalizing the covariance term. Specifically, only the second-order statistics of the reference image are used to construct the normalization factor. By incorporating the same statistics computed from the distorted image into normalization, the problem of over-enhancement may be alleviated. In general, quality assessment of image enhancement is a challenging problem (Fang et al. 2015; Wang et al. 2015), and to the best of our knowledge, all existing full-reference IQA models fail to reward properly-enhanced cases, while penalizing over-enhanced cases.
– *Luminance and color* artifacts are perceived in final images that are associated with many IQA models. Two causes seem plausible. First, methods such as GMSD discard luminance information. Second, methods such as MS-SSIM and NLPD are originally designed for grayscale images only. Applying them to RGB channels separately fails to take into account hue and saturation information. Transforming to a perceptually better color space, and making use of knowledge of color distortions (Rajashekar et al. 2009) offers an pportunity for improvement.

6.3 Combining with Adversarial Loss

In the field of image restoration and generation, many state-of-the-art algorithms are based on adversarial training (Goodfellow et al. 2014), demonstrating impressive capabilities in synthesizing realistic visual content. The output of the adversarial loss is the probability of an image being computer-generated, but this does not confer capabilities for no-reference IQA modeling, as confirmed by a low SRCC of 0.366 on the LIVE dataset (Sheikh et al. 2006). Nev-



(a) Original: PSNR↑ / SSIM↑ LPIPS↓ / DISTS↓ **(b)** MAE: 27.16 / 0.795 0.328 / 0.226 **(c)** MS-SSIM: 27.01 / 0.799 0.321 / 0.222 **(d)** VIF: 17.72 / 0.659 0.432 / 0.298

(e) CW-SSIM: 25.79 / 0.738 0.369 / 0.258 **(f)** MAD: 25.96 / 0.753 0.313 / 0.179 **(g)** FSIM: 26.11 / 0.766 0.313 / 0.197 **(h)** GMSD: 20.56 / 0.750 0.321 / 0.215

(i) VSI: 25.87 / 0.755 0.336 / 0.232 **(j)** NLPD: 27.16 / 0.793 0.324 / 0.224 **(k)** LPIPS: 25.90 / 0.758 0.219 / 0.123 **(l)** DISTS: 25.22 / 0.740 0.236 / 0.107

Fig. 18 Another set of super-resolution results optimized for different IQA models

ertheless, adversarial loss may be useful at the algorithm level, meaning that given a set of images generated by a computational method, the average probability quantitatively measures the capability of the method in generating photorealistic high-quality images. In this subsection, we explored the combination of the adversarial loss and a top-performing IQA measure for additional perceptual gains.

We chose the task of blind image deblurring, and fine-tuned a state-of-the-art model—DeblurGAN-v2 (under the configuration of Inception-ResNet) (Kupyn et al. 2019). The original loss function for the generator is

$$\ell_o = 0.5 \times \ell_{\text{MSE}} + 0.006 \times \ell_{\text{VGG}} + 0.01 \times \ell_{\text{Adv}}. \quad (13)$$

The first and second terms are the MSE on pixels and responses of conv3_3 of VGG19 (Simonyan and Zisserman 2015), respectively, and ℓ_{Adv} is a variant of the adversarial loss (Kupyn et al. 2019). We selected the best-performing IQA model—DISTS-for this experiment. We followed the same training strategy, but modified the loss function of the generator to be

$$\ell_n = \ell_{\text{DISTS}} + 0.001 \times \ell_{\text{Adv}}, \quad (14)$$

where ℓ_{DISTS} denotes the DISTS index. An immediate advantage of this replacement is that the number of hyperparameters is reduced, making manual hyperparameter adjustment easier. After fine-tuning, the average DISTS value decreases from 0.22 to 0.18 on the Köhler test dataset (Köhler et al. 2012). Figure 14 shows two visual examples, from which we find that the fine-tuned results have sharper edges and enhanced contrast, indicating that perceptual gains may be obtained by DISTS on the two examples.

7 Conclusions

We have conducted a comprehensive study of perceptual optimization of four low-level vision tasks, guided by eleven full-reference IQA models. This provides an alternative means of testing the perceptual relevance of IQA models in a practical setting, which we believe is an important complement to the conventional methodology for IQA model

Fig. 19 Another set of compression results optimized for different IQA models

evaluation. Subjective testing led to several useful findings. First, through perceptual optimization, we generated a number of distortions (different from those used in existing IQA databases), which may easily fool the respective models or models of similar design philosophies (see Table 1). It should be noted that the emergence of specific distortions is in principle dependent on the experimental choices (e.g., initialization strategy, model architecture, and optimization technique). Second, although they underperformed the DNN-based models on three of four applications, the standard full-reference IQA models (MS-SSIM and MAE) are still valuable tools for optimizing image processing systems due to their robustness and simplicity. Third, more recent IQA models with surjective mappings may still be used to monitor image quality and to optimize the parameter settings of image processing methods, but in a limited and well-controlled space. Last, the two DNN-based models (LPIPS and DISTS) offered the best overall performance in our experiments, but their high computational complexity and lack of interpretability may hinder their use.

Our work has interesting connections to two separate lines of research. First, inspired by the philosophy of "analysis by synthesis" (Grenander 1970), Wang and Simoncelli (2008) introduced the maximum differentiation competition methodology to automatically synthesize images for efficiently comparing IQA models. Given two IQA models, MAD generates samples in the space of all possible images that best discriminate the two models. However, the synthesized images may be highly unnatural, and in this case, of limited practical importance. Ma et al. (2020) alleviated this issue by manually constraining the search space to a finite image set of practical interest. Our approach combines the best aspects of these two methods, in the sense that the test images for model comparison are automatically generated by the trained networks, but arise as solutions of real-world vision tasks and are thus of practical importance. Second, the existence of type II adversarial examples (Szegedy et al. 2013) has exposed the vulnerability of many computer vision algorithms, where a tiny change to the input that is imperceptible to the human eye would cause the algorithm to make classification mistakes. In our case, weaknesses in an IQA model are exposed through optimized images that may be interpreted as type I "adversarial" examples of the model: a significant change is made to the original image that sub-

stantially degrades its perceptual quality, but the model still claims that this image is of high quality.

The analysis of our experimental results suggests several desirable properties that should be included in future IQA methods. First, the transformation used in the IQA model should be perceptual, mapping the input images into a space where a simple distance measure (e.g., Euclidean) matches human judgements of image quality. This is in the same spirit that color scientists pursue perceptually uniform color spaces, and is an underlying principle of a number of existing models (e.g., NLPD). Zhang et al. (2018) and Ding et al. (2020) demonstrated that a cascade of linear convolution, downsampling, and rectified nonlinearity optimized for high-level vision tasks may be a good candidate. Second, the IQA model should enjoy unique optima (i.e., the underlying mapping should be injective) to guarantee that images close to optimal are visually similar to the original. This criterion was respected by early models (e.g., MS-SSIM), but was largely overlooked in recent IQA model development. Third, the IQA model should be continuous and differentiable, with well-behaved gradients, to aid optimization in complex situations (e.g., training DNNs with millions of parameters). Last but not least, the IQA model should be computationally efficient, enabling real-time quality assessment and perceptual optimization. To the best of our knowledge, although many current IQA models possess subsets of these properties, no current IQA model satisfies them all.

Acknowledgements The authors would like to thank all subjects who participated in our subjective study during this period of the coronavirus pandemic. This work was supported in part by the National Natural Science Foundation of China (62071407 to KDM and 62022002 to SQW), the CityU SRG-Fd and APRC Grants (7005560 and 9610487 to KDM), the Hong Kong RGC Early Career Scheme (9048122 to SQW), and the Howard Hughes Medical Institute (investigatorship to EPS).

Appendix 1: Perceptual Correlation Comparison of IQA Models

A conventional method for evaluating IQA models is to compute their agreement with subjective scores in one or more standardized IQA databases [e.g., LIVE (Sheikh et al. 2006), CSIQ (Larson and Chandler 2010) or TID2013 (Ponomarenko et al. 2015)], consisting of artificially distorted images. Many existing IQA models achieve impressive correlation with these databases (see Table 2), but their performance in assessing the perceptual quality of images produced by low-level vision algorithms has not been tested. In this appendix, we tested them on multiple human-rated image generation/restoration databases, including a denoising database—FLT (Egiazarian et al. 2018), two motion deblurring databases—Liu13 (Liu et al. 2013) and Lai16 (Lai et al. 2016), two super-resolution databases—Ma17 (Ma et al. 2017a) and QADS (Zhou et al. 2019), a dehazing database—SHRQ (Min et al. 2019), a depth image-based rendering database—Tian19 (Tian et al. 2018), two texture synthesis databases—SynTex (Golestaneh et al. 2015) and TQD (Ding et al. 2020), and a patch similarity database—BAPPS (Zhang et al. 2018). The details of these databases are summarized in Table 3.

Tables 4 and 5 show the performance comparisons of 13 IQA methods in terms of the SRCC and 2AFC scores. As suggested in Zhang et al. (2018), the 2AFC score is computed by: $pq + (1-p)(1-q)$, where p is the percentage of human votes and $q = \{0, 1\}$ is the vote of an IQA model. When q agrees with the majority of human votes, the 2AFC score is larger, indicating better performance. We find that the overall performance of all models is lower compared to that in the standard IQA databases (see Table 2), indicating the difficulty of generalizing to unseen distortions. Moreover, DNN-based measures are relatively better than knowledge-driven models in these application-oriented databases, but there is still significant room for improvement.

Figure 15 shows a quality assessment example of real-world super-resolution methods. Here we only compared the most widely used measures (PSNR and SSIM), and the two that performed best both on optimization and assessment (LPIPS and DISTS). It is not surprising that PSNR and SSIM have the poor correlation with human opinions, as they focus more on signal fidelity than perceptual quality (Blau and Michaeli 2018). LPIPS and DISTS perform better, but the former is somewhat oversensitive to texture substitution. As many recent image restoration algorithms succeed in generating richer textures, DISTS holds much promise for use in quality assessment for such applications (Figs. 16, 17, 18 and 19).

References

Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., & Gool, L. V. (2019). Generative adversarial networks for extreme learned image compression. In *IEEE conference on computer vision and pattern recognition* (pp. 221–231).

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016) End-to-end optimization of nonlinear transform codes for perceptual quality. In *Picture coding symposium* (pp. 1–5).

Ballé, J., Laparra, V., & & Simoncelli, E. P. (2017). End-to-end optimized image compression. In *International conference on learning representations* (pp. 1–27).

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018) Variational image compression with a scale hyperprior. In *International conference on learning representations* (pp. 1–47).

Blau, Y., & Michaeli, T. (2018). The perception–distortion tradeoff. In *IEEE conference on computer vision and pattern recognition* (pp. 6228–6237).

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

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Burt, P., & Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, *31*(4), 532–540.

Chang, H. W., Yang, H., Gan, Y., & Wang, M. H. (2013). Sparse feature fidelity for perceptual image quality assessment. *IEEE Transactions on Image Processing*, *22*(10), 4007–4018.

Channappayya, S. S., Bovik, A. C., Caramanis, C., & Heath, R. W. (2008). SSIM-optimal linear image restoration. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 765–768).

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, *16*(8), 2080–2095.

Daly, S. J. (1992). Visible differences predictor: An algorithm for the assessment of image fidelity. *Human Vision, Visual Processing, and Digital Display*, *3*, 2–15.

Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. CoRR arXiv:2004.07728.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision* (pp. 184–199).

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224.

Egiazarian, K., Ponomarenko, M., Lukin, V., & Ieremeiev, O. (2018). Statistical evaluation of visual quality metrics for image denoising. In *IEEE International conference on acoustics, speech and signal processing* (pp. 6752–6756).

Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, *15*(12), 3736–3745.

Fang, Y., Ma, K., Wang, Z., Lin, W., Fang, Z., & Zhai, G. (2015). No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters*, *22*(7), 838–842.

Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., & Freeman, W. T. (2006). Removing camera shake from a single photograph. *ACM Transactions on Graphics*, *25*(3), 787–794.

Girod, B. (1993). *What's wrong with mean-squared error?*. Cambridge, MA: MIT Press.

Glasner, D., Bagon, S., & Irani, M. (2009). Super-resolution from a single image. In *IEEE international conference on computer vision* (pp. 349–3560).

Golestaneh, S. A., Subedar, M. M., & Karam, L. J. (2015). The effect of texture granularity on texture synthesis quality. *Applications of Digital Image Processing XXXVIII*, *9599*, 356–361.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014) Generative adversarial nets. In*Conference on neural information processing systems* (pp. 2672–2680).

Grenander, U. (1970). A unified approach to pattern analysis. *Advances in Computers*, *10*, 175–216.

Hel-Or, Y., & Shaked, D. (2008). A discriminative approach for wavelet denoising. *IEEE Transactions on Image Processing*, *17*(4), 443–457.

Huang, J. B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *IEEE conference on computer vision and pattern recognition* (pp. 5197–5206).

ITU-R. (2002). *Methodology for the subjective assessment of the quality of television pictures*. Geneva: International Telecommunication Union.

Johnson, J., Alahi, A., & Li, F. F. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711).

Jordan, C. (1881). Sur la série de fourier. *Comptes Rendus de l'Académie des Sciences Paris*, *2*, 228–230.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations* (pp. 1–15).

Köhler, R., Hirsch, M., Mohler, B., Schölkopf, B., & Harmeling, S. (2012). Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *European conference on computer vision* (pp. 27–40).

Kovesi, P. (1999). Image features from phase congruency. *Journal of Computer Vision Research*, *1*(3), 1–26.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., & Matas, J. (2018). DeblurGAN: Blind motion deblurring using conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition* (pp. 8183–8192).

Kupyn, O., Martyniuk, T., Wu, J., & Wang, Z. (2019). DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *IEEE international conference on computer vision* (pp. 8878–8887).

Lai, W. S., Huang, J. B., Hu, Z., Ahuja, N., & Yang, M. H. (2016). A comparative study for single image blind deblurring. In *IEEE conference on computer vision and pattern recognition* (pp. 1701–1709).

Laparra, V., Ballé, J., Berardino, A., & Simoncelli, E. P. (2016). Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging*, *16*, 1–6.

Laparra, V., Berardino, A., Ballé, J., & Simoncelli, E. P. (2017). Perceptually optimized image rendering. *Journal of the Optical Society of America A*, *34*(9), 1511–1525.

Larson, E. C., & Chandler, D. M. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, *19*(1), 1–21.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., & Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, X., & Orchard, M. T. (2001). New edge-directed interpolation. *IEEE Transactions on Image Processing*, *10*(10), 1521–1527.

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In *IEEE conference on computer vision and pattern recognition workshop* (pp. 136–144).

Liu, A., Lin, W., & Narwaria, M. (2012a). Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, *21*(4), 1500–1512.

Liu, T. J., Lin, W., & Kuo, C. C. J. (2012b). Image quality assessment using multi-method fusion. *IEEE Transactions on Image Processing*, *22*(5), 1793–1807.

Liu, Y., Wang, J., Cho, S., Finkelstein, A., & Rusinkiewicz, S. (2013). A no-reference metric for evaluating the quality of motion deblurring. *ACM Transactions on Graphics*, *32*(6), 175–186.

Lubin, J. (1993). *The use of psychophysical data and models in the analysis of display system performance*. Cambridge, MA: MIT Press.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *Astronomical Journal*, *79*, 745.

Ma, C., Yang, C. Y., Yang, X., & Yang, M. H. (2017a). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, *158*, 1–16.

Ma, K., Duanmu, Z., & Wang, Z. (2018). Geometric transformation invariant image quality assessment using convolutional neural networks. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 6732–6736).

Ma, K., Duanmu, Z., Wang, Z., Wu, Q., Liu, W., Yong, H., et al. (2020). Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 851–864.

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., et al. (2017b). Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, *26*(2), 1004–1016.

Ma, K., Liu, X., Fang, Y., & Simoncelli, E. P. (2019). Blind image quality assessment by learning from multiple annotators. In *IEEE international conference on image processing* (pp. 2344–2348).

Ma, K., Yeganeh, H., Zeng, K., & Wang, Z. (2015). High dynamic range image compression by optimizing tone mapped image quality index. *IEEE Transactions on Image Processing*, *24*(10), 3086–3097.

Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., & Van Gool, L. (2018) Conditional probability models for deep image compression. In *IEEE conference on computer vision and pattern recognition* (pp. 4394–4402).

Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., et al. (2019). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Transactions on Multimedia*, *21*(9), 2319–2333.

Pan, J., Sun, D., Pfister, H., & Yang, M. H. (2016). Blind image deblurring using dark channel prior. In *IEEE conference on computer vision and pattern recognition* (pp. 1628–1636).

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., et al. (2015). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing Image Communication*, *30*, 57–77.

Portilla, J., Strela, V., Wainwright, M. J., & Simoncelli, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, *12*(11), 1338–1351.

Prashnani, E., Cai, H., Mostofi, Y., & Sen, P. (2018). PieAPP: Perceptual image-error assessment through pairwise preference. In *IEEE conference on computer vision and pattern recognition* (pp. 1808–1817).

Rajashekar, U., Wang, Z., & Simoncelli, E. P. (2009). Quantifying color image distortions based on adaptive spatio-chromatic signal decompositions. In *IEEE international conference on image processing, IEEE* (pp. 2213–2216).

Raphan, M., & Simoncelli, E. P. (2008). Optimal denoising in redundant representations. *IEEE Transactions on Image Processing*, *17*(8), 1342–1352.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of Optical Society of America*, *62*(1), 55–59.

Safranek, R. J., & Johnston, J. D. (1989). A perceptually tuned subband image coder with image dependent quantization and post-quantization data compression. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 1945–1948).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

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

Sheikh, H. R., Bovik, A. C., & De Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, *14*(12), 2117–2128.

Sheikh, H. R., Wang, Z., Bovik, A. C., & Cormack, L. (2006). Image and video quality assessment research at LIVE. http://live.ece.utexas.edu/research/quality/.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE conference on computer vision and pattern recognition* (pp. 1874–1883).

Simoncelli, E. P., & Adelson, E. H. (1996). Noise removal via bayesian wavelet coring. In *IEEE international conference on image processing, IEEE* (Vol. 1, pp. 379–382).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations* (pp. 1–14).

Snell, J., Ridgeway, K., Liao, R., Roads, B.D., Mozer, M.C., & Zemel, R.S. (2017). Learning to generate images with perceptual similarity metrics. In: *IEEE Transactions on Image Processing*, pp 4277–4281

Sun, J., Xu, Z.,& Shum, H. Y. (2008). Image super-resolution using gradient profile prior. In *ieee conference on computer vision and pattern recognition, IEEE* (pp. 1–8).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. In *International conference on learning representations* (pp. 1–10).

Tao, X., Gao, H., Shen, X., Wang, J., & Jia, J. (2018). Scale-recurrent network for deep image deblurring. In *IEEE conference on computer vision and pattern recognition* (pp. 8174–8182).

Teo, P. C., & Heeger, D. J. (1994). Perceptual image distortion. *Human Vision, Visual Processing, and Digital Display V*, *2179*, 127–141.

Tian, S., Zhang, L., Morin, L., & Déforges, O. (2018). A benchmark of DIBR synthesized view quality assessment metrics on a new database for immersive media applications. *IEEE Transactions on Multimedia*, *21*(5), 1235–1247.

Timofte, R., De Smet, V., & Van Gool, L. (2013). Anchored neighborhood regression for fast example-based super-resolution. In *IEEE international conference on computer vision* (pp. 1920–1927).

Timofte, R., Agustsson, E., Van Gool, L., Yang, M. H., & Zhang, L. (2017). NTIRE 2017 challenge on single image super-resolution: Methods and results. In *IEEE conference on computer vision and pattern recognition* (pp. 114–125).

Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In *IEEE international conference on computer vision, IEEE* (pp. 839–846).

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. In *IEEE conference on computer vision and pattern recognition* (pp. 9446–9454).

Vukadinovic, V., & Karlsson, G. (2009). Trade-offs in bit-rate allocation for wireless video streaming. *IEEE Transactions on Multimedia*, *11*(6), 1105–1113.

Wang, S., Ma, K., Yeganeh, H., Wang, Z., & Lin, W. (2015). A patch-structure representation method for quality assessment of contrast changed images. *IEEE Signal Processing Letters*, *22*(12), 2387–2390.

Wang, S., Rehman, A., Wang, Z., Ma, S., & Gao, W. (2011). SSIM-motivated rate-distortion optimization for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(4), 516–529.

Wang, T., Zhang, L., Jia, H., Li, B., & Shu, H. (2016). Multiscale contrast similarity deviation: An effective and efficient index for perceptual image quality assessment. *Signal Processing: Image Communication*, *45*, 1–9.

Wang, X., Yu, K., Wum S., Gu, J., Liu, Y., & Dong, C.(2018). ESRGAN: Enhanced super-resolution generative adversarial networks. In *European conference on computer vision workshops*

Wang, Z., & Bovik, A. C. (2011). Reduced- and no-reference image quality assessment: The natural scene statistic model approach. *IEEE Signal Processing Magazine*, *28*(6), 29–40.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Z., & Simoncelli, E. P. (2005). Translation insensitive image similarity in complex wavelet domain. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 573–576).

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

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *IEEE Asilomar conference on signals, system and computers* (pp. 1398–1402).

Watson, A. (1993). DCTune: A technique for visual optimization of DCT quantization matrices for individual images. *Society for Information Display Digest of Technical Papers*, *24*, 946–949.

Watson, A. B., Yang, G. Y., Solomon, J. A., & Villasenor, J. (1997). Visibility of wavelet quantization noise. *IEEE Transactions on Image Processing*, *6*(8), 1164–1175.

Wiener, N. (1950). *Extrapolation, interpolation and smoothing of stationary time series: with engineering applications*. Cambridge, MA: MIT Press.

Xue, W., Mou, X., Zhang, L., & Feng, X. (2013). Perceptual fidelity aware mean squared error. In *IEEE international conference on computer vision* (pp. 705–712).

Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2014). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, *23*(2), 684–695.

Yang, C. Y., & Yang, M. H. (2013). Fast direct super-resolution by simple functions. In *IEEE international conference on computer vision* (pp. 561–568).

Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, *19*(11), 2861–2873.

Ye, P., Kumar, J., & Doermann, D. (2014). Beyond human opinion scores: Blind image quality assessment based on synthetic scores. In *IEEE conference on computer vision and pattern recognition* (pp. 4241–4248).

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155.

Zhang, L., Shen, Y., & Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, *23*(10), 4270–4281.

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, *20*(8), 2378–2386.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhang, W., Liu, Y., Dong, C., & Qiao, Y. (2019a). RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 3096–3105).

Zhang, W., Ma, K., Zhai, G., & Yang, X. (2019b). Learning to blindly assess image quality in the laboratory and wild. CoRR arXiv:1907.00516.

Zhang, X., Feng, X., Wang, W., & Xue, W. (2013). Edge strength similarity for image quality assessment. *IEEE Signal Processing Letters*, *20*(4), 319–322.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, *3*(1), 47–57.

Zhou, F., Yao, R., Liu, B., & Qiu, G. (2019). Visual quality assessment for super-resolved images: Database and method. *IEEE Transactions on Image Processing*, *28*(7), 3528–3541.