# Perceptual image quality assessment using a normalized Laplacian pyramid

*Valero Laparra, Center for Neural Science, New York University, USA and Universitat de València, Spain*
*Johannes Ballé, Center for Neural Science and Howard Hughes Medical Institute, New York University, USA*
*Alexander Berardino, Center for Neural Science, New York University, USA*
*Eero P. Simoncelli, Center for Neural Science, Courant Institute of Mathematical Sciences and Howard Hughes Medical Institute, New York University, USA*

## Abstract

*We present an image quality metric based on the transformations associated with the early visual system: local luminance subtraction and local gain control. Images are first decomposed using a Laplacian pyramid, which subtracts a local estimate of the mean luminance at multiple scales. Each pyramid coefficient is then divided by a local estimate of amplitude (weighted sum of absolute values), where the weights are optimized for prediction of local amplitude using (undistorted) images from a separated database. The quality of a distorted image, relative to its undistorted original, is the root mean squared error in this "normalized Laplacian" domain. We show that both luminance subtraction and amplitude division stages lead to significant reductions in redundancy, relative to the original image pixels. We also show that the resulting quality metric provides a better account of human perceptual judgements than either MS-SSIM or a recently-published gain-control metric based on oriented filters.*

## Introduction

Many problems in image processing rely, at least implicitly, on a measure of image quality. Although mean squared error (MSE) is the near-universal choice, it is well known that it is not very well matched to the distortion perceived by human observers [1, 2]. Objective measures of perceptual image quality attempt to correct this by incorporating known characteristics of human perception, such as the dependence of contrast sensitivity on spatial frequency (see reviews [3, 4]). In many cases, this is accomplished by transforming the reference and distorted images into a new format that mimics physiological representations of the early stages of the visual system, and quantifying the root mean squared error within that "perceptual" space. But existing examples skip over the effects of the earliest part of the visual system (the retina and thalamus) and build their transformations based on the properties of primary visual cortex (area V1). Specifically, they typically include multi-scale oriented filtering followed by local gain control to normalize response amplitudes (e.g. [5, 6, 7]).

While these models are motivated by physiology, they also are well matched to the statistical properties of natural images, consistent with theories of biological coding efficiency and redundancy reduction [8, 9]. In particular, application of Independent Component Analysis (ICA) [10], which seeks a linear transformation optimizing statistical independence of the data dimensions, produces oriented filters resembling V1 receptive fields. Local gain control, in a form known as "divisive normalization" that is often used to describe sensory neurons [11], has been shown to decrease the dependencies between neighboring coefficients [12, 13, 14, 15].

To date, the most widely used measure of perceptual distortion is the structural similarity metric (SSIM) [16], which is designed to be invariant to so-called nuisance variables (local mean, local standard deviation) while retaining sensitivity to the remaining "structure" of the image. SSIM is generally used within a multi-scale representation (MS-SSIM), so as to handle features of all sizes [17]. While SSIM is informed by the invariances of human perception, the form of its computation (a product of the correlations between mean-subtracted, variance-normalized, and structure terms) has no obvious mapping onto physiological representation. Nevertheless, the computations that underlie the embedding of those invariances – subtraction of the local mean, and division by the local standard deviation – are reminiscent of the response properties of neurons in the retina and thalamus. In particular, responses of these cells are often modeled as bandpass filters ("center-surround") whose responses are rectified and subject to gain control according to local luminance and contrast (e.g., [18]).

Here, we define a new quality metric, computed as the root mean squared error of an early visual representation based on center-surround filtering followed by local gain control. The filtering is performed at multiple scales, using the Laplacian pyramid [19]. While the model architecture and choice of operations are motivated by the physiology of the early visual system, we use a statistical criterion to select the local gain control parameters. Specifically, the weights used in computing the gain signal are chosen so as to minimize the conditional dependency of neighboring transformed coefficients. Despite the simplicity of this representation, we find that it provides an excellent account of human perceptual data, outperforming MS-SSIM, as well as V1-inspired models, in predicting the human quality judgments in the TID 2008 database [20].

## Normalized Laplacian model

Our model is comprised of two stages (figure 1): the image $\boldsymbol{x}$ is subjected to local luminance removal, which is implemented by subtracting a local estimate of the mean, followed by a local gain control, which is implemented by dividing by a local estimate of fluctuation around the mean. The perceptual metric is then simply the root mean squared error in this transformed domain.

We view the local luminance subtraction and contrast normalization as a means of reducing redundancy in natural images. Most of the redundant information in natural images is local; i.e.,

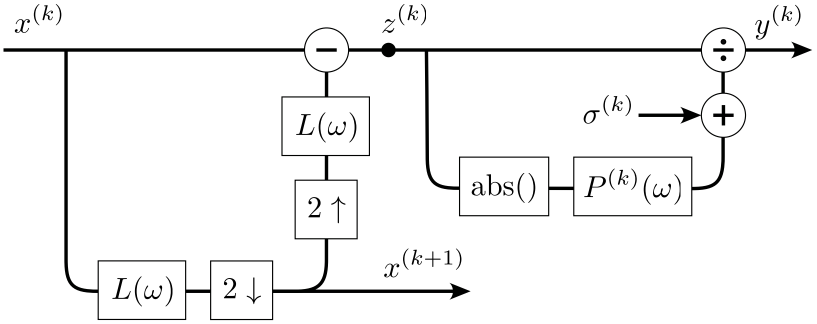

***Figure 1.*** *Normalized Laplacian model diagram, shown for a single scale $(k)$. Starting from image, $x^{(k)}$ ($k=1$ corresponds to the original image), the intermediate image $z^{(k)}$ is computed by subtracting the local mean (eq. 2). This is accomplished using the standard Laplacian pyramid construction: convolve with lowpass filter $L(\omega)$, downsample by a factor of two in each dimension, upsample, convolve again with $L(\omega)$, and subtract from the input image $x^{(k)}$. This intermediate image is then normalized by an estimate of local amplitude , obtained by computing the absolute value, convolving with scale-specific filter $P^{(k)}(\omega)$, and adding the scale-specific constant $\sigma^{(k)}$ (eq. 3)). The downsampled image $x^{(k+1)}$ forms the starting image for scale $(k+1)$.*

the distribution of an image pixel ($x_i$) conditioned on all others can be well approximated by the conditional

$$p(x_i|\boldsymbol{x}_{Ni}), \tag{1}$$

where $\boldsymbol{x}_{Ni}$ is its immediate neighborhood. This is a form of Markov property and motivates our use of local mean and amplitude estimates. In each stage of the model, a parametric estimate of a statistic of the central pixel is gathered from its neighbors, and then removed; in the first stage, this statistic is the mean $f_L$:

$$z_i = x_i - f_L(\boldsymbol{x}_{Ni}), \tag{2}$$

and in the second stage, the amplitude $f_C$:

$$y_i = z_i / f_C(\boldsymbol{z}_{Ni}). \tag{3}$$

Decomposition of an example image is shown in figure 2. All transformations in the model are translation invariant (i.e., the parameters of the two operations are identical for all locations). This considerably reduces the number of parameters of the model.

### Luminance subtraction stage

We used the Laplacian pyramid [19] to implement luminance subtraction. This effectively decomposes the image using a multiscale array of linear "difference of Gaussians" bandpass filters. At each scale, a lower resolution (blurred) version of the image is computed, and subtracted, implementing equation 2 (see fig. 1). The process is then applied recursively to the blurred (and downsampled) image. This linear stage has no free parameters, except for the number of scales, $N$, which is chosen according to the resolution of the images (for examples in this paper $N = 6$).

### Contrast normalization stage

As an estimate of the local amplitude around the mean, we use a linear combination of rectified neighbors:

$$f_C^{(k)}(\boldsymbol{z}_{Ni}) = \sigma^{(k)} + \sum_{j \in Ni} p_j^{(k)} \left| z_j^{(k)} \right|, \tag{4}$$

where $\boldsymbol{p}^{(k)}$ is the vector of weights used at scale $k$. We constrain the elements of this vector to be non-negative and introduce a small constant $\sigma^{(k)}$ such that $f_C^{(k)}(\boldsymbol{z}_{Ni})$ is guaranteed to be positive for all neighborhoods, avoiding division by zero. For each scale, the constant is set to the average of the absolute value:

$$\sigma^{(k)} = \frac{1}{Ns^{(k)}} \sum_{i=1}^{Ns^{(k)}} \left| z_i^{(k)} \right|, \tag{5}$$

where $N_s^{(k)}$ is the number of coefficients in the subband at scale $k$. The weight vector is chosen as the solution of the optimization problem

$$\hat{\boldsymbol{p}}^{(k)} = \arg\min_{\boldsymbol{p}} \sum_{i=1}^{N_s^{(k)}} \left( \left| z_i^{(k)} \right| - f_C(\boldsymbol{z}_{Ni}^{(k)}) \right)^2. \tag{6}$$

Note that $\sigma^{(k)}$ may be considered as an initial approximation for the absolute value of $z$, and the weighted sum acts to adaptively tune this value. All parameters are optimized over a large set of (undistorted) images from the McGill natural image database [21]. This optimization was performed only on these undistorted images, with no access to information about the type of distortions nor perceptual data on which we subsequently tested the model.

### Distance metric

Finally, our proposed perceptual metric is given by:

$$D(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sqrt{N_s^{(k)}}} \left\| \boldsymbol{y}^{(k)} - \tilde{\boldsymbol{y}}^{(k)} \right\|_2, \tag{7}$$

where $\boldsymbol{y}^{(k)}$ and $\tilde{\boldsymbol{y}}^{(k)}$ denote vectors containing the transformed reference and distorted image data, respectively. Note that we compute root mean squared error for each scale, and then average over these, effectively giving larger weight to the lower frequency coefficients (which are fewer in number, due to subsampling).

## Results

To evaluate our model, we fit the parameters of the gain control (eq. 6) using the McGill image dataset [21] and then evaluated

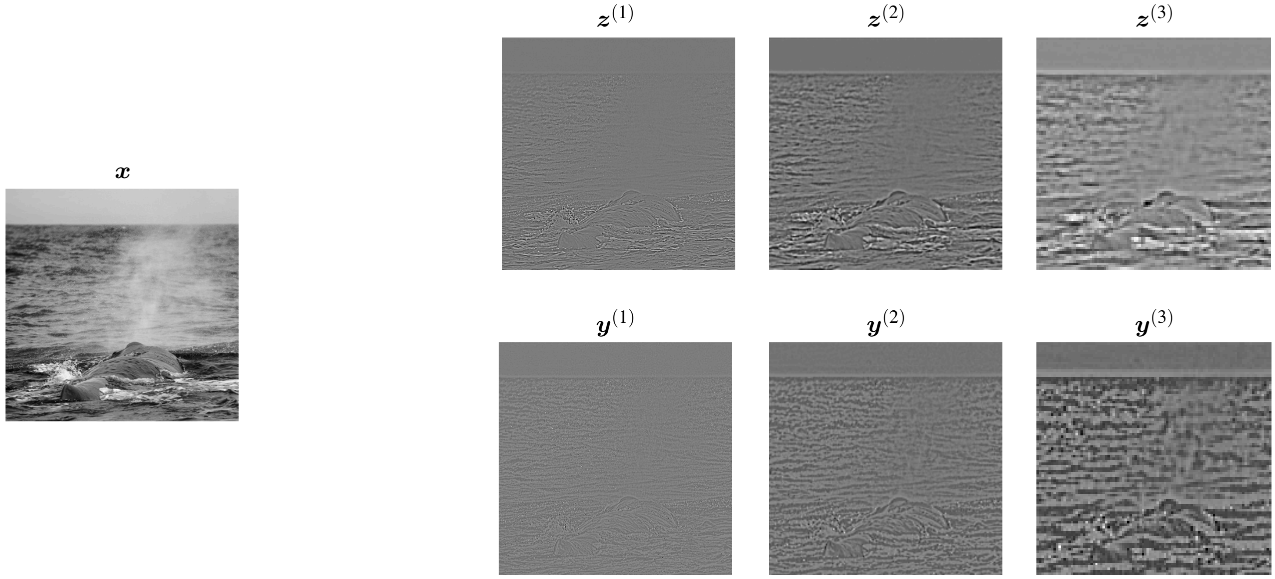

***Figure 2.*** *Representation of an example image. $\boldsymbol{x}$ is the original image (left). $\boldsymbol{z}$ is the decomposition of the image using the Laplacian pyramid (three scales shown), each image corresponding to a different scale. Note that the Laplacian pyramid includes downsampling in each scale. The examples shown here have been upsampled for visualization purposes. $\boldsymbol{y}$ are the corresponding locally contrast-normalized images.*

the resulting model in two different ways [1]. First, we examined the ability of each stage of the normalized Laplacian transform to reduce redundant information between the central coefficient and its neighbors. Second, we compared the model distances to human perceptual responses over a large set of distorted images.

## Mutual information measurements

Figure 3 illustrates the reduction of redundant information at each stage of the model. Each image shows the empirical pairwise mutual information [22] between a given coefficient (central pixel of each image) and each of its neighbors. Mutual information has been computed using one million samples from the reference images in the TID database [20]. The figure reports the results for the first scale – results for the other scales are similar. The information reduction from both stages of processing is seen to be quite substantial – a factor of roughly six and three, respectively.

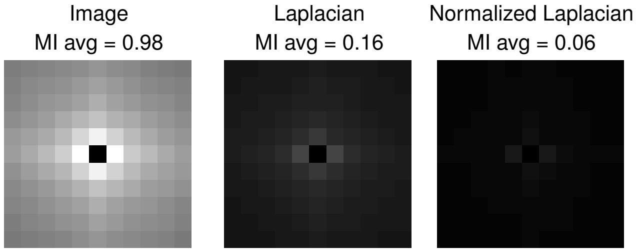

***Figure 3.*** *Local mutual information between neighbors. The brightness of each pixel is proportional to the mutual information between a central coefficient and the neighbor at that relative location. Values are estimated from one million image patches. The average mutual information over the whole neighborhood is given above each panel.*

## Image quality assessment

In this experiment, we analyze how well our perceptual metric correlates with human reports of perceptual distortion. We use a grayscale version of the TID database [20] which consists of 1700 different distorted images (17 different distortion types, at 4 different strengths, for each of 25 original images), each with its own mean opinion score (MOS). The MOS represents the mean distortion rating of all participants who assessed a particular distorted image. For reference, we also show the results of measuring the RMSE between the reference and the distorted image in the image domain, as well as in the Laplacian domain. We compare the performance of our metric with the multi-scale version of the most widely used perceptual metric, MS-SSIM [17]. We also compare with a metric based on a V1-based model [7], which uses an oriented wavelet decomposition followed by divisive normalization. The parameters of this model were chosen based on a separate set of perceptual measurements.

Figure 4 shows DMOS (which is inversely proportional to the MOS) against the predictions of each of the distance metrics. We present three different numerical evaluations of the predictive ability of each metric. The first ($\rho_1$) is the Pearson correlation between each metric and the DMOS (correlations are, up to the sign, identical for MOS and DMOS). To compute the second ($\rho_2$) and the third (RMSE), we first fitted generalized logistic functions with four parameters to the measurements (black line). Then, we computed the DMOS prediction from that curve for each image and evaluated its correlation and the RMSE respectively.

As expected, the RMSE in the image domain gives the worst correlation with human perception. This is a classical result and is the primary motivation for seeking better perceptual metrics. MS-SSIM and the V1 model perform comparably, with the V1 model exhibiting slightly better correlation ($\rho_1$), but MS-SSIM providing slightly better prediction error. The normalized Lapla-

[1] Code with the implementation of the proposed model can be found at `http://www.cns.nyu.edu/~valero/NLP/NLP.html`

cian metric achieves notable improvements in Pearson correlations and prediction error. Note that this is particularly surprising given that both the V1 model and MS-SSIM are optimized for perceptual performance, whereas the normalized Laplacian model parameters were optimized for statistical performance on an independent database of (undistorted) natural images. In addition, the logistic regression for the normalized Laplacian model (as well as the V1 model) is almost linear. Finally, note that most of the performance is derived from the nonlinear normalization stage: the unnormalized Laplacian offers only a modest improvement over RMSE in the image domain.

## Discussion

We have presented a perceptual quality metric computed as the root mean squared error of images represented in a nonlinear multi-scale decomposition in which the local mean and amplitude have been removed, with parameters optimized to remove redundancy in natural images. We have shown that this representation accomplishes a significant reduction of redundancy, and transforms the data to a more perceptually relevant space. In particular, the model provides a better account of human perceptual quality judgements than either the widely-used MS-SSIM metric, or a biologically-inspired V1 model based on locally normalized responses of oriented filters. We expect this performance gap could be increased by choosing model parameters that optimize the fit to the human distortion ratings.

A number of previous image quality metrics have used local gain control [5, 6, 7], but all of them did so in the context of an oriented linear transform. Despite a large body of work that has been interpreted as evidence that oriented linear filters are an optimal choice for capturing statistical regularities in images [23, 24], several articles have suggested that this optimum is shallow [25], and that non-oriented filters are nearly as effective [26, 27]. The comparisons of Fig. 4 suggest that the non-oriented bandpass representation that we propose here may offer a better substrate for a quality metric than an oriented representation. But this is a preliminary finding, and a more thorough comparison is needed. An interesting possibility is that a cascaded representation, in which the normalized Laplacian pyramid is followed by further decomposition into oriented subands (and possibly another stage of local gain control), would be consistent with the stages of the human visual system, and may prove an even stronger platform on which to build a quality metric.

It is worth emphasizing that the normalized Laplacian model parameters are optimized to minimize redundancies in the representation of undistorted natural images. Thus, although they embed no specific knowledge of the types of distortions on which the model is tested, they do capture important information about the statistical properties of natural images. This suggests that the model might be useful as a platform for a no-reference image quality metric. Specifically, one could use a measure of mutual information (or another measure of statistical independence) of the normalized Laplacian representation of an image, to quantify its "naturalness" (and thus, the level of distortion), similar to the work in [28].

Finally, the TID database provides a useful, but limited, means of assessing the performance of a metric in matching human judgements. The set of distortions includes only those encountered in typical image processing settings, and the human responses are quite variable across observers. A more directed assessment can arise from examining artificial images, synthesized to maximize or minimize the distortions of one metric while holding constant another (MAD competition [29]). For example, images that maximize/minimize our metric while adhering to a fixed RMSE distortion in the pixel domain would provide a direct visualization of the types of error that the metric deems "worst" and "best". And such "adversarial" images generated with the normalized Laplacian metric while holding the MS-SSIM constant (or vice versa), would provide a direct visualization of how the two models differ, and thus an effective means of distinguishing the advantages and disadvantages of each.

## References

[1] Girod, B., “Digital images and human vision,” ch. What’s Wrong with Mean-squared Error?, 207–220, MIT Press (1993).

[2] Wang, Z. and Bovik, A., “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” *Signal Processing Magazine* **26**, 98–117 (Jan. 2009).

[3] Eskicioglu, A. M. and Fisher, P. S., “Image quality measures and their performance,” *IEEE Trans. Communications* **43**, 2959–2965 (Dec. 1995).

[4] Eckert, M. P. and Bradley, A. P., “Perceptual quality metrics applied to still image compression,” *Signal Processing* **70**, 177–200 (Nov. 1998).

[5] Teo, P. and Heeger, D., “Perceptual image distortion,” in [*Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*], **2**, 982–986 vol.2 (Nov 1994).

[6] Watson, A. B. and Solomon, J. A., “Model of visual contrast gain control and pattern masking,” *Journal of Optical Society of America* **14**(9), 2379–2391 (1997).

[7] Laparra, V., Muñoz Marí, J., and Malo, J., “Divisive normalization image quality metric revisited,” *JOSA A* **27**(4), 852–864 (2010).

[8] Barlow, H. B., “Possible principles underlying the transformation of sensory messages,” in [*Sensory Communication*], Rosenblith, W. A., ed., 217–234, MIT Press, Cambridge, MA (1961).

[9] Simoncelli, E. P. and Olshausen, B., “Natural image statistics and neural representation,” *Annual Review of Neuroscience* **24**, 1193–1216 (May 2001).

[10] Bell, A. J. and Sejnowski, T. J., “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation* **7**, 1129–1159 (Nov. 1995).

[11] Carandini, M. and Heeger, D. J., “Normalization as a canonical neural computation.,” *Nature Reviews Neurosci* **13**, 51–62 (Jan. 2012).

[12] Schwartz, O. and Simoncelli, E., “Natural signal statistics and sensory gain control,” *Nat. Neurosci.* **4**(8), 819–825 (2001).

[13] Malo, J., Epifanio, I., Navarro, R., and Simoncelli, E. P., “Nonlinear image representation for efficient perceptual coding,” *IEEE Transactions on Image Processing* **15**(1), 68–80 (2006).

[14] Malo, J. and Laparra, V., “Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images,” *Neural Computation* **22**(12), 3179–3206 (2010).

[15] Lyu, S., “Dependency Reduction with Divisive Normalization: Justification and Effectiveness,” *Neural Computation* **23**(11), 2942–2973 (2011).

[16] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Im. Proc.* **13**, 600–612 (Apr 2004).

[17] Wang, Z., Simoncelli, E. P., and Bovik, A. C., “Multi-scale struc-

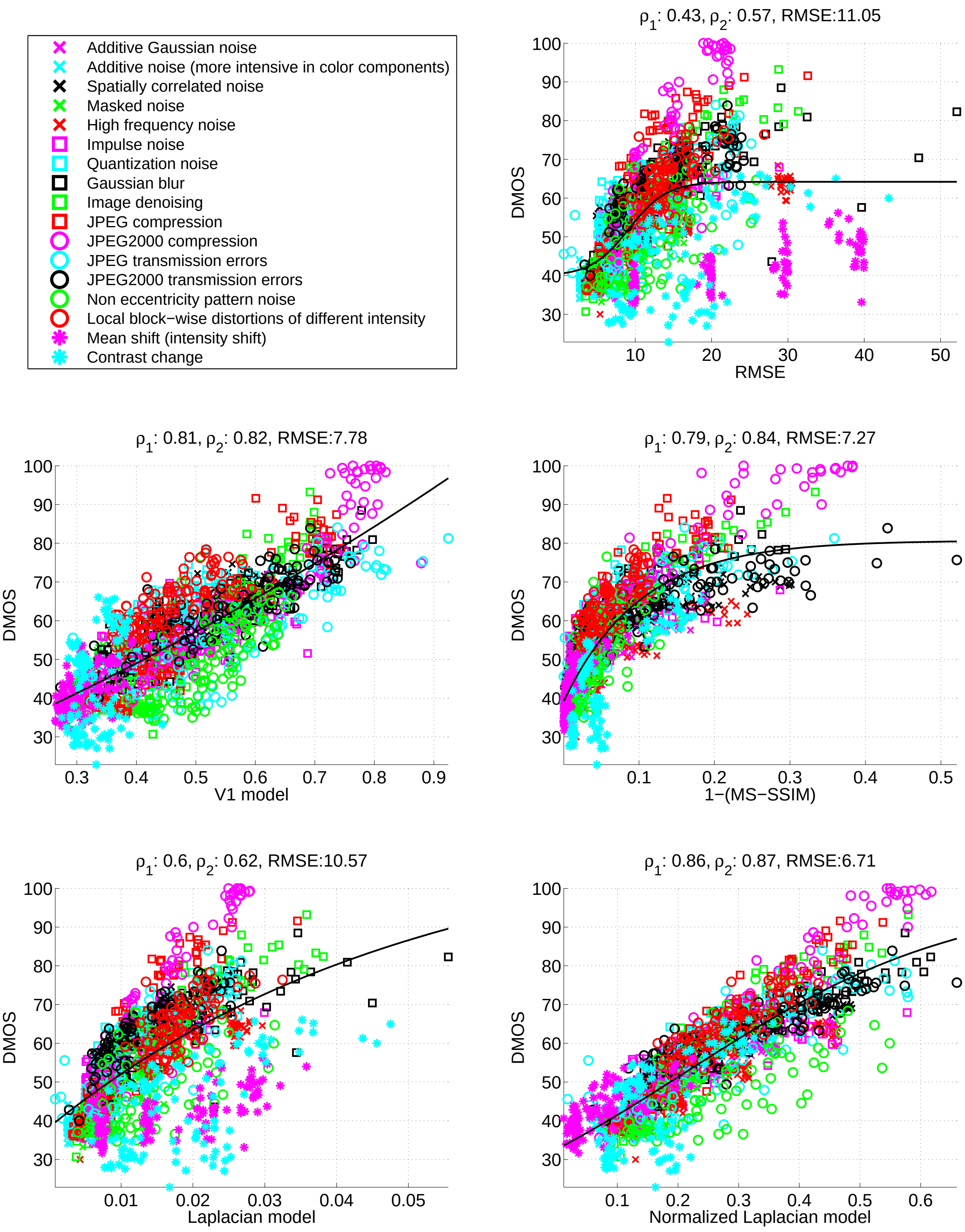

***Figure 4.*** *Comparison of metrics to human perceptual data. Each plot shows the inverse of the mean opinion score of human observers (DMOS) as a function of prediction of a quality metric, for 1700 images corrupted by different types and magnitudes of distortion (see key, first row left). Performance of each metric is summarized with three numbers (provided above each plot): the Pearson correlation before fitting a logistic function ($\rho_1$), the Pearson correlation ($\rho_2$) and the prediction error (RMSE) after fitting a logistic function (black line). First row right: root mean square error (RMSE) in the image domain. Second row left: MSE in a normalized oriented V1 model [7]. Second row right: multi-scale structural similarity index (MS-SSIM) [16]. Third row left: RMSE in the Laplacian pyramid domain. Third row right: RMSE in the normalized Laplacian domain (eq. 7).*

tural similarity for image quality assessment," in [*in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*], 1398–1402 (2003).

[18] Mante, V., Bonin, V., and Carandini, M., "Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli," *Neuron* **58**(4), 625 – 638 (2008).

[19] Burt, P. J., Edward, and Adelson, E. H., "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications* **31**, 532–540 (1983).

[20] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F., "Tid2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics* **10**, 30–45 (2009).

[21] Olmos, A. and Kingdom, F. A. A., "A biologically inspired algorithm for the recovery of shading and reflectance images," *Perception* **33**, 1463–1473 (2004).

[22] Cover, T. M. and Thomas, J. A., [*Elements of Information Theory 2nd Edition*], Wiley-Interscience, 2 ed. (July 2006).

[23] Olshausen, B. and Field, D., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, 607–609 (1996).

[24] Bell, A. and Sejnowski, T., "The "independent components" of natural scenes are edge filters," *Vision Research* **37**(23), 3327–3338 (1997).

[25] Baddeley, R., "Searching for filters with "interesting" output distributions: An uninteresting direction to explore," *Network* **7**, 409–421 (1996).

[26] Bethge, M., "Factorial coding of natural images: how effective are linear models in removing higher-order dependencies?," *Journal of the Optical Society of America A* **23**(6), 1253–1268 (2006).

[27] Lyu, S. and Simoncelli, E. P., "Nonlinear extraction of 'independent components' of natural images using radial Gaussianization," *Neural Computation* **21**, 1485–1519 (Jun 2009).

[28] Mittal, A., Moorthy, A. K., and Bovik, A. C., "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process* , 4695–4708 (2012).

[29] Wang, Z. and Simoncelli, E. P., "Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision* **8**(12), 8 (2008).

## Author Biography

*Valero Laparra was born in València (Spain) in 1983, and received a B.Sc. degree in Telecommunications Engineering (2005), a B.Sc. degree in Electronics Engineering (2007), a B.Sc. degree in Mathematics degree (2010), and a PhD degree in Computer Science and Mathematics (2011). He is a postdoc in the Image Processing Laboratory (IPL) at Universitat de València, and currently a visiting postdoc in the Laboratory for Computer Vision at NYU, USA. More details in http://www.uv.es/lapeva.*

*Johannes Ballé received his Dipl.-Ing. degree in Computer Engineering from RWTH Aachen University, Germany in 2007 and his Dr.-Ing. degree in Electrical Engineering (summa cum laude) from the same university in 2012, specializing in image and video compression. He is currently working as a research associate at New York University's Center for Neural Science. His research interests include image statistics and processing, computational vision, information theory, as well as machine learning.*

*Alexander Berardino received his B.A. in Neuroscience with a concentration in computation and cognitive systems from Boston University in 2010. He subsequently spent 2 years as a research fellow in Biomedical Engineering at B.U. (2010-2012). He is currently pursuing his Ph.D. in Neural Science at New York University, in the lab of Eero Simoncelli. His research interests include the intersection of image statistics and biological computation, as well as the analysis of mid-to-high level biological visual representations.*

*Eero P. Simoncelli (S'90–M'91–SM'04–F'09) received the B.S. degree summa cum laude in physics from Harvard University in 1984, studied applied mathematics at University of Cambridge for a year and a half, and received M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, in 1988 and 1993, respectively. He was an Assistant Professor of Computer and Information Science at the University of Pennsylvania from 1993 to 1996. In 1996, he moved to New York University, where he is currently a Professor in neural science and mathematics. In 2000, he became an Investigator of the Howard Hughes Medical Institute under their new program in computational biology. His research interests span a wide range of topics in the representation and analysis of visual images, in both machine and biological systems.*