Towards Perceptually Plausible Training of Image Restoration Neural Networks

Ali Ak, Patrick Le Callet

Towards Perceptually Plausible Training of Image Restoration Neural Networks

1st Ali Ak
Image Perception Interaction Team
LS2N, University of Nantes
Nantes, France
ali.ak@univ-nantes.fr

2nd Patrick Le-Callet
Image Perception Interaction Team
LS2N, University of Nantes
Nantes, France
patrick.lecallet@univ-nantes.fr

***Abstract*—Learning-based black-box approaches have proven to be successful at several tasks in image and video processing domain. Many of these approaches depend on gradient-descent and back-propagation algorithms which requires to calculate the gradient of the loss function. However, many of the visual metrics are not differentiable, and despite their superior accuracy, they cannot be used to train neural networks for imaging tasks. Most of the image restoration neural networks rely on mean squared error to train. In this paper, we investigate visual system based metrics in order to provide perceptual loss functions that can replace mean squared error for gradient descent-based algorithms. We also share our preliminary results on the proposed approach.**

***Index Terms*—visual perception, neural networks, just noticeable difference**

I. Introduction

Visual metrics has numerous use cases in visual processing domain. They play an important role in the development, evaluation, and optimization of many visual processing algorithms. There are various approaches to develop visual metrics. While some of the metrics focus on signal driven calculations [1] [8] [10], some focus on modeling the visual system [2] [3]. Metrics which relies on signal driven calculations model the quality perception as a continuous function. On the other hand, Visual Model based metrics, such as VDP [3] and HDR-VDP [2], can predict the perceptual quality of the images more accurately and tuned on the Just Noticeable Differences around near threshold values. Although they are more accurate, they have a high computational complexity since they are derived from different components of Human Visual System (HVS) where the data is collected from a set of psychophysical measurements. Additionally, this complexity results in non-differentiable models which prevents them to be used in many visual processing applications.

Visual model based metrics can estimate the probability of the detection of differences between a pair of images for an observer. They are generally tuned towards the near-threshold Just Noticeable Differences. This provides an accurate judgment of perception by providing information regarding statistically significant differences between a pair of stimuli.

978-1-7281-3975-3/19/$31.00 ©2019 European Union

HVS is not sensitive to the low amount of differences on a pair of stimuli. Several factors affect the perception of a difference. Under certain conditions, the minimum amount of difference required to be perceived is called the Just Noticeable Difference(JND). However, most of the existing visual metrics provide a continuous quality score for a varying amount of distortions. Visual model based metrics such as HDR-VDP provides more perceptually plausible predictions by focusing on just noticeable differences.

Neural networks which are trained for image restoration tasks such as denoising, compression, super-resolution, view synthesis uses loss functions during the training. Most of the current approaches rely on MSE. This results in blurry high-frequency details, and noisy flat regions depending on the task.

There are several attempts to tackle the lack of perceptual training of image restoration tasks. Zhao et al. provide an in-depth analysis of loss functions in image restoration tasks [4]. In their paper, authors also provide a differentiable version of SSIM as an alternative loss function which also fails to provide perceptually plausible results.

Lately, Perceptual loss has been introduced by Johnson et al [6]. They used layer activations of deep neural networks that have been trained for image classification tasks on millions of images such as VGG [12] and ResNet [13]. Activations of different layers have been used as a loss function in image restoration tasks which provides perceptually more plausible results compared to MSE.

Goodfellow et al. proposed a unique network scheme called Generative Adversarial Networks(GANs) [5]. GANs are deep neural network architectures contains two individual networks, computing with each other. Generative network produces an image to fool the discriminator that it is a real image while discriminator tries to predict the inputted image is from the real distribution or a generated one. This training scheme allows for capturing any distribution of data. For image restoration tasks, GANs combined with traditional loss functions provide perceptually plausible results compared to former alternatives. To keep output image similar to input, GANs need to incorporate with a loss function such as perceptual loss or MSE.

Despite the success of learning-based black-box approaches at image restoration tasks, there is still a need for perceptually plausible training schemes. Complexity and non-differentiable

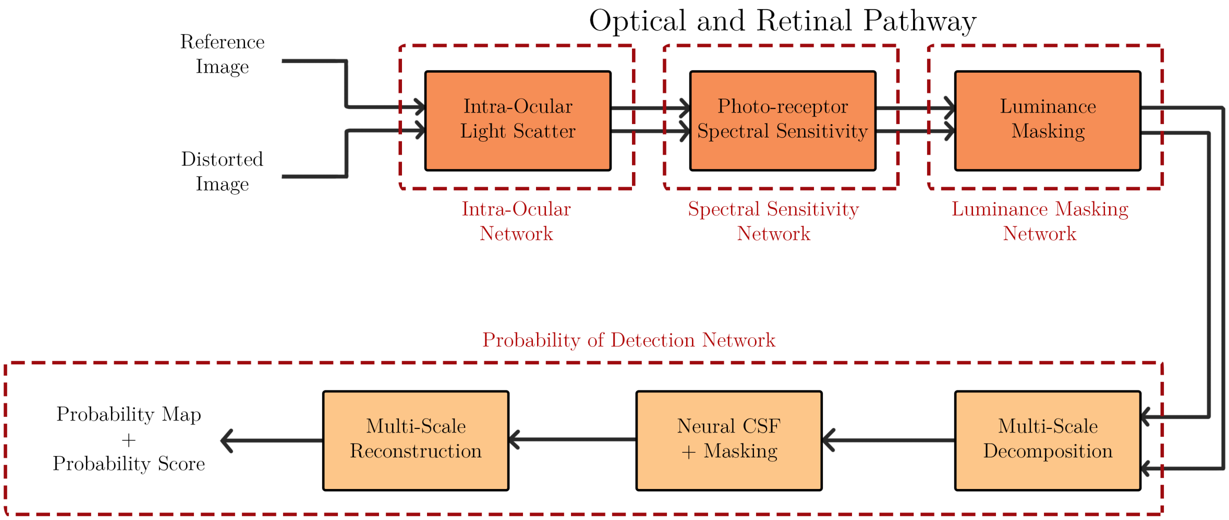
Optical and Retinal Pathway
Reference Image
Distorted Image
Intra-Ocular Light Scatter
Photo-receptor Spectral Sensitivity
Luminance Masking
Intra-Ocular Network
Spectral Sensitivity Network
Luminance Masking Network
Probability of Detection Network
Probability Map + Probability Score
Multi-Scale Reconstruction
Neural CSF + Masking
Multi-Scale Decomposition

Fig. 1. HDR-VDP [2] box diagram with our proposed scheme. Red dashed rectangles represent individual networks in our approach. Each colored box represents different modules from HDR-VDP [2]. Optical and Retinal Pathway of HDR-VDP is color coded with darker orange.

nature of the HVS based visual metrics prevent them from being used as a loss function for the training of neural networks. In this paper, we will investigate different modules from HVS based visibility metric, HDR-VDP [2], and how we can replicate them with neural networks. Later, we will share our preliminary results and conclude with our findings.

II. HDR-VDP Overview

While an arbitrary amount of distortion is perceivable under specific viewing conditions, it might not be noticeable under another. There are many factors involved in the perception of distortions [16], such as viewing density, display luminance, background luminance, and individual observer differences. Many of these factors have been well explored and modeled to some extent in the literature.

We will investigate some of these factors and how they are modeled in HDR-VDP [2]. HDR-VDP is a comprehensive model of the human visual system. It is built with several modules, such as Intra-ocular Light Scattering, Photo-receptor Spectral Sensitivity, Luminance Masking, Multi-scale Decomposition, Neural Contrast Sensitivity, Masking, Multi-scale Reconstruction. It is designed to estimate perceivable differences for near-threshold distortions.

Working with gamma-corrected images as input and assuming that the pixel values are scaled perceptually uniform may mislead the perceptual evaluation of produced images. The following formula is used to convert the pixel intensity values into absolute luminance values.

$$Luma = 0.2126 \times R + 0.7152 \times G + 0.0722 \times B$$

$$L = (L_{Peak} - L_{Black}) \times Luma^{(g)} + L_{Black}$$

Where R, G, B values are first converted into grayscale ($Luma$) and according to display device's Peak Luminance (L_{Peak}), Black level Luminance (L_{Black}) and gamma (g), final Luminance values can be calculated for each pixel in terms of cd/m^2.

Optical and Retinal Pathway of the HVS modeled as a combination of separate modules in HDR-VDP [2]. The first module is intra-ocular light scatter in which the scattered light is passing through the cornea, lens, and eye chamber is modeled [7]. This process results in a loss of contrast of the perceived image(scene).

After the intra-ocular light scatter module, photo-receptor spectral sensitivities are modeled as non-linear functions. It describes the probability of L, M, S cones, and rods on the retina sensing each particular wavelength.

Photo-receptor spectral sensitivity module is followed by Luminance masking in HDR-VDP [2]. Luminance masking accounts for the local luminance adaptation of photo-receptors. Most of the visual metrics assume a global adaptation of global adaptation to luminance values, while photo-receptors has a highly non-linear response to different luminance levels [10]. This adaptational sensitivity to luminance is captured in the contrast sensitivity function, and it is derived from the measurements by Mantiuk et al [15]. Afterward, achromatic responses are calculated as the sum of L, M cones, and rod responses acquired from Luminance adaptation module.

Psychophysical studies suggest that HVS responses to each spatial frequency and orientation with a different sensitivity [16]. Many visual metrics employ one of several multi-scale approaches. In Daly's VDP Cortex Transform [17] is used while Steerable Pyramid [18] is used in HDR-VDP [2]. After multi-scale decomposition, neural CSF function is applied to each spatial frequency and orientation bands individually.

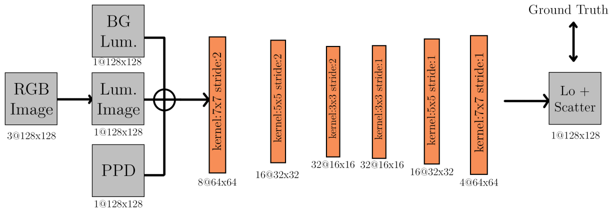

Fig. 2. Intra-Ocular Light Scatter Neural Network box diagram.

Neural CSF is signal independent and can be derived from CSF measurements which are acquired from psychophysical experiments. Before applying, optical and luminance dependent components are excluded (by division) from fitted CSF function since they are already modeled in earlier stages. Contrast masking is also applied in three different parts as self masking and masking across orientations and masking due to two neighboring frequencies. Afterward, the psychometric function is applied, and probability summation is done as summation across all frequencies.

This visual model provides a probability of detection map as well as taking the max value over an image, produces a probability of detection score. Due to the lack of benefit in our task, we skipped the quality score branch of the HDR-VDP. Authors also provided an extension for the quality score.

III. Optical and Retinal Pathway Neural Network Implementation

Optical and Retinal Pathway modules are jointly modeled and trained with a combined loss function of each three individual modules. General scheme of the model is summarized in Fig. 1. Each network receives its inputs from the previous module as it has been implemented in HDR-VDP. An equally weighted loss function has been used during the training where the ground truth data has been acquired from HDR-VDP. In each of the Optical and Retinal Pathway neural networks, reference images and test images are inputted separately as it is implemented in HDR-VDP. For the following modules, we have used an auto-encoder neural network architecture since we want each module to represent their corresponding white-box outputs accurately. Using auto-encoder for each module in the Optical and Retinal Pathway allowed us to keep the input and output size the same. Additionally, it allowed us to train all three networks with a combined loss function, and by restricting the networks on individual tasks, it allowed us to subdivide the problem. In each of the following subsections, we will provide our incentives for the neural network structures and provide details about the individual neural network structures.

A. Intra-Ocular Light Scatter

In HDR-VDP, this module is modeled as a modulated transfer function. Intra-ocular light scatter module works on spectral values, and in an ideal model, there should be a small wavelength dependency, however, most studies show that there is little to none wavelength dependency in intra-ocular light scatter [8]. This allows using the same MTF for each input radiance map. In HDR-VDP, authors used the psychophysical measurements to fit the function. Then, this function is used to create MTF kernels for the required size.

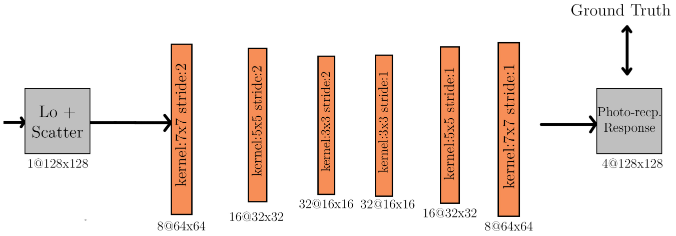

Fig. 3. Photo-receptor Spectral Sensitivity Neural Network box diagram.

MTF kernel is multiplied by the input spectral values in the frequency domain. As it can be observed in Fig. 2, we have used six convolution layers in this network to capture the same receptive field with the MTF kernel in HDR-VDP. We have used an auto-encoder structure and used bilinear interpolation instead of deconvolution in the decoder network to prevent checkerboard artifacts [19]. ReLu activation functions have been used for non-linearity. This helped us to achieve a computationally more efficient and differentiable replicate of the module with similar accuracy.

B. Photo-receptor Spectral Sensitivities

In HDR-VDP, Photo-receptor Spectral sensitivities are modeled as non-linear functions for each photo-receptor type separately over the wavelength range. Photo-receptor sensitivities are acquired via multiplying the fitted functions with each channel.

As it can be observed from Fig. 3, in our neural network implementation, the output of the Intra-Ocular Light Scatter Network is inputted directly to this network as it is implemented in HDR-VDP. We have modeled this module as a combination of convolution layers and ReLu activation functions (for non-linearity) in the network. Combination of ReLus provided us a fair estimation of the non-linearity of spectral sensitivity curves.

C. Luminance Masking

To learn the non-linear adaptation responses, we have used a similar auto-encoder scheme again. Instead of ReLu, we have used SeLu activation functions as non-linearity after each convolution layers. We have adapted SeLu slope by adjusting α value in the SeLu function according to the data acquired from the HDR-VDP [2]. We have combined Luminance Masking and Achromatic Responses modules together in our implementation since Achromatic responses are implemented as an equally weighted summation of luminance masking output channels in HDR-VDP [2].

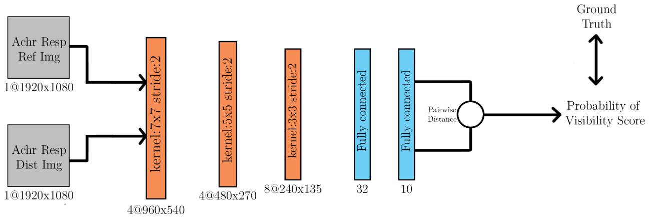

Fig. 4. Probability of detection neural network architecture. Both reference and distorted images go trough the network separately and network is trained against the pairwise distance between the extracted vectors.

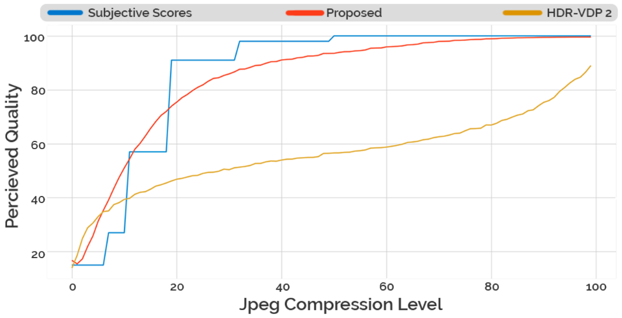

Fig. 5. Predicted Visibility Scores from proposed metric vs HDR-VDP on a source image from MCL-JCI dataset.

IV. Probability of Detection Network

Instead of predicting a probability of detection map, in our first implementation, we chose to acquire just a probability of detection score. This allowed us to try and train image restoration networks by just replacing common loss functions.

Currently, we are working on implementing Steerable Pyramid spatial frequency scales and orientation filters. Instead of training the network freely, we aim to use the existing knowledge from the white-box approach, which we already used as ground truth.

To train our probability of detection neural network, we have used MCL-JCI dataset [20]. MCL-JCI dataset contains 50 source images compressed where each of them compressed by JPEG algorithm with 100 different Q levels. Subjective tests are conducted on the dataset where subjects are asked if they notice any distortion between a reference image and a distorted version of it. JND staircase functions are acquired via subjective experiment. After experimenting with the dataset, to focus on near-threshold distortions, we have used only the first JND step information. We have extracted 50×100 image pairs to compare with $1920 \times 1080px$ size. We have assigned each image pair with a score between [0,1] depending on the agreement of the population about the visibility of the distortion.

We have used a Siamese neural network architecture in our implementation. Siamese networks have several advantages. They don't require as much as data compared to other neural networks architectures. They learn similarity functions, so instead of a classification task, they are trained on discriminating which directly related to our task.

Achromatic responses of both reference and distorted images, which acquired from the previous neural networks, are inputted to the probability of detection network. The network is trained with data generated from MCL-JCI dataset. Due to large size of input images ($1920 \times 1080px$), we used a narrow architecture. Pairwise distance between the output vector of input images is trained against the visibility score extracted from MCL-JCI dataset.

V. Training Details

Optical and Retinal Pathway implementation contains three similar auto-encoder neural networks. Each auto-encoder produces an intermediate result. Their ground truth data is acquired from HDR-VDP [2]. We have created a dataset by using HDR-VDP. We have created a dataset of 300000 images with $128 \times 128px$ size, and varying background luminance, pixel per degree. We have used 0.0005 learning rate during the training and adaptive momentum optimizer(Adam) for the training. We have used 16 as batch size. We have trained the network over 500 epochs.

After training of the Optical and Retinal Pathway, we have used the network in evaluation mode to process the input reference and distorted images for the training of probability of detection network. We have used the same learning rate of 0.0005 and Adam optimizer for the training. Due to the high resolution of input images, we have used 8 batch size during training. We have used Pytorch library for the implementation.

VI. Preliminary Results

We have processed 100 images ($1920 \times 1080px$ sized) with both HDR-VDP 2 original implementation and our proposed method. On average we have acquired a 120 times speed improvement with the same computer specifications.

In Fig. 5, we can see the visibility score predictions of HDR-VDP and proposed approach. The horizontal axis represents the source image with 100 Q levels of JPEG distortions. From left to right, the Q parameter of the JPEG compression algorithm increases. On the vertical axis, quality score predictions are displayed. Blue line is subjective scores from MCL-JCI dataset, the red line is our prediction, and the yellow line is HDR-VDP 2 predictions. As can be observed, our predictions are better aligned with subjective opinions.

After confirming our perceived quality predictions, we have used our model to train image restoration neural network algorithms. In Fig. 6, We have shared our preliminary results for a denoising network on MNIST dataset. Input images with noise, ground truth images, the output of a neural network trained with our model as an objective function and the same network trained to minimize MSE has been provided. The neural network has an auto-encoder structure, and while trained with our proposed metric and MSE, all the hyper-parameters are kept same. The network trained with proposed metric outputs smoother results. Solid white regions have

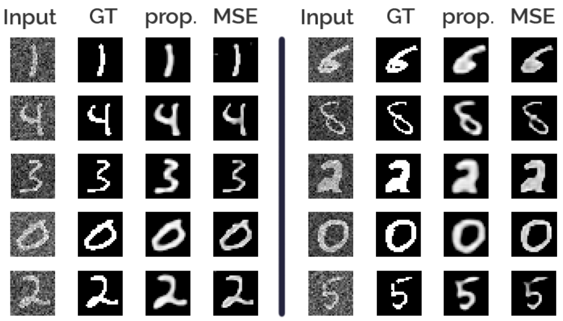

Fig. 6. Denoising neural network trained on MNIST dataset with two different objective functions. On the 3rd columns, outputs of the neural network trained with proposed metric is displayed, and on the 4th column, outputs of the same neural network trained with MSE is displayed.

a more uniform pixel intensity with the proposed objective function compared to the results of the network trained with MSE.

VII. Conclusion

Developing a perceptually plausible visual metric is a challenging task. Due to the complex nature of HVS, proposed metrics have mathematically complex and non-differentiable models. This prevents to use existing visual metrics to train neural networks with gradient-descent based optimization methods. In this paper, we examined the shortcomings of existing loss functions and discussed some of the proposed solutions to overcome these shortcomings. Then, we investigated several critical elements of existing visual system based models and presented our approach to implement them in a neural network scheme. We shared our preliminary results on training a denoising auto-encoder with the current state of the model as well as JND based dataset MCL-JCI. As future work, we plan to integrate a visual difference probability map into our visual model network and improve our training scheme of image restoration networks. Ultimately, we plan to achieve a perceptually plausible GAN-like training scheme. We also work on using the proposed metric as a loss function in different image restoration tasks, such as super-resolution, compression, etc.

Acknowledgment

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision)

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process., vol. 13, no. 4,

[2] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph., 30(4):40:140:14, July 2011.

[3] Scott Daly. 1993. The visible differences predictor: an algorithm for the assessment of image fidelity. In Digital images and human vision, Andrew B. Watson (Ed.). MIT Press, Cambridge, MA, USA 179-206.

[4] H. Zhao, O. Gallo, I. Frosio and J. Kautz, ”Loss Functions for Image Restoration With Neural Networks,” in IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47-57, March 2017.

[5] J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Vol. 2. MIT Press, Cambridge, MA, USA, 2672-2680.

[6] Johnson J, Alahi A, Li FF (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, pp 694711

[7] Ritschel, T., Ihrke, M., Frisvad, J. R., Coppens, J., Myszkowski, K., Seidel, H.-P. 2009. Temporal Glare: Real-Time Dynamic Simulation of the Scattering in the Human Eye. Computer Graphics Forum 28, 2, 183192.

[8] Whitaker, D., Steen, R., Elliott, D. 1993. Light scatter in the normal young, elderly, and cataractous eye demonstrates little wavelength dependency. Optometry and Vision Science 70, 11, 963968.

[9] Stockman, A., Sharpe, L. 2000. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. Vision Res 40, 13, 17111737.

[10] T. O. Aydin, R. K. Mantiuk, and H.-P. Seidel. Extending quality metrics to full luminance range images. Proceedings ofSPIE, 6806:68060B68060B10, 2008.

[11] Tariq, Taimoor Luis Gonzalez, Juan Kim, Munchurl. (2019). A HVS-inspired Attention Map to Improve CNN-based Perceptual Losses for Image Restoration.

[12] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[13] Wu, Sai, Mengdan Zhang, Gang Chen and Kan Chen. A New Approach to Compute CNNs for Extremely Large Images. CIKM (2017).

[14] Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 586-595.

[15] Mantiuk, R., Daly, S., Myszkowski, K., Seidel, H. 2005. Predicting visible differences in high dynamic range images: model and its calibration. In Proc. SPIE, vol. 5666, 204 214.

[16] Foley, J. 1994. Human luminance pattern-vision mechanisms: masking experiments require a new model. Journal of the Optical Society of America A 11, 6, 17101719.

[17] Watson, A. 1987. The cortex transform: Rapid computation of simulated neural images. Computer Vision, Graphics, and Image Processing 39, 3, 311327.

[18] Simoncelli, E., Freeman, W. 2002. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In Proceedings., International Conference on Image Processing, IEEE Comput. Soc. Press, vol. 3, 444447.

[19] Odena, et al., ”Deconvolution and Checkerboard Artifacts”, Distill, 2016. http://doi.org/10.23915/distill.00003

[20] Jin, Lina Lin, Joe Yu-chieh Hu, Sudeng Wang, Haiqiang Wang, Ping Katsavounidis, Ioannis Aaron, Anne Kuo, C.-C. Jay. (2016). Statistical Study on Perceived JPEG Image Quality via MCL-JCI Dataset Construction and Analysis. Electronic Imaging. 2016. 1-9. 10.2352/ISSN.2470-1173.2016.13.IQSP-222.