Visibility Metric for Visually Lossless Image Compression

Nanyang Ye[1], María Pérez-Ortiz[1,2], and Rafał K. Mantiuk[1]

[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom
Email: {yn272, rafal.mantiuk}@cl.cam.ac.uk

[2]Department of Computer Science, University College London, London, United Kingdom
Email: maria.perez@ucl.ac.uk

***Abstract*—Encoding images in a visually lossless manner helps to achieve the best trade-off between image compression performance and quality and so that compression artifacts are invisible to the majority of users. Visually lossless encoding can often be achieved by manually adjusting compression quality parameters of existing lossy compression methods, such as JPEG or WebP. But the required compression quality parameter can also be determined automatically using visibility metrics. However, creating an accurate visibility metric is challenging because of the complexity of the human visual system and the effort needed to collect the required data. In this paper, we investigate how to train an accurate visibility metric for visually lossless compression from a relatively small dataset. Our experiments show that prediction error can be reduced by 40% compared with the state-of-the-art, and that our proposed method can save between 25%-75% of storage space compared with the default quality parameter used in commercial software. We demonstrate how the visibility metric can be used for visually lossless image compression and for benchmarking image compression encoders.**

***Index Terms*—Visually lossless image compression, visibility metric, deep learning, transfer learning**

I. Introduction

The enormous amount of image and video data on the Internet poses a significant challenge for data transmission and storage. While significant effort has been invested in better image and video coding methods, improving these methods according to the human visual system's perception of visual quality.

Image and video compression methods can be categorized as lossy or lossless. While lossless methods retain original information up to the bit-precision of the digital representation, they are several times less efficient than lossy methods. Visually lossless methods lie in between lossy and lossless compression: they introduce compression distortions, but ensure that those distortions are unlikely to be visible. Visually lossless compression was first introduced to compress medical images and handle the increasing amount of data in clinics' picture archiving [1]. By selecting a fixed compression parameter or modifying the compression encoders, visually lossless compression has been shown effective for medical images in the gray-scale domain [2]. However, previous research on visually lossless compression is largely content and encoder dependent, which means we cannot apply medical compression methods to the compression of general image content using popular compression methods, such as JPEG. The goal of this paper is to devise a content and encoder independent visually lossless compression method for natural images.

While image quality metrics (IQMs) are meant to predict the perceived magnitude of impairment [3], visibility metrics predict the probability of noticing the impairment. IQMs are thus well suited for predicting strong, suprathreshold distortions. However, visibility metrics tend to be more accurate when predicting barely noticeable artifacts. Eckert *et al*. found that objective quality scores produced by IQMs (e.g. mean squared error), are correlated with compression quality setting [4]. However, simple IQMs have been found inaccurate [5].

The visually lossless threshold (VLT) is the encoder's parameter setting that produces the smallest image file while ensuring visually lossless quality. In this paper, we propose to train a visibility metric to determine the VLT. The proposed flow is shown in Figure 1. The original image is compressed at several quality levels by a lossy compression method, such as JPEG or WebP. Then, decoded and original images are compared by the visibility metric to determine the quality level at which the probability of detecting the difference (p_{det}) is below a predetermined threshold.

This paper extends our previous work on a general-purpose visibility metric [5] and focuses on visually lossless compression. The novel contributions are:

1) We present a new dataset of 50 images in which JPEG and WebP images are manually adjusted to be encoded at the VLT.
2) We significantly improve the predictive performance of the CNN-based visibility metric using pre-training.
3) We demonstrate the method utility in visually lossless compression and for comparing compression methods.

II. Quality vs. Visibility metrics

In this section we clarify the difference between full-reference IQMs and visibility metrics, as these are often confused. IQMs predict a single quality score (mean opinion score, MOS) for a pair of reference and test images [6]. Such a score should be correlated with mean opinion scores

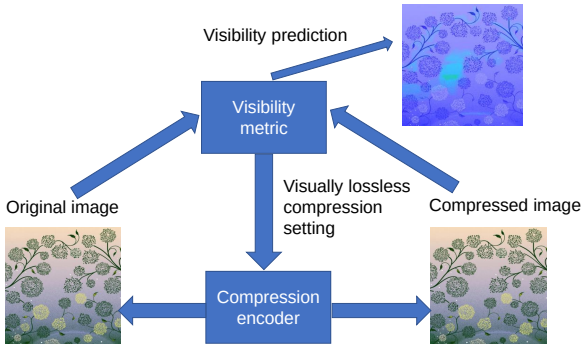

Fig. 1. Proposed flow of our method for visually lossless compression based on visibility metrics.

collected in subjective image quality assessment experiments. An extensive review of IQMs, including SSIM, VSI, FSIM, can be found in the survey [3].

However, it is challenging to estimate VLT accurately from traditional subjective assessment experiments, since mean opinion scores only capture overall image quality and do not explain where the distortions are visible in an image. This will be demonstrated with experiments in Section IV.

Visibility metrics predict the probability that the average observer [7] or a percentage of the population [5] will notice the difference between a pair of images. This probability is predicted locally, for pixels or image patches. The maximum value from such a probability map is computed to obtain a conservative estimate of the overall detection probability. Visibility metrics are either based on models of low-level human vision [7], [8] or rely on machine learning, trained using locally labeled images [5]. This paper improves on the CNN-based metric from [5] with the help of a newly collected dataset and a new pre-training approach.

III. Visually Lossless Image Compression Dataset

With the aim of validating visibility metrics, we collect a visually lossless image compression (VLIC) dataset, containing images encoded with JPEG and WebP[1] codecs.

A. Stimuli

The VLIC dataset consists of 50 reference scenes coming from previous image compression and image quality studies. The stimuli are taken from the Rawzor's free dataset (14 images)[2], CSIQ dataset (30 images) [9] and the subjective quality dataset in [10] (where we randomly selected 6 images from the 10 images in the dataset). For Rawzor's dataset, images were cropped to 960x600 pixels to fit on our screen. These images provide a variety of contents, including portraits, landscapes, daylight and night scenes. The images were selected to be a representative sample of Internet content. For JPEG compression, we use the standard JPEG codec (libjpeg[3]). For WebP compression, we use the WebP codec (libwebp[4]). Half of the reference scenes is compressed with JPEG and the other half with WebP, each into 50 different compression levels.

B. Experiment Procedure

The experimental task was to find the compression level at which observers cannot distinguish between the reference image and the compressed image.

1) Experiment stages: The experiment consisted of two stages. In the first stage, observers were presented with reference and compressed images side-by-side, and asked to adjust the compression level of the compressed image until they could not see the difference (method-of-adjustment). Compression levels found in the first stage were used as the initial guess for a more accurate 4-alternative-force-choice procedure (4AFC), used in the second stage. In this second stage, observers were shown 3 reference and 1 distorted images in random order and asked to select the distorted one. QUEST method [11] was used for adaptive sampling of compression levels and for fitting a psychometric function. Between 20 and 30 4AFC trials were collected per participant for each image. We decided to use side-by-side presentation rather than flicker technique from IEC 29170-2 standard [12] because the latter leads to overly conservative VLT as visual system is very sensitive to flicker. The side-by-side presentation is also closer to the scenario that users are likely to use in practice when assessing compression quality. On average, it took between 2 and 5 minutes for each observer to complete measurements for a single image.

2) Viewing Condition: The experiments were conducted in a dark room. Observers sat 90 cm from a 24 inch, 1920x1200 resolution NEC MultiSync PA241W display, which corresponds to the angular resolution of 60 pixels per visual degree. The viewing distance was controlled with a chinrest.

3) Observers: Observers were students and university staff with normal or corrected to normal vision. We collected data from 19 observers aged between 20 and 30 years. Around 10 measurements were collected for each image.

IV. Quality Metrics for Visually Lossless Image Compression

In this section, we demonstrate the weakness of a state-of-the-art image quality metric in visually lossless image compression. We use the Feature SIMilarity index metric (FSIM, [13]) and Butteraugli [14] to predict quality at different compression levels on the images from our VLIC dataset. For better clarity, we present results for 5 randomly select images. The results, shown in Figure 2, indicate that FSIM can reflect the changes in the visual quality but it is impossible to select a single value that would reliably predict the VLT (the dashed vertical lines). Butteraugli, which was specifically designed for finding VLT, gives slightly better prediction, but still not good enough. This demonstrates that the typical quality metrics are not accurate enough to predict VLT. We

[1] https://developers.google.com/speed/webp/
[2] http://imagecompression.info/test_images/
[3] https://github.com/LuaDist/libjpeg
[4] https://github.com/webmproject/libwebp

made a similar observation when testing other quality metrics, both full-reference and non-reference.

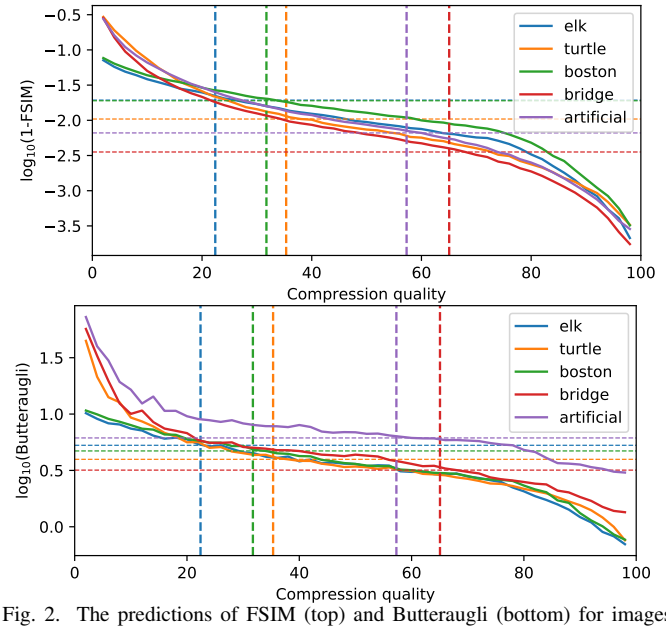

Fig. 2. The predictions of FSIM (top) and Butteraugli (bottom) for images compressed with increasing compression quality. The vertical dashed lines denote the visually lossless threshold (VLT) from the experiment with the same color denoting the same image. The predictions of existing metrics are not good indicators of VLT.

V. Deep neural network for visually lossless compression

In this section, we summarize the architecture from our previous work [5] and explore pre-training strategies to achieve an accurate metric even when only using a small dataset.

A. Architecture of the visibility metric

The DNN visibility metric's architecture is the same as in our previous work [5] (Figure 3), which is a convolutional-deconvolutional neural network. For VLT, we use the visibility threshold of 0.5, corresponding to 50% of the population not being able to tell the difference between the pristine and compressed images. Note that different thresholds could be used if more/less conservative VLT is needed.

B. Training and validation

We trained our network on the local visibility dataset (LocVis)[5], which consisted of test, reference images and maps with the probability of detection experimentally determined for image pixels. The network was trained using a probabilistic loss function, accounting for measurement noise, as explained in [5]. For training, we used a batch size of 48 with 50000 iterations and the Adam optimizer with a learning rate of 0.0001. We implemented our experiments in Tensorflow 1.8.

To find VLT for a particular image, the prediction of p_{det} was computed for 50 quality levels from 2-98. The prediction, shown as a blue line in in Figure 4 (for *big_building* image), often results in non-monotonic function. Instead of using binary search, we searched from high to low quality to find the quality level ($q1$) at which p_{det} raised above the predetermined threshold (0.5 for us). Then, we searched from low to high quality to find the quality level ($q2$) at which p_{det} dropped below the threshold. The mean of these two levels was taken as the metric's prediction of VLT for this image.

[5]LocVis dataset: https://doi.org/10.17863/CAM.21484

TABLE I
Pre-training cross-fold validation result (Results that do not have statistically significant differences are underlined)

RMSE of VLT Prediction		
No-pretraining	Butteraugli-pretraining	HDR-VDP-2-pretraining
24.82 ± 5.42	22.48 ± 3.35	**12.62 ± 0.56**

Kim *et al*. demonstrated that PSNR scores could be used to pre-train DNN quality metrics, which is later fine-tuned on a smaller, human-labeled dataset [15]. Inspired by this idea, we used existing visibility metrics, HDR-VDP-2 and Butteraugli, to generate the additional set of 3000 images with local visibility marking, which greatly increases the amount of training data. The images were taken from TID2013 dataset [16], which consisted of 25 scenes affected by 24 different distortions at several distortion levels. We first pre-trained visibility metric on the newly generated dataset and then fine-tuned the CNN weights on the LocVis dataset with accurate human markings.

We compare the performance of three training strategies: 1) without pre-training; 2) pre-trainng on images labelled with Butteraugli; 3) pre-training on images labelled with HDR-VDP-2. For that, we divide the LocVis dataset into 5 equal parts and train CNN visibility metric on the 4 parts using leave-one-out approach, which gives us 5 different trained versions of the metric. For each or those, we compute RMSE of VLTs against our VLIC dataset. The results, shown in Table I, are reported as the mean and standard deviation of the 5 trained versions for each training strategy. This validation procedure allows us to reduce random effects.

The validation results are shown in Table I. We compare the performance of the metric trained without pre-training, pre-trained using the dataset generated with HDR-VDP-2, and with Butteraugli. From Table I, we find that pre-training reduces both the mean and standard deviation of RMSE. This suggests that pre-training enhances the generalization ability of the neural network. We observe a much larger improvement for HDR-VDP-2 pre-training. The statistical significance of the difference is confirmed by a two-sample t-test, illustrated as an underline in Table I.

C. Comparison with other methods

In this section, we compare the VLT predictions of our method with other visibility metrics [5]. We use the best performing version of our metric: with HDR-VDP-2 pre-training, followed by fine-tuning on the LocVis dataset. The results are shown in Table II. We test all the metrics on the newly collected VLIC dataset (Section III), which was not used in training. The proposed method reduces RMSE by 40%

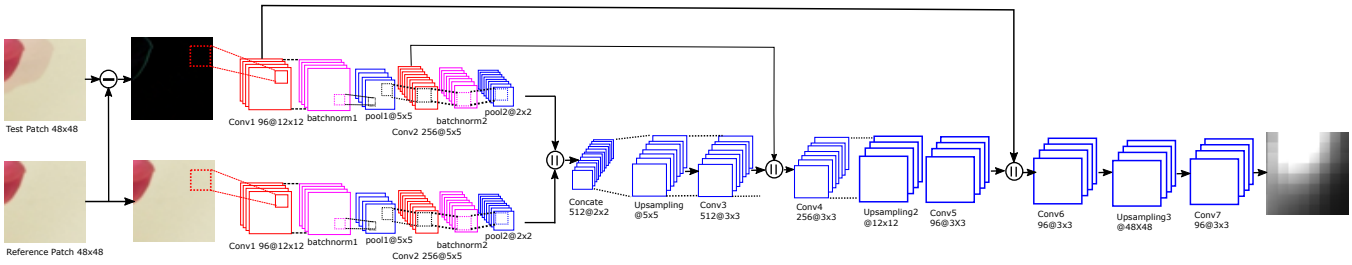

Fig. 3. The CNN architecture. We use the non-Siamese fully-convolutional neural network to predict visibility maps.

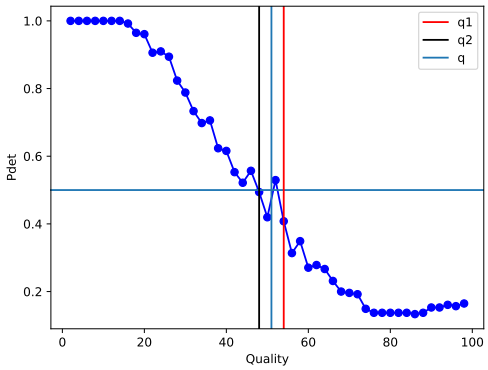

Fig. 4. Illustration of the procedure used to determine the visually lossless threshold using the visibility metric.

TABLE II
EXPERIMENT RESULTS ON THE VISUALLY LOSSLESS IMAGE COMPRESSION DATASET

RMSE of VLT Prediction			
Proposed	CNN [5]	Butteraugli [14]	HDR-VDP-2 [7]
12.38	20.33	20.91	43.33

compared with the best-performing CNN-based metric from our previous work.

To illustrate visually lossless compression, we show randomly selected examples of pristine and visually lossless compressed images in Figure 5. The compressed images appear identical to the reference since they are compressed at the VLT level, but provide substantial saving as compared to the default encoder setting of 90.

Fig. 5. The pairs of reference and compressed images in which the compression quality was adjusted using the proposed metric to be at the visually lossless level. The values in parenthesis denote saving as compared to JPEG and WebP with the fixed quality of 90. (Better viewed in zoom-in mode)

VI. APPLICATIONS

We demonstrate the utility of our metric in two applications: visually lossless compression and benchmarking of lossy image compression. For this, we randomly selected 1000 high quality images from the Flicker8K dataset[6], which were encoded with JPEG quality of 96.

A. Visually lossless image compression

For visually lossless compression, we used the procedure from Section V-B to find the VLT with the probability of detection 0.25. This ensured that only 25% of the population had a chance of spotting the difference between the compressed and original images. The threshold was found separately for JPEG or WebP. Then, we computed the decrease in file size between our visually lossless coding and JPEG or WebP, both set to quality 90. We choose the quality of 90 as many applications often use it as a default setting. We plot the histogram of per-image file size saving in Figure 6. The figure shows that our metric can save between 25% to 75% of file size for most images in the dataset for both compression methods. Note that the negative number in the plots indicate that some images need to be compressed with higher quality than quality 90.

B. Benchmarking lossy image compression encoders

To demonstrate that our metric can be utilized for benchmarking lossy image compression methods, we encoded im-

[6]http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html

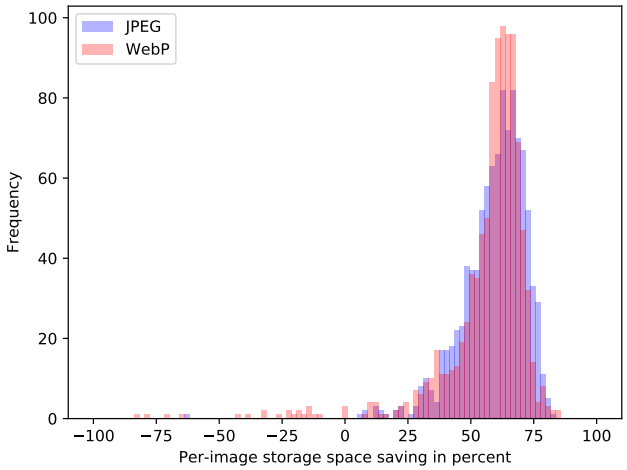

Fig. 6. Histogram of per-image storage saving as compared to quality 90 setting.

ages from Flicker8K dataset at different bit-rates with JPEG and WebP. Then, for each decoded images, we used our metric to predict the probability of detection and plot the results in Figure 7. The figure shows that WebP can encode low bit-rate images with less noticeable artifacts than JPEG. However, the advantage of WebP is lost at higher bit-rates. Compared to quality metrics, our visibility metric can express the difference in terms of probability of detecting artifacts, rather than in terms of an arbitrary scaled quality value.

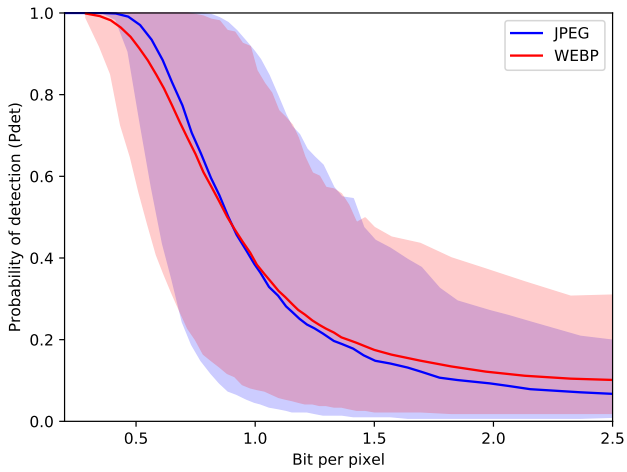

Fig. 7. Relationship between probability of detection and file size. We use bits-per-pixel because of different size of images in the dataset. The shaded region marks the 95% confidence interval.

VII. Conclusions

In this paper, we proposed a visibility metric trained for visually lossless image compression and showed the benefits of pretraining such network with previously proposed image quality metrics. We have shown that when combined with lossy image compression methods, we can save significant amount of storage and transmission space. We also shown that the metric can be applied to benchmarking of image compression methods.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725253–EyeCode).

References

[1] D. Wu, D. M. Tan, and H. R. Wu, "Visually lossless adaptive compression of medical images," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, vol. 1, Dec 2003, pp. 458–463 Vol.1.

[2] H. S. Woo, K. J. Kim, T. J. Kim, S. Hahn, B. Kim, Y. H. Kim, and K. H. Lee, "JPEG 2000 compression of abdominal CT: Difference in tolerance between thin- and thick-section images," *American Journal of Roentgenology*, vol. 189, no. 3, pp. 535–541, Sep 2007. [Online]. Available: https://doi.org/10.2214/AJR.07.2304

[3] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297 – 312, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320311000204

[4] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.*, vol. 70, no. 3, pp. 177–200, Nov. 1998. [Online]. Available: http://dx.doi.org/10.1016/S0165-1684(98)00124-8

[5] K. Wolski, D. Giunchi, N. Ye, P. Didyk, K. Myszkowski, R. Mantiuk, H.-P. Seidel, A. Steed, and R. K. Mantiuk, "Dataset and metrics for predicting local visible differences," *ACM Transactions on Graphics*, in press.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[7] 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.*, vol. 30, no. 4, pp. 40:1–40:14, Jul. 2011. [Online]. Available: http://doi.acm.org/10.1145/2010324.1964935

[8] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. MIT Press, 1993, vol. 1666, no. 1992, pp. 179–206.

[9] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy." *J. Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/jei/jei19.html#LarsonC10

[10] C. Strauss, F. Pasteau, F. Autrusseau, M. Babel, L. Bedat, and O. Deforges, "Subjective and objective quality evaluation of LAR coded art images," in *2009 IEEE International Conference on Multimedia and Expo*, June 2009, pp. 674–677.

[11] A. B. Watson and D. G. Pelli, "Quest: A bayesian adaptive psychometric method," *Perception & Psychophysics*, vol. 33, no. 2, pp. 113–120, Mar 1983. [Online]. Available: https://doi.org/10.3758/BF03202828

[12] ISO/IEC, "ISO/IEC 29170-2:2015 Information technology – Advanced image coding and evaluation – Part 2: Evaluation procedure for nearly lossless coding," 1995. [Online]. Available: https://www.iso.org/standard/66094.html

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *Image Processing, IEEE Transactions on*, vol. 20, pp. 2378 – 2386, 09 2011.

[14] J. Alakuijala, R. Obryk, O. Stoliarchuk, Z. Szabadka, L. Vandevenne, and J. Wassenberg, "Guetzli: Perceptually guided JPEG encoder," *CoRR*, vol. abs/1703.04421, 2017. [Online]. Available: http://arxiv.org/abs/1703.04421

[15] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, Nov 2017.

[16] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Image database TID2013," *Image Commun.*, vol. 30, no. C, pp. 57–77, Jan. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.image.2014.10.009