Predicting visible image differences under varying display brightness and viewing distance

Nanyang Ye
Department of Computer Science and Technology
University of Cambridge, UK
yn272@cam.ac.uk

Krzysztof Wolski
MPI Informatik
Saarbrücken, Germany
kwolski@mpi-inf.mpg.de

Rafał K.Mantiuk
Department of Computer Science and Technology
University of Cambridge, UK
rkm38@cam.ac.uk

Abstract

Numerous applications require a robust metric that can predict whether image differences are visible or not. However, the accuracy of existing white-box visibility metrics, such as HDR-VDP, is often not good enough. CNN-based black-box visibility metrics have proven to be more accurate, but they cannot account for differences in viewing conditions, such as display brightness and viewing distance. In this paper, we propose a CNN-based visibility metric, which maintains the accuracy of deep network solutions and accounts for viewing conditions. To achieve this, we extend the existing dataset of locally visible differences (LocVis) with a new set of measurements, collected considering aforementioned viewing conditions. Then, we develop a hybrid model that combines white-box processing stages for modeling the effects of luminance masking and contrast sensitivity, with a black-box deep neural network. We demonstrate that the novel hybrid model can handle the change of viewing conditions correctly and outperforms state-of-the-art metrics.

1. Introduction

A number of applications in computer vision, computer graphics and image processing can benefit from the knowledge whether introduced changes in images are visible to the human eye, or not. For example, we could use such a metric to determine the maximum image compression level for visually lossless compression, the best resolution or compression method for textures used in computer graphics rendering, or to evaluate image reconstruction methods.

Different from (full reference) image quality or similarity metrics, which predict a single value that represents an overall image quality, visibility metrics predict a visibility map which provides local information about probability of perceiving the difference between a pair of images. Visibility metrics tend to offer higher accuracy for near-threshold distortions, which are crucial for the applications in which no visible artifacts can be tolerated. Visibility metrics can also predict the location of visible artifacts in images. In contrast, image quality metrics are better at estimating the distortion magnitude for supra-threshold distortions.

Most existing image visibility metrics, such as Sarnoff Visual Discrimination model (VDM) [16], Visual Difference Predictor (VDP) [8], and High Dynamic Range VDP (HDR-VDP) [18] are white-box models, which are designed to model the low-level perception mechanisms of human visual system (HVS). Because of their white-box nature, these models can generalize well to new conditions, such as different viewing distances or absolute luminance levels. However, because of the limited number of trainable parameters and their complexity, they cannot be trained to fit complex multi-modal data distributions as effectively as black-box machine learning-based models. The work done in [24] demonstrated that CNN-based visibility predictor achieves higher performance than the existing white-box metrics. However, this deep learning solution was trained for and could predict visibility only for a fixed viewing condition: a display with the peak luminance of 110 cd/m^2 and the angular resolution of 40 pixels per visual degree (ppd).

In this work, we extend the work of Wolski *et al*. [24] so that the proposed visibility metric can account for a range of display brightness levels and angular resolutions. We achieve this by combining white-box models of luminance masking and spatial resampling with a black-box CNN-

based model, based on the architecture from [24].

To obtain sufficient data for training under different absolute luminance levels and viewing distances, we use HDR-VDP[18], an existing white-box visibility metric, improved by retraining from [24], to generate predictions for a large number of images affected by JPEG and WebP image compression. Then we use a human-labeled dataset to fine-tune the metric and validate the results. The human-labeled dataset consists of both existing local visibility dataset (LocVis [1]) and a newly-collected dataset of 264 images labeled under different viewing conditions. The combined dataset (LocVisVC) is available online[2]. The code of the visibility metric can be found in the GitHub repository[3].

The main contributions of our paper are:

1. We provide a local visibility dataset that is measured under varying luminance and viewing distance conditions.

2. We propose a hybrid visibility metric that combines white-box perceptual processing with a black-box neural network to account for absolute luminance and viewing distance and achieves the best performance.

3. We find an efficient method to train the metric with a limited dataset, in which we take advantage of existing white-box metrics to label a large dataset used for pre-training.

2. Related work

In this section, we provide background information on modeling of basic HVS characteristics that are important for image perception on displays with variable brightness and observer distance. We also survey existing image quality and visibility metrics that attempt to model such HVS characteristics. Finally, we discuss relatively rare attempts to employ modern machine learning in visibility metrics.

2.1. Contrast perception models

The visual sensitivity of HVS varies as a function of a number of factors such as luminance, contrast, spatial frequency, luminance adaptation, color, and spatial image content. In this work, we explicitly model the first two factors in a white-box fashion, which we briefly discuss in this section. We expect that color perception and more advanced concepts of spatial vision, such as visual masking, can be more efficiently learned by the network in a black-box manner without any domain specific knowledge. Our strategy is to explicitly model the easy-to-capture HVS characteristics, while leaving the capture of more involved effects to machine learning.

[1] https://doi.org/10.17863/CAM.21484

[2] https://doi.org/10.17863/CAM.37996

[3] https://github.com/ynyCL/DPVM

Luminance masking The human eye is not equally sensitive to all luminance levels. In dark conditions, much smaller luminance differences can be distinguished than in bright conditions, but the sensitivity to contrast ($\Delta L/L$) also gets worse at low light. This effect is often called luminance masking or luminance self-masking [25]. In terms of distortion perception, this means that the same magnitude of distortion can be differently perceived as a function surrounding luminance. To make the luminance values more perceptually uniform, luminance can be transformed into the logarithmic domain. However, the logarithmic transform, known as Fechner's law, does not model precisely the HVS sensitivity to light changes [17]. Typically the threshold vs. intensity (t.v.i.) or contrast sensitivity function is used to determine the smallest noticeable difference in luminance across the luminance range, and build a function that maps physical luminance values into approximately perceptually uniform units [4, 19, 20]. Overall, luminance masking is well understood and easy to model [9, 19, 18, 1, 20], so we include it explicitly into our visibility metric.

Contrast Sensitivity Function Perceived contrast depends not only on its magnitude but also on the spatial frequency of a contrast pattern. Contrast Sensitivity Function (CSF) [2] specifies the detection threshold for a stimulus as a function of its spatial frequencies that effectively are projected on the retina and increase proportionally to the observation distance. Due to the inverted "U" shape of the CSF, image elements represented by low (high) spatial frequencies might become visible (invisible) with the increase of the observation distance. The concept of hybrid images employed in arts and media [22], where the observer sees completely different content as a function of his or her distance to the image, is a dramatic demonstration of immense CSF impact on visual perception. This has strong consequences in image distortion perception as well, where the visibility of distortions varies as a function of the observation distance in an easy to model way [8, 19, 18]. As an additional factor CSF changes as a function of luminance adaptation, which means that artifacts visibility in darker image regions might be further reduced [8, 19, 13].

2.2. Image metrics

Image quality metrics Quality metrics are intended to estimate the magnitude of image distortion as a single mean opinion score value. We recommend the readers more complete surveys on quality metrics [15, 7], and this section we discuss only sparse metric examples that attempt to model the display brightness and observer distance. High Dynamic Range Video Quality Measure (HDR-VQM) [21] is proposed to address the change of physical luminance in the images. HDR-VQM employs the perceptual uniform transformation [1] to convert the physical luminance to the

perceptual uniform values and use log-Gabor filters [10] to compute the subband difference. However, HDR-VQM does not account for the observer distance.

Visibility metrics Visibility metrics are intended to predict the probability that a human observer detects a difference at a particular image location. Due to the limited size of training datasets, the majority of existing visibility metrics are built using models of the HVS, which restrict the number of tunable parameters that need to be trained. Early visibility metrics model human's sensitivity to spatial contrast, and typically properly account for changes in the display brightness and observer distance. The sCIELab metric uses a spatio-chromatic CSF to filter the CIELab encoded pixels for computing the visibility map [27]. A power function that is used in such luminance encoding models luminance masking. More complex examples of visibility metrics include VDM [16], VDP [8], and HDR-VDP [18] that apart from luminance masking and CSF modeling, also account for visual contrast masking. However, the threshold elevation or transducer models [23] used for this purpose appear to be overly simplified for complex images, which often leads to inaccuracies in the visibility prediction [6]. Recent machine learning models are more flexible in modeling these important HVS characteristics [24], and in particular, their interactions as a function of complex image content.

CNN-based visibility metrics The effectiveness of Convolutional Neural Network (CNN)-based methods has widely been demonstrated in quality evaluation [5, 11, 3, 26] and more sparsely in visibility prediction [24]. The latter visibility metric uses a convolutional-deconvolutional architecture and its prediction correlates well with human experiment results. However, this CNN-based visibility metric cannot deal with the change of luminance or distance, which largely prohibits the practical use of this metric as many barely-noticeable distortions can change their visibility significantly with the change of luminance or distance as we show in the next section.

3. Data collection

The aim of the experiment was to collect data on distortion visibility under different viewing conditions: varying display peak luminance and viewing distance.

Stimuli We randomly selected 66 images that from the LocVis dataset [24]. The selected scenes covered many types of distortions, such as compression, synthetic perception patterns and artifacts from image-based rendering methods. All the selected scenes had up to 3 levels of distortion, for example three different amplitudes of noise or different JPEG compression levels).

Experimental Procedure The visibility of image differences can be measured with different experiment setups, such as a side-by-side presentation, flickering between distorted and reference images, and no-reference presentation [6]. As the HVS is very sensitive to temporal changes, the flicker mode results in overly conservative estimates. Therefore, we selected the side-by-side presentation, which avoids this problem, and is also more relevant for many applications.

Observers were asked to paint freely all the visible distortions using a custom painting interface. To speed up the process and to increase the coherency of collected data multiple levels of distortion magnitude proposed in [24] were used.

Display and viewing conditions The experiment took place in a room with dimmed lights. The display was positioned to minimize screen reflections. The images were shown on a $23''$, 1920×1200 pixels resolution Acer GD235HZ display set the the sRGB color profile. The screen was calibrated using a Minolta LS100 luminance meter to two different peak luminance conditions: 10 cd/m^2 and 220 cd/m^2. To achieve the luminance of 10 cd/m^2, the display was dimmed and a 0.6 Neutral Density (ND) filter, reducing the light by a factor of 4, was put on the screen. These two setups cover the luminance range found in most of the displays [4]. The observers viewed the display at two distances, 40 cm and 86 cm, which correspond to angular resolutions of 30 and 60 pixels per visual degree.

Observers In total, 46 observers, aged between 23 and 29 years old, were recruited among computer science and other field students. All observers were paid for their participation and had normal or corrected-to-normal vision. They were naïve about the purpose of the experiment. To reduce the effect of fatigue, the experiment was split into several sessions, where each session lasted less than one hour.

Results Figure 1 illustrates the trends of visibility changes under different luminance and distance conditions. With the increase of luminance and the decrease of ppd (decreasing ppd is equivalent to decreasing distance), distortions become more visible, which agrees with empirical observations and previous research [18]. This also confirms the need for a visibility metric that accounts for both absolute luminance and a viewing distance.

[4] https://www.laptopmag.com/benchmarks/display-brightness

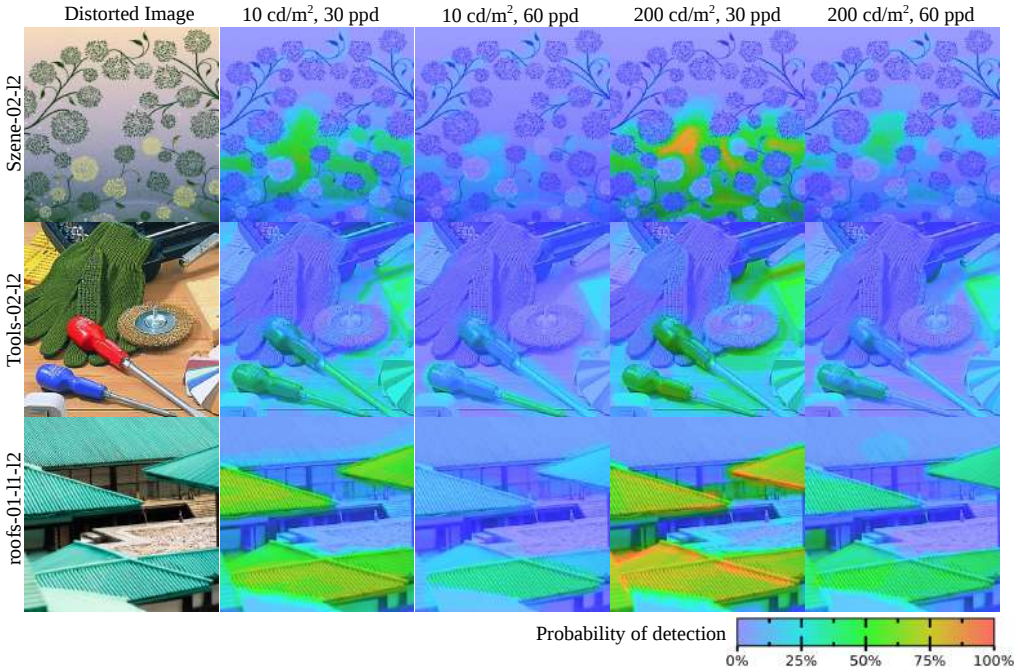

Figure 1. Examples of images and subjective data from LocVisVC dataset. Decreasing ppd (decreasing distance between the observer and the display) for the same luminance condition increases the visibility of artifacts. When luminance is increased keeping the same ppd condition the visibility of artifacts also increases.

Pre-training dataset As the manually labeled datasets were insufficient for training, we also prepared a dataset with synthetic labels, generated with the HDR-VDP visibility metric. We used 200 high-quality photographs obtained directly from camera RAW files. All photographs were resized to the maximum resolution of 1920×1080. The images were then distorted by encoding and decoding using JPEG[5] and WebP[6] image compression at the quality settings of 20, 50 and 90. We then randomly selected 50 images as the base scenes for our dataset. Each of these images was converted into linear colorimetric units using the display model (explained in Section 4.1) assuming the peak luminance of 10 cd/m^2, 110 cd/m^2 and 220 cd/m^2. The visibility map for these images was then predicted for the angular resolutions of 30, 40, 50 and 60 pixels per visual degree, producing in total 600 labeled images. A summary of the dataset can be found in Table 1.

4. Metric Architecture

Most NN-based metrics rely on existing architectures, which are trained in an end-to-end manner. In our case, both the viewing distance and the display peak brightness are significant factors that affect predictions. Both parameters could be fed to the network in a standard manner, hoping that the network will learn the correct relationships. However, such a solution requires a large quantity of subjective data, which cannot be easily collected for our task in a reasonable time. To address this challenge, we design a hybrid architecture, in which the viewing distance and the display peak luminance are modeled explicitly as a pre-processing stage of the CNN-based metric. The architecture of the proposed metric and the data pre-processing are illustrated in Figure 3 and described in the following sections.

4.1. Display model

Since modern displays differ substantially in their peak brightness, it is important to model how much light their emit. As an example, some mobile displays can reach the peak luminance of 900 cd/m^2 and can be dimmed to as low light levels as 3 cd/m^2. The visibility of image distortions is very different between both cases. To model the amount of the emitted light, we use the standard gain-gamma-offset display model:

$$L = (L_{\text{peak}} - L_{\text{black}}) \left(\frac{I}{255} \right)^{2.2} + L_{\text{black}}, \qquad (1)$$

where I is the input pixel value, L_{peak} is the peak luminance of the display, and L_{black} is the luminance of black level (light emitted from pixels set to black). Each image provided to the metric is first transformed from pixel values to colorimetric red, green and blue values using the display model from the equation above.

[5]https://github.com/LuaDist/libjpeg
[6]https://developers.google.com/speed/webp

Subset name	Scenes	Images	Distortion levels	Level generation method	Peak luminance	ppd
MIXED	20	59	2-3	blending	110 cd/m^2	40
PERCEPTIONPATTERNS	12	34	1,3	blending	110 cd/m^2	40
ALIASING	14	22	1-3	varying sample number	110 cd/m^2	40
PETERPANNING	10	10	1	n/a	110 cd/m^2	40
SHADOWACNE	9	9	1	n/a	110 cd/m^2	40
DOWNSAMPLING	9	27	3	varying shadow map resolution	110 cd/m^2	40
ZFIGHTING	10	10	1	n/a	110 cd/m^2	40
COMPRESSION	25	71	2-3	varying bit-rates	110 cd/m^2	60
DEGHOSTING	12	12	1	n/a	100 cd/m^2	60
IBR	18	36	1,3	varying key frame distances	110 cd/m^2	40
CGIBR	6	6	1	n/a	110 cd/m^2	40
TID2013	25	261	n/a	n/a	100 cd/m^2	40
VIEWCOND	26	**264**	1-3	n/a	10, 200 cd/m^2	30, 60
PRETRAIN	200	**600**	3	JPEG and WebP compression	10, 110, 200 cd/m^2	30,40,50,60

Table 1. The subsets of the dataset used for training. VIEWCOND is the newly measured LocVisViewCond dataset. PRETRAIN is the HDR-VDP generated synthetic dataset for pre-training. The other sets are from the original LocVis dataset.

4.2. Viewing distance

An intuitive way to account for the viewing distance is to provide to the model an image with the fixed angular resolution. As the contrast sensitivity of visual system is mostly dependent on the spatial frequency content in cycles per visual degree (cpd), the constant angular resolution ensures that spatial frequencies remain the same regardless of the viewing distance. The angular resolution of an image can be computed as:

$$r = \frac{N_x}{h_{deg}} \quad [ppd] \,, \tag{2}$$

where N_y is the display vertical resolution expressed in pixels and the display height in visual degrees is given by:

$$h_{deg} = 2 \arctan\left(\frac{h_{mm}}{2\, d_{mm}}\right) \,, \tag{3}$$

where d_{mm} is the viewing distance expressed in millimeters. The display height expressed in millimeters can be found from:

$$h_{mm} = \sqrt{\frac{(25.4\, s_{diag})^2}{1 + \left(\frac{N_x}{N_y}\right)^2}} \,, \tag{4}$$

where s_{diag} is the display diagonal length expressed in inches. Once we know the angular resolution of the input image, we resample it so that it has the angular resolution of 60 ppd. 60 ppd is the highest resolution in our dataset and also a reasonable limit for most visual task, since the sensitivity of visual system drops rapidly below 30 cpd [2].

Since resampling alone cannot account for all frequency-dependent effects, such as the shift of peak sensitivity with luminance, we also introduce the *ppd* parameter to the latent code. This is achieved by concatenating a slice with replicated *ppd* values to the feature maps generated by the encoders (see Figure 3).

4.3. Luminance masking

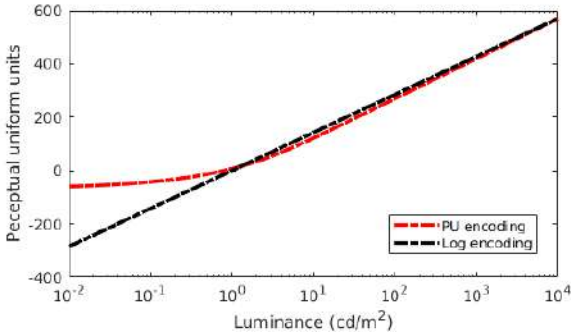

Figure 2. PU and logarithmic transform functions, for converting absolute light levels into approximately perceptually uniform values, which could be input to a CNN.

Since differences are less visible at lower absolute luminance levels, we need to account for this drop of visual system sensitivity. Luminance masking can be modeled by a transfer function derived from the contrast sensitivity function of visual system [18, 1]. The transfer function we use is also known as Perceptually Uniform (PU) encoding [1], as it transforms physical luminance into approximately perceptually uniform units. The PU encoding is defined as an integral of inverse of detection thresholds:

$$P(L) = \int_{L_{min}}^{L} \frac{1}{T(l)} dl \tag{5}$$

where L_{min} is the minimum luminance to be encoded. The detection thresholds $T(L)$ are modeled as a function of absolute luminance L:

$$T(L) = S \cdot \left(\left(\frac{C_1}{L}\right)^{C_2} + 1\right)^{C_3} \tag{6}$$

Where S is the absolute sensitivity constant, L is the luminance, and C_1, C_2, C_3 are parameters obtained by fitting to contrast sensitivity measurements. We use the parameters from [18].

For comparison, we also experiment with the logarithmic encoding of luminance, as it is the first-order approximation of the visual system response, which accounts for the Fechner law. We show both perceptual encoding functions in Figure 2.

4.4. CNN architecture

The CNN architecture of the proposed metric is based on the one proposed in [24]. Although image metrics are often modeled using Siamese architectures [11], the CNN we employ has two independent branches, which encode different information: the first branch encodes the difference between test and reference images (after pre-processing steps) and the second branch encodes the reference image. Such independent branches, shown in Figure 3, are used to improve detection of small image differences. In contrast to CNN architectures used for classification or detection tasks, which need to be robust to noise, our model needs to be particularly sensitive to small variations in input.

Each branch of the encoder uses two convolutional layers of the AlexNet [14]. Two branches and the *ppd* value are concatenated together, as explained in Section 4.2. The patch with predicted probability of detection map is generated by two deconvolution layers. More formally, we denote the perceptually encoded color images of difference and reference patches as D and R, respectively. We also define mapping functions F_{wconv^d} and F_{wconv^r} to represent the convolutional operations for two branches, in which $wconv^d$ and $wconv^r$ are weights for the difference and reference encoding branches, respectively. We also denote the w_{dec} as the weights for deconvolutional operations with skip connections. Our metric can then be expressed as:

$$P_w(D,R) = F_{w_{dec}}\left(F_{w_{conv^d}}(D) \textcircled{\scriptsize ||} F_{w_{conv^r}}(R) \textcircled{\scriptsize ||} r\right), \quad (7)$$

where $\textcircled{\scriptsize ||}$ represents the concatenation operation of the output of the difference branch, reference branch, and the slice with the replicated *ppd* values r.

To predict a visibility map for an image of arbitrary size, we slice the image into 48×48 pixel patches with 42-pixel overlap, infer visibility for each patch and compute the final visibility map by averaging the predictions from the overlapping patches. Predicting a visibility map usually takes 2-4 seconds for 1920×1080 image using NVidia GTX 1080Ti GPU.

5. Training

For training the new Deep Photometric Visibility Metric (DPVM), we use the probabilistic loss function from [24], as it provides a principled way of modeling the experimental data. The probabilistic loss function models the marking task as a stochastic process accounting for the mistakes, lack of attention and limited number of observations. This allows us to capture the uncertainty in the human-labeled dataset. After the pre-processing steps, we split images into 48×48 pixel non-overlapping patches. We remove the patches where there is no difference between their distorted and reference versions. We implement the CNN in Tensorflow 1.10.1 [7]. We use The adaptive momentum optimizer (Adam) with a learning rate $1e^{-5}$ and a batch size of 48 is used for optimization.

We split the training process into two stages.

Stage 1: Pre-training with HDR-VDP As the collected dataset contains only limited variation in viewing distance and display peak luminance levels, we supplement our training with over 13 million patches that have been automatically labeled by a white-box visibility metric — HDR-VDP. The generation of this PRETRAIN dataset was explained in Section 3. The idea is inspired by the work of Kim *et al*., who demonstrated that PSNR scores can be used to pre-train CNN-based quality metrics [12]. Similarly, we run 20000 iterations of training on the PRETRAIN dataset, which is followed by fine-tuning in Stage 2. Although the labels generated by HDR-VDP can be inaccurate, they capture general relationship between input and output patches and therefore prime the CNN to capture the relationships, which could be missing in manually labeled data.

Stage 2: Fine-tuning At this stage, we initialize the neural network with weights from the first stage and use the manually labeled datasets for training.

6. Results

To validate prediction performance, we randomly split the LocVisVC dataset into 5 folds, ensuring that each scene is in a single fold, and run a 5-fold cross-validation. We report the mean and standard error of the likelihood used for the loss function (the higher likelihood indicated the higher accuracy).

PU vs. logarithmic encoding First, we compare performance when either PU encoding or a logarithmic function is used to account for luminance masking. The likelihood for the PU encoding (0.877 ± 0.015) was substantially higher than for logarithmic function (0.705 ± 0.02). This

[7] https://www.tensorflow.org

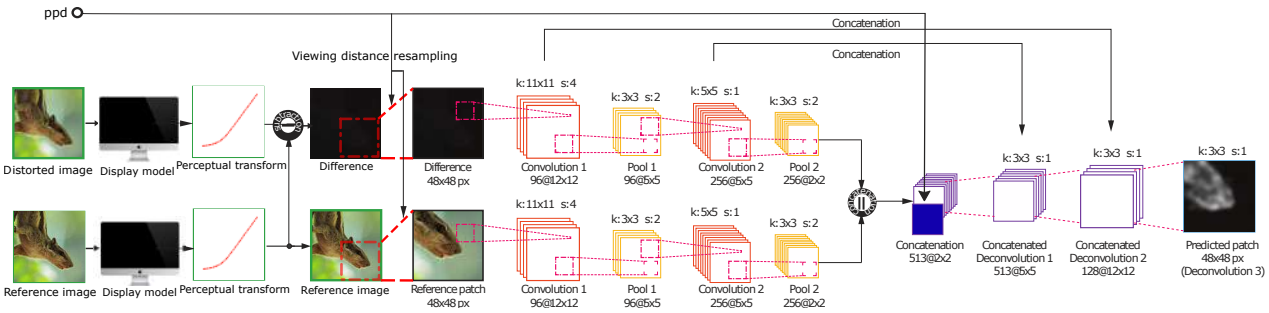

Figure 3. CNN visibility metric architecture.

suggests that luminance masking is a significant effect in our dataset, which cannot be easily learned by the black-box CNN. Given sufficient data, we could expect similar performance for both luminance encodings. This results demonstrates that when the data is limited, the combination white-box preprocessing with black box learning is more efficient strategy.

HDR-VDP pre-training Next, we investigate the effect of pre-training on the metric performance. We run pre-training for the number of iterations ranging from 10,000 to 50,000, followed by fine-tuning of 50,000 steps, and report the results in Figure 4. The figure shows that pre-training always resulted in higher accuracy, but the performance dropped after about 20,000 iterations. This shows that the amount of pre-training needs to be carefully controlled to retain the ability of the network to effectively learn from the human-labelled data. In the following experiments, we use 20,000 iteration for pre-training.

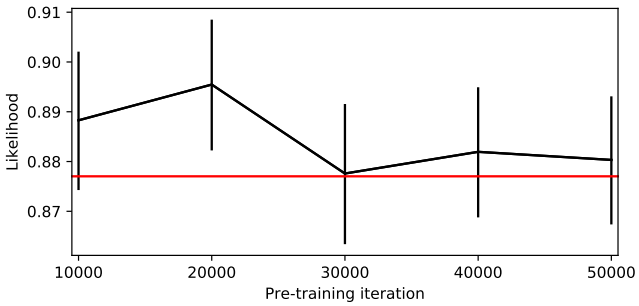

Figure 4. The effect of pre-training iterations on the performance. The red line denotes the result without pre-training. The error bars denote standard errors. The higher likelihood, the better is accuracy.

Metric comparison Finally, we compare the proposed DPVM metric to the HDR-VDP, which is the only visibility metric that can account for the viewing conditions. For fair comparison, we retrain HDR-VDP-2.2 on the same dataset as used for the training the CNN-based metric. The result of cross-validation is shown individually for each subset in Figure 5. The likelihood of the proposed DPVM metric is significantly higher for each subset, demonstrating the CNN-based metric can be trained with higher accuracy.

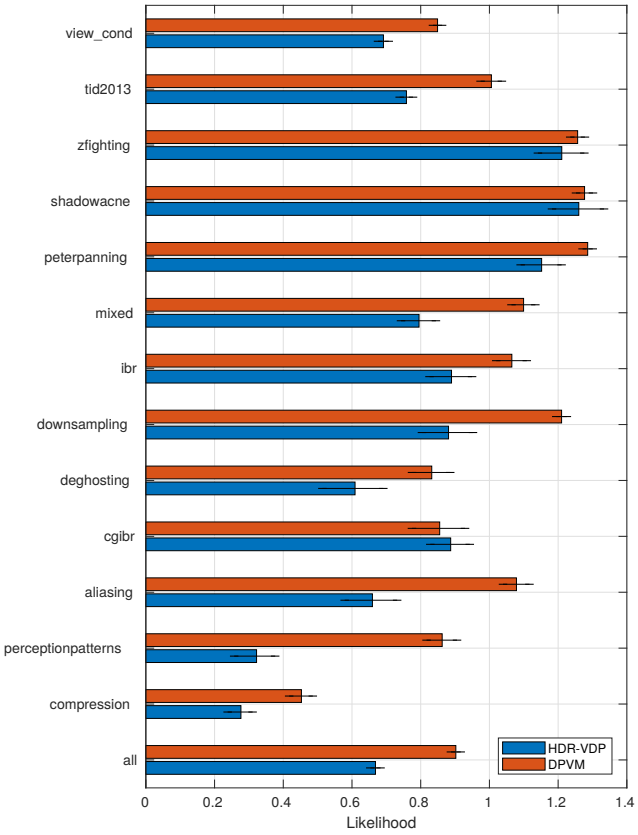

Figure 5. Metric cross-validation results for each subset and for the entire dataset.

An examples of metric predictions and user markings are shown in Figure 6. We can observe there that similar to HDR-VDP, DPVM can account for the change of viewing distance and absolute luminance as shown in rows 1–3. Pre-training with HDR-VDP also helps improve the generalization performance for most cases.

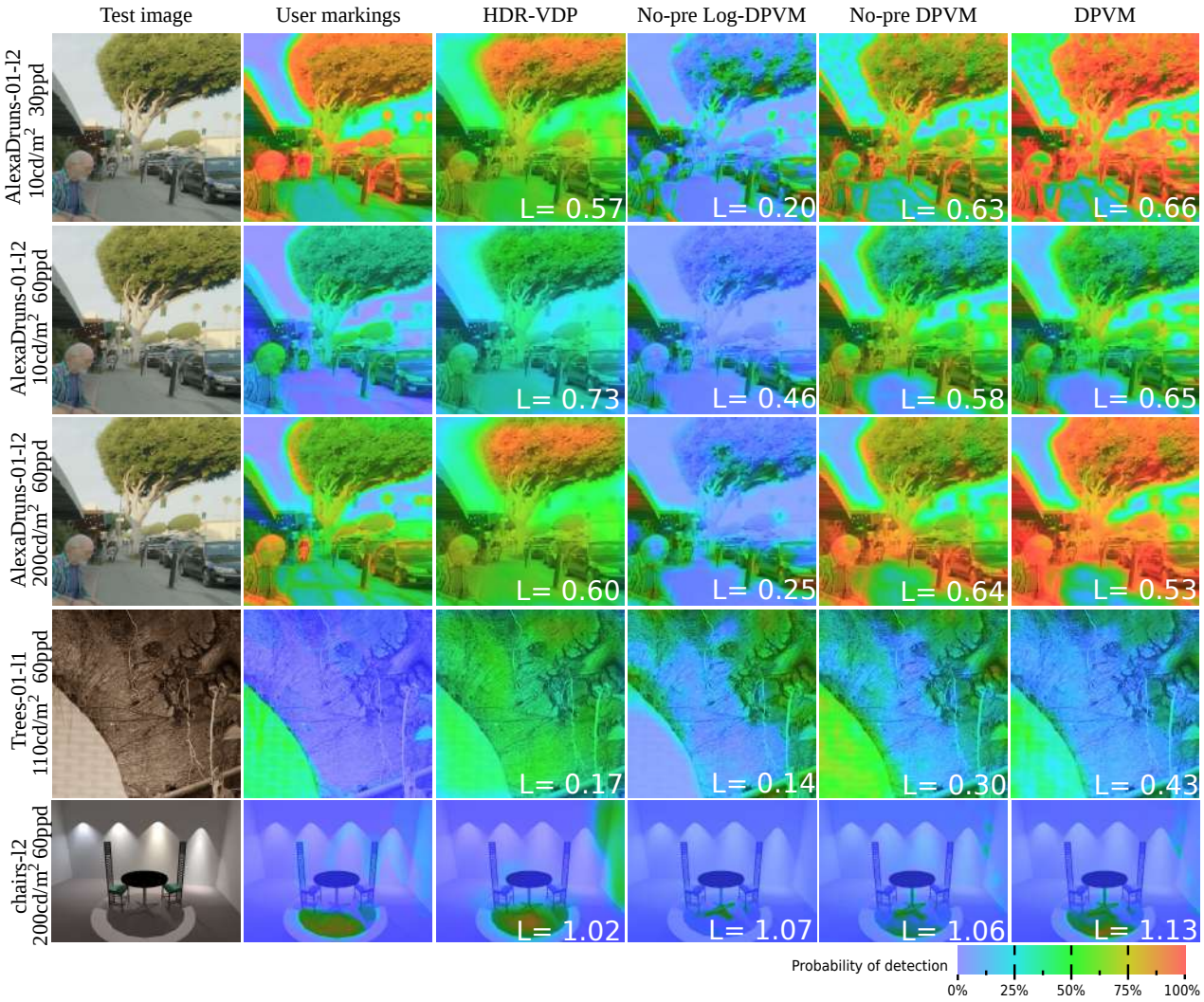

Figure 6. Distorted images, users' markings and metrics' predictions examples from the dataset. L is the likelihood, the higher the better. No-pre prefix means without HDR-VDP pre-training.

7. Conclusions

In this work, we collect a visibility dataset under varying viewing conditions. We propose a hybrid architecture that incorporates a simplified white-box model of visual processing, followed by a black-box deep neural network. Given limited data, we pre-train the our model on a dataset generated with an existing, white-box visibility metric. We demonstrate that the proposed deep visibility metric, combined with our training strategy, can account for the change of viewing conditions and can outperforms the state-of-the-art metric in cross-validation on our new dataset.

8. Acknowledgement

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreements n° 725253EyeCode, n°765911RealVision). The project was also supported by the Fraunhofer and Max Planck cooperation program within the German pact for research and innovation (PFI).

References

[1] T. O. Aydin, R. K. Mantiuk, and H.-P. Seidel. Extending quality metrics to full luminance range images. *Proceedings of SPIE*, 6806:68060B–68060B–10, 2008.

[2] P. Barten. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE Press, 1999.

[3] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. On the use of deep learning for blind image quality assessment. *arXiv:1602.05531*, 2016.

[4] H. Blackwell. Contrast thresholds of the human eye. *Journal of the Optical Society of America*, 36(11):624–632, 1946.

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

[6] M. Čadík, R. Herzog, R. K. Mantiuk, K. Myszkowski, and H.-P. Seidel. New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 31(6):147, 2012.

[7] D. M. Chandler. Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, page Article ID 905685, 2013.

[8] S. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In A. B. Watson, editor, *Digital Images and Human Vision*, volume 1666, pages 179–206. MIT Press, 1993.

[9] DICOM PS 3-2004. Part 14: Grayscale standard display function. In *Digital Imaging and Communications in Medicine (DICOM)*. National Electrical Manufacturers Association, 2004.

[10] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, Dec 1987.

[11] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[12] 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*, 34(6):130–141, Nov 2017.

[13] K. J. Kim, R. Mantiuk, and K. H. Lee. Measurements of achromatic and chromatic contrast sensitivity functions for an extended range of adaptation luminance. In B. E. Rogowitz, T. N. Pappas, and H. de Ridder, editors, *Human Vision and Electronic Imaging*, page 86511A, 2013.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[15] W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *J. Visual Communication and Image Representation*, pages 297–312, 2011.

[16] J. Lubin. *Vision models for target detection and recognition*, chapter A Visual Discrimination Model for Imaging System Design and Evaluation, pages 245–283. World Scientific, 1995.

[17] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel. Predicting visible differences in high dynamic range images: model and its calibration, 2005.

[18] 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):40:1–40:14, July 2011.

[19] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perception-motivated high dynamic range video encoding. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 23(3):730–738, 2004.

[20] S. Miller, M. Nezamabadi, and S. Daly. Perceptual Signal Coding for More Efficient Usage of Bit Codes. *SMPTE Motion Imaging Journal*, 122(4):52–59, 2013.

[21] M. Narwaria, M. P. D. Silva, and P. L. Callet. Hdr-vqm: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46 – 60, 2015.

[22] A. Oliva, A. Torralba, and P. G. Schyns. Hybrid images. *ACM Trans. Graph.*, 25(3):527–532, 2006.

[23] H. R. Wilson. A transducer function for threshold and suprathreshold human vision. In *Biological Cybernetics*, volume 38, pages pp. 171 – 178, 1980.

[24] 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.

[25] W. Zeng, S. Daly, and S. Lei. An overview of the visual optimization tools in JPEG 2000. *Signal Processing: Image Communication*, 17(1):85–104, 2002.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.

[27] X. Zhang and B. A. Wandell. A spatial extension of CIELAB for digital color-image reproduction. *Journal of the Society for Information Display*, 5(1):61, 1997.

A. Effects of the additional ppd feature layer

To test whether adding the angular resolution (*ppd*) to the latent code can model the non-linear effects correctly, we compare the performance of the architecture with or without the ppd feature layer. We denote the architectures as DPVM-Vanilla or DPVM-PPD for the case without or with the ppd feature layer. The means and standard errors of likelihoods for 5-fold cross validation are shown in Table 2. From Table 2 we can observe that introducing the additional ppd feature layer improves the performance. The reason for choosing the middle of the neural network for concatenating the ppd feature layer is two fold: Firstly, concatenating the ppd feature layer in the middle will only introduce a 2X2 feature layer, which will not increase the number of parameters to fit greatly. Secondly, there are three deconvolutional layers with ReLU activations that can provide enough capacity to model the non-linear effects of the angular resolution.

Method	Likelihood
DPVM-Vanilla	0.8556 ± 0.015
DPVM-PPD	0.8772 ± 0.016

Table 2. Effects of the additional ppd feature layer.