Assessment of multi-exposure HDR image deghosting methods

Kanita Karaduzovic-Hadziabdic[1,*], Jasminka Hasic Telalovic[1,], Rafał K. Mantiuk[2,]

Abstract

To avoid motion artefacts when merging multiple exposures into a high dynamic range image, a number of HDR deghosting algorithms have been proposed. However, these algorithms do not work equally well on all types of scenes, and some may even introduce additional artefacts. As the number of proposed deghosting methods is increasing rapidly, there is an immediate need to evaluate them and compare their results. Even though subjective methods of evaluation provide reliable means of testing, they are often cumbersome and need to be repeated for each new proposed method or even its slight modification. Because of that, there is a need for objective quality metrics that will provide automatic means of evaluation of HDR deghosting algorithms. In this work, we explore several computational approaches of quantitative evaluation of multi-exposure HDR deghosting algorithms and demonstrate their results on five state-of-the-art algorithms. In order to perform a comprehensive evaluation, a new dataset consisting of 36 scenes has been created, where each scene provides a different challenge for a deghosting algorithm. The quality of HDR images produced by deghosting method is measured in a subjective experiment and then evaluated using objective metrics. As this paper is an extension of our conference paper, we add one more objective quality metric, UDQM, as an additional metric in the evaluation. Furthermore, analysis of objective and subjective experiments is performed and explained more extensively in this work. By testing correlation between objective metric and subjective scores, the results show that from the tested metrics, that HDR-VDP-2 is the most reliable metric for evaluating HDR deghosting algorithms. The results also show that for most of the tested scenes, Sen et al.'s deghosting method outperforms other evaluated deghosting methods. The observations based on the obtained results can be used as a vital guide in the development of new HDR deghosting algorithms, which would be robust to a variety of scenes and could produce high quality results.

Keywords: HDR imaging, deghosting algorithms, subjective and objective evaluation

2010 MSC: 00-01, 99-00

1. Introduction

To capture the high dynamic range present in most real world scenes, several methods have been proposed: CCD sensors [1, 2], HDR CMOS sen-

*Corresponding author

[1]International University of Sarajevo, BiH

[2]The Computer Laboratory, University of Cambridge, UK

sors, specialized hardware [3, 4, 5], HDR cameras such as SpheronVR GmbH and Panoscan MK-3, and more recently, methods that reconstruct HDR image from a single shot with spatially-varying pixel exposures using commercial cameras [6, 7]. However, the most popular and affordable method in generating HDR images is multi-exposure technique [8, 9, 10], where a sequence of differently exposed low dynamic range (LDR) images is merged to produce an HDR image. By capturing the same scene with a sequence of differently exposed images, each image may have different pixels that are over-, or underexposed as well as the pixels that are properly exposed. High dynamic range image can be generated by combining different exposures to only use well exposed pixels from each image. The following HDR formula computes an HDR image, H_{ij} , from a sequence of LDR images as the weighted average of pixels across N exposures:

$$H_{ij} = \frac{\sum_{k=1}^{N} \frac{f^{-1}(z_{ij}^k) w(z_{ij}^k)}{\Delta t_k}}{\sum_{k=1}^{N} w(z_{ij}^k)}, \qquad (1)$$

where z_{ij}^k is the pixel value at location (i, j) in exposure k, $w(z_{ij})^k$ is the weight corresponding to that pixel, Δt_k is the exposure time for image k. f^{-1} represents the inverted camera response function. Since camera response function is often unknown, it must be first estimated by one of the established techniques [9, 10, 8]. Pixels in the resultant HDR image contain values that are approximately proportional to the luminance of the original scene. Under- and over-exposed pixels can be excluded from the final image by appropriately selecting the weights $w(z_{ij})^k$ [10, 8]. Weighting functions may also be used to reduce ghosting and lower the noise in the generated HDR image [11].

Multi-exposure techniques [8, 9, 10] for generating HDR images work well for static scenes taken on a tripod. However, most everyday photographs contain moving objects and are captured by a hand-held camera. To merge such photographs into HDR images, a number of multi-exposure HDR deghosting algorithms have been proposed [12, 13, 14, 15, 11, 16]. The main goal of these algorithms is to produce a good quality HDR image without motion artefacts, which are usually described as 'ghosting'. As those algorithms often fail to remove all ghosting artefacts and can introduce new distortions, this brings the need to evaluate and compare their results. Subjective quality assessments provide a reliable means of image quality evaluation. However, they are often costly and demanding to perform. Objective quality assessment methods provide computational and automated means of measuring performance of different algorithms.

In [17], Tursun et al. perform a comprehensive survey and classification of approximately 50 HDR deghosting algorithms. They also performed a subjective study to evaluate various state-of-the-art deghosting algorithms. The authors identified the need to evaluate HDR deghosting algorithms by using objective quality metrics as an important part of future work. In their most recent work [18], they proposed a reduced reference objective quality metric to evaluate HDR deghosting algorithms. Hanhart et al. [19] performed an extensive benchmarking of objective quality metrics for HDR image quality assessment. Objective metrics were benchmarked on a dataset of compressed HDR im-

ages [20]. Their findings showed that HDR visual detection predictor (HDR-VDP-2) [21] and HDR video quality metric (HDR-VQM) [22] are most reliable predictors of perceived quality. Karaduzovic et al. [23] performed subjective evaluation of HDR deghosting algorithms and proposed a methodology for subjective evaluation. In another paper, Karaduzovic et al. [24] analyze deghosting HDR algorithms based on expert evaluation. Tursun et al. [25], performed another subjective evaluation of deghosting algorithms for HDR images. They observed that the evaluated deghosting algorithms remove ghost artefacts, but at the same time, they also introduce noise and texture smoothing artefacts.

We further the work of setting the evaluation environment for assessing artefacts that may result in multi-exposure HDR deghosting algorithms. The work contains both subjective and objective assessment of HDR deghosting methods. It includes the first evaluation of objective quality metrics for assessment of multi-exposure HDR image deghosting methods, which is the main difference with previous similar works. In particular, 6 objective metrics are evaluated to test whether they are suitable for prediction of artefacts generated by HDR deghosting methods. In the assessment, the study also contains the most recent objective quality metric, unified deghosting quality metric (UDQM), proposed by Tursun et al. [18]. The metric is especially designed for evaluation of HDR deghosting methods. In [18] the success of UDQM is validated by performing correlation with subjective results. However, the comparison of UDQM results with other objective quality metrics has been done by presenting the outputs of only two other objective metrics, i.e. Liu et al.'s [26] and dynamic range independent metric (DRIM) [27]. DRIM produces as result three distortion maps without any quality value, which are thus difficult to interpret. Furthermore, in this study, a benchmark dataset of 36 scenes that contains raw and jpg ground truth and test multi-exposures has been created. To our knowledge, no other dataset contains a pair of ground truth and test multi-exposure sequence for evaluation of HDR image deghosting methods. The availability of ground truth sequences makes the dataset available to be used with full-reference metrics as well. This benchmark dataset further enriches our multi-exposure dataset described in [24], which deals with 41 real life scenes featuring live objects, but lacks raw and ground truth exposures, as well as the datasets created by [17] and [18] which contain 10 and 16 scenes respectively without the ground truth multi-exposure sequence.

This paper is an extended version of our conference paper [28] and it includes the following contributions:

- Creation of a dataset of 36 carefully selected test and reference images. The dataset can be used as a benchmark dataset to evaluate and compare deghosting algorithms using both subjective and objective means of assessment. For objective assessment, full-reference metrics can also be used as the dataset contains both test and reference multi-exposure sequences which are often timely to obtain.

- An in-depth evaluation of several metrics for evaluating multi-exposure HDR deghosting al-

gorithms (Sen et al. [14], Silk and Lang [16], Hu et al. [15], Photomatix Pro (version 4.2.6) and Photoshop CS5 Extended (version 12.0). We assess the performance of following objective metrics: perceptually uniform peak signal-to-noise ratio (PU2PSNR) [29], perceptually uniform structural similarity index metric (PU2SSIM) [29], Weber root mean square error (Weber RMSE), HDR-VDP-2 (version 2.2.1) [21], unified deghosting quality metric (UDQM) [18], and Liu et al.'s (LR) objective equality metric [26] for motion deblurring.

- Measurement of the success of objective quality metrics by performing subjective evaluation of five algorithms to test whether they can be used to predict deghosting artefacts. The most reliable metric is then selected by computing expected values of Spearman and Pearson correlation coefficients between the two scores computed by bootstrapping.

In particular, the extensions and changes as compared to the conference paper [28] paper are the following:

- Addition of one more objective quality metric (i.e. UDQM [18]) to the assessment.

- Addition of summary of noted comments about HDR image deghosting algorithms from observers during the subjective experiments (Table 3).

- In addition to analyzing subjective experiment results by scaling the pairwise comparison data in Just-Noticable-Difference (JND) units, the results of the subjective experiments were also used to compute statistical significance of the differences between the algorithms by performing multiple comparison test (Figures 3 and 4).

- Spearman and Pearson correlation scores where computed for each scene category by grouping JND values across image sets, and computing correlation with each objective quality metric results, which were also grouped across image sets for each scene category. For completeness, we also computed the correlation scores for each scene.

- Computing *expected* values for Spearman and Pearson correlation coefficients obtained by bootstrapping JND scores, instead of averages. We also include the graph with confidence intervals for Spearman and Pearson scores also computed by bootstrapping.

- This work also includes additional figures that visualize the artefacts generated by the evaluated HDR image deghosting methods and give further insights into the results.

2. HDR Deghosting Algorithms

As already mentioned, merging multiple exposures using Equation 1. will produce motion artefacts, because the equation is based on the assumption that pixels in all exposures are perfectly aligned. As a result a large number of algorithms with different approach to the deghosting problem have been proposed in literature. In [17], Tursun et al. provide a detailed taxonomy of deghosting algorithms. HDR deghosting algorithms can be classified into the following categories: 1) *Global align-*

ment algorithms which address artefacts that result from global camera motion. Such algorithms in general assume that the scene is static. 2) *Moving object algorithms* that address artefacts that result from moving objects in the scene. Such algorithms usually assume that the camera is static. In case the dynamic scene was captured by a hand-held camera, such algorithms usually perform global registration, often by applying one of the existing global alignment methods. The main difference between different moving object algorithms is detection of motion regions and an approach taken to remove ghosting. Following approaches handle moving objects in the scene:

- Rejection of ghost regions algorithms: Such algorithms replace detected motion pixels either with pixels from only one exposure [30, 31], or from multiple exposures [32, 33]. The main drawback of these methods is that if the object in motion contains HDR content, then the dynamic range of the moving object is reduced.

- Reconstruction of motion pixels algorithms: such algorithms align detected objects in motion by searching for the best corresponding pixels in other exposures. Two approaches in finding these correspondences are optical-flow based approach [34, 35], and a patch-based approach [14, 15]. In general, these algorithms are computationally expensive, due to the intensive pixel or patch-based operations.

- Completely removing moving objects from the scene algorithms: these algorithms distinguish moving objects from the static background. The easiest approach is simply to discard motion pixels in HDR merging phase.

In this work we evaluate state-of-the-art algorithms that belong to the *moving object algorithms*, in particular, the methods that fall into the first two categories of such algorithms (i.e. rejection of ghost regions algorithms and reconstruction of motion pixels algorithms). This section provides a brief overview of evaluated HDR deghosting methods. Please refer to the state-of-the-art report [17] for a comprehensive review of approximately 50 HDR deghosting algorithms. In their study, the authors also perform a subjective evaluation of various state-of-the-art algorithms: Grosch [36], Khan et al. [37], Sen et al. [14], Silk and Lang [16], Hu et al. [15]. Since the algorithm by Grosch [36] did not perform well in their evaluation, we did not include it in our study. Because Khan et al. algorithm removes moving objects, the results could not be compared with the reference image (containing those objects in motion) and therefore the method could not be assessed with our existing dataset. The remaining three algorithms Sen et al. [14], Silk et al. [16], Hu et al. [15], are included in our evaluation. Furthermore, we also add to our evaluation two widely used HDR deghosting algorithms integrated into commercial software packages: Photomatix Pro (version 4.2.6) and Photoshop CS5 Extended (version 12.0). An HDR image generated by merging a sequence of RAW images using Robertson et al. method [38] without deghosting is also included in the evaluation as a control condition.

Sen et al.'s [14] algorithm is a patch based method that deals with dynamic scenes with vary-

ing complexity. The main goal of the algorithm is to generate a good quality HDR image from multi-exposure sequence of LDR images that are aligned to the reference image, L_{ref}. The method minimizes an energy function composed of two terms. The first term uses the most well exposed pixels from the reference image. The second term constraints the ill exposed pixels from the reference image to match other exposures by applying a modified bidirectional similarity energy function (EMBDS), which is based on BDS proposed by Simakov et al. [39]. The two terms are balanced by applying per pixel weighting. The weights of ill exposed pixels in the reference image are decreased, whereas the weights of the pixels in the second term are increased. The method optimizes energy function by introducing auxiliary LDR images. The algorithm simultaneously solves for HDR image and auxiliary images using an iterative approach. The iterative approach performs joint optimization of image alignment and HDR merge process until all the auxiliary exposures are correctly aligned to the reference exposure and a deghosted HDR image is produced. With this approach, generated HDR image uses information from all exposures and is aligned to the reference exposure. This approach requires linearized LDR images.

Another patch based algorithm proposed by Hu et al. [15], generates a sequence of registered images from a stack of misaligned images of dynamic scenes captured with a hand-held camera. The method uses an iterative approach to register a sequence of input exposures to a reference image L_{ref}. Initially, the algorithm automatically selects a reference image to be the image with most well-exposed pixels. Then, for each input LDR source image S, the algorithm generates a new latent image L, which looks like the reference image L_{ref}, but is exposed like S. Each latent image is then updated by applying the PatchMatch algorithm by finding corresponding patches between the latent image L and input image S. For well exposed pixels latent images are similar to L_{ref}. For over and under exposed patches, PatchMatch algorithm is modified to find a patch in the input images. During the HDR reconstruction, the algorithm propagates the intensity and gradient information in order to preserve as much detail as possible.

Silk et al.'s [16] method handles dynamic regions by performing change detection on exposure normalized images. In order to improve change detection results, the method performs super pixel segmentation. 'Pairwise down weighting' is applied, where inter-frame change masks are refined to lower the contribution of motion regions to the HDR weighted average on a per frame basis. The method also provides a solution to handle 'fluid' motion, (i.e. non-rigid motion such as foliage blowing in the wind). It is used to handle motion displacements when motion occurs throughout a large portion of input images. 'Fluid' motion is performed by selecting pixels from the best exposure from input images to replace the fluid motion areas.

Granados et al. [11] developed an HDR deghosting method that deals with HDR deghosting by modeling noise distribution of color values measured by the camera. The algorithm initially performs registration of possible misalignments due to the camera motion using global homography computed with RANSAC [40] from SURF [41] key

points matches. Then, an estimation of the readout noise and camera gain is achieved, and an HDR image is constructed by a weighted average from the consistent subset of LDR images. The authors define a pair of pixels in multi-exposed sequence of images to be consistent, if their corresponding colors correspond to the same irradiance and thus refer to the same static object. This is done by analyzing noise distribution of color values. Next, Markov Random Field prior is used to reconstruct irradiance of estimated static objects. The resultant algorithm is robust to high image noise, and does not require the selection of a reference image nor background estimation. However, since the dynamic content is handled by selecting pixels from a single LDR image, the dynamic range of moving HDR content cannot be enhanced using the proposed method. The method requires a multi-exposure sequence of raw images. Because the method was patent pending and hence the source code was not available at the time of conducting this work, we did not include it in our analysis.

The details of algorithms integrated in Photoshop and Photomatix software packages are proprietary.

3. Objective Quality Assessment of HDR Images

In image processing and computer graphics, performance of algorithms often needs to be evaluated and compared with state-of-the-art results. Quality assessment methods are usually used for such purposes. Subjective quality assessments provide a reliable means of image quality evaluation. However, they are often costly and demanding to perform. Objective quality assessment methods provide computational means of measuring the performance of different algorithms. Besides evaluation, quality metrics may also provide further insight into the evaluated algorithm.

In this paper, we assess the performance of several objective quality metrics to test whether they can be used to predict the quality of HDR deghosting algorithms: PU2PSNR, PU2SSIM [29], HDR-VDP2 [21], Weber RMSE, UDQM [18] and LR [26] metrics. The first four metrics are full-reference metrics, UDQM is a reduced reference metric and LR metric is a no-reference metric designed for evaluating motion deblurring. Since most common deblurring artefacts identified by Liu et al. [26] are very similar to the artefacts that may be generated by HDR deghosting methods [24] we also included this metric in our assessment. Additionally, we also tested the performance of LR metric as a full-reference metric (using a deghosted and a reference images as inputs to the metric, rather than the ghosted and deghosted images which are inputs into Liu's no-reference metric.). Only the performance of Liu et al.'s no-reference metric has been considered in [18] for evaluating multi-exposure HDR images generated by HDR deghosting methods, however the performance of remaining metrics for evaluation of HDR deghosting algorithms has not been studied yet.

PU2PSNR and PU2SSIM are extensions of two popular quality metrics PNSR and SSIM [42]. The

PSNR is computed as:

$$PSNR(x,y) = 20\ \log_{10}\frac{peak}{\sqrt{MSE(x,y)}}[dB],$$
$$\text{where} \quad MSE(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{2}$$

where x and y correspond to the pixel values in the reference, and reconstructed image respectively. *peak* refers to the maximum luminance value, and MSE refers to the mean square error between two input images. Lower the MSE, higher the PSNR, and thus better the quality of the reconstructed image.

SSIM is another widely used quality metric:

$$SSIM(x,y) = l(\mu_x,\mu_x)^\alpha c(\sigma_x,\sigma_x)^\beta s(\sigma_x,\sigma_x)^\gamma, \tag{3}$$

where x and y correspond to the pixel values in reference and reconstructed image respectively. μ and σ are mean and standard deviation of input images. SSIM index is then computed by a weighted combination of luminance comparison function l, contrast comparison function c, and structure comparison function. The above formula for SSIM index is designed for LDR images and must be adapted to be applicable to HDR images.

The above formulas for PSNR and SSIM are designed for LDR images and must be adapted to be applicable to HDR images. In [29], Aydın et al. proposed an extension to these two quality metrics, which enables comparing HDR images at all luminance levels visible to the human eye without affecting their results for typical CRT display luminance levels. The proposed extension applies a perceptually uniform (PU) encoding transfer function that transforms luminance values in the range from 10^{-5} to $10^{8} cd/m^2$ into approximately perceptually uniform code values. The obtained code values are used in the quality metric instead of gamma corrected RGB or luma values. This extension applied to the PSNR and SSIM metrics make them suitable for evaluation of HDR images. Therefore, before using the well established PSNR and SSIM metrics, in this work, PU encoding transfer function is applied to each deghosted HDR image, and the obtained approximately perceptually uniform code values are then applied to PSNR and SSIM metrics.

HDR-VDP-2 [21] metric is based on a visual model that can predict visibility and quality difference between a reference and test image pairs. The metric works with the full range of luminance values that are present in real-world scenes (i.e. HDR images). For visibility differences, the metric produces difference and probability maps between test and reference images. Difference map provides the information how well will the observer notice the difference between two images, (red color indicates high probability, green color low probability). Probability map produces the probability of detection map, which shows where and how likely a difference between two images will be noticed by an observer. Difference map shows the contrast-normalized per-pixel difference weighted by the probability of detection. For quality differences, the metric produces a mean-opinion score ('Q' score) which computes the quality degradation of a test image with respect to the reference image.

Recently, Tursun et al. [18] proposed an objective quality metric, UDQM, specially designed for evaluating HDR deghosting algorithms. The metric is a reduced reference metric whose inputs are a sequence of multi-exposures, acquisition settings (ex-

posure time, ISO and f-number), and a deghosted image. Based on the most common artefacts generated by HDR deghosting methods, the metric uses several objective quality metrics which are specially designed and tuned to evaluate such artefacts. An overall quality score is computed as a weighted sum of proposed individual metrics.

Liu et al. [26] developed a no-reference metric, LR, designed for evaluating motion deblurring. The metric is based on a set of 8 optimally selected features designed to measure the most common deblurring artefacts (ringing, noise and residual blur). The proposed metric is than trained to obtain the optimal weights for each selected feature.

4. Dataset

Since algorithm performance may be scene dependent, we created a dataset particularly designed to provide a comprehensive set of challenging scenes for evaluating deghosting algorithms. In [24], Karaduzovic et al. identified the most common artefacts that could be introduced by a deghosting process: motion artefacts, loss of dynamic range (i.e. amount of details visible), noise and color artefacts. The identified artefacts and the Middlebury dataset proposed by Baker et al. [43] were used as a guideline for creating the dataset, which resulted in carefully categorized scene types. The dataset contains 36 scenes organized into 4 different *image sets*. Each image set refers to a specific lighting condition under which 9 categorized scenes, each with different type of motion, have been captured in a controlled environment. Scene type categorization are listed in Table 1. For each scene, both *test* and *reference* multi-exposure sequence were captured. Test exposure sequence refers to the sequence of multi-exposures that contain either objects or camera motion. Reference exposure sequence refers to the sequence of multi-exposures where all pixels are perfectly aligned (i.e. ground truth sequence). For exposure sequences with objects in motion, the position in the middle of the motion was selected for the reference exposure sequence. The availability of a pair of ground truth and test multi-exposure sequences is a unique feature of the dataset because it makes the dataset suitable to be used with well-established full-reference metrics suitable for HDR images (e.g. PU2PSNR, PU2SSIM, HDR-VDP-2). However, because multi-exposure sequences need to be captured in a controlled environment, the dataset does not contain very complex non-controllable motion such as people and pets. Figure 1 shows 4 representative scenes used in the experiments.

In order to evaluate multi-exposure HDR deghosting methods, other than having different motion types captured under various lighting conditions, the captured dataset should also contain scenes with wide dynamic range. Furthermore, multi-exposure stack should also have saturated pixels because such pixels make the correspondence, and hence the deghosting process, more difficult.

Dynamic range of our captured HDR ground truth images ranges from $2.42 - 3.89$ orders of magnitude (see Table 2). It is measured as the logarithm of the ratio between the brightest and the darkest luminance present in the scene: $log10(Ypeak/Ynoise)$. In order to account for a reliable noise level in an HDR image, we applied

a Gaussian smoothing filter to the HDR image, and then computed the maximum (i.e. $Ypeak$) and minimum ($Ynoise$) luminance stored in the ground truth HDR image for each scene. We also computed the number of saturated pixels (i.e. percentage of pixels where at least one of the RGB channels has pixel value greater than 0.996) in the multi-exposure stack for each scene (see Table 1) Complete dataset containing both RAW and JPG images is available for the research community [44].

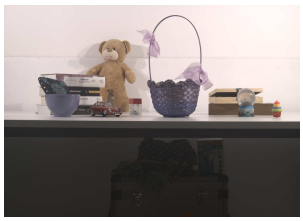
4.1. Acquisition

Scenes marked with * in Table 1 indicate dynamic scenes captured on a tripod where objects were moved between LDR image capture to simulate motion. Each image set consists of five exposures with one f-stop exposure time difference. The first three sets were captured in a dark room where for *set 1*, the only source of illumination was coming from a Halogen 300 Watt spot light, positioned at 45 degrees to the table containing objects in motion and two 60 Watt light bulbs-white positioned at 45 degrees on the other side of the table; for *set 2*, the light source was coming from a 2×300 Watt Halogen spot photographic light positioned at 45 degrees to the table containing objects in motion on both sides; for *set 3*, the light source was coming from a table lamp with 60 Watt light bulb; *set 4* was captured in a room where the camera was pointing towards a large window. Figure 1 shows 4 representative scenes used in the experiments. RAW images with linear response curves were captured to minimize the internal camera image processing by Canon EOS DSLR 1000D camera. In order to avoid camera motion, camera was remotely controlled by gPhoto2 (version 2.5.5. http://gphoto.sourceforge.net) and mounted on a tripod. The only scenes where camera was not mounted on a tripod were *handheld* and *multi-view* scenes. Since most deghosting algorithms are computationally expensive, and often fail to process higher resolution images the resolution of all images was rescaled to 1953×1301. This resolution is half the resolution of 16-bit tiff images obtained from captured RAW images. To get the best algorithm performance, we used linear 16-bit images as inputs to the algorithms. Whenever possible, we tried to use the fine-tuning options suggested by the authors. For subjective experiments, generated HDR images were tonemapped by applying a customized tone mapping operator (TMO) [23] based on the fast bilateral filter [45]. The main goal of this TMO is to reproduce details exactly as they were captured in HDR images and compress low-frequence content to fit within a dynamic range of the display.

5. Subjective Experimental Setup

20 participants with computer graphics background, aged between 22 and 41, performed the pairwise comparison experiment. All participants reported normal or corrected to normal vision. Each participant was presented with randomized all possible comparison pairs of the same scene processed with a different deghosting algorithm. Psychtoolbox-3 (http://psychtoolbox.org) was used to design the experimental stimuli and program the experiment. For 6 evaluated algorithms (5 deghosting and 1 without deghosting) and 36 scenes, the to-

Example scene, image set 1


Example scene, image set 2


Example scene, image set 3


Example scene, image set 4

Figure 1: Four example scenes used in the experiments, one scene from each image set.

Table 1: Different types of scenes contained within each image set. Scenes marked with * indicate dynamic scenes captured on a tripod where objects were moved between LDR image capture to simulate motion. Columns 3, 4, 5 and 6 show the percentage of saturated pixels in multi-exposure stack for each scene.

Scene name	**Scene description**	image set1	image set2	image set3	image set4
*complex motion	Highly dynamic scene with small/large motion displacement of small/large objects, non-rigid motion, occlusion, and several independently moving objects.	0 - 2	0 - 40	0 - 13	0 - 35
handheld	Static scene captured with a handheld camera	0 - 4	0 - 34	0 - 12	0 - 23
*lolm	Large object displacement with large motion.	0 - 2	0 - 42	0 - 7	0 - 24
*losm	Large object displacement with small motion.	0 - 3	0 - 42	0 - 9	0 - 31
multiview	Multi-view sequence of a static scene.	0	0 - 37	0 - 12	0 - 6
*nrm	Motion of non-rigid and high texture objects.	0 - 1	0 - 44	0 - 9	0 - 24
*occlusion	Scene containing occlusion.	0	0 - 41	0 - 7	1 - 30
*solm	Small object displacement with large motion.	0 - 5	0 - 45	0 - 9	1 - 33
*sosm	Small object displacement with small motion.	0 - 5	0 - 44	0 - 9	1 - 35

Table 2: Dynamic range of the captured HDR ground truth image in orders of magnitude, measured as the logarithm of the ratio between the brightest and the darkest luminance present in the scene: $log10(Ypeak/Ynoise)$.

image set 1	**complex**	**handheld**	**lolm**	**losm**	**multiview**	**nrm**	**occlusion**	**solm**	**sosm**
dynamic range ($log10$)	3.056	3.20	2.84	2.82	2.96	3.18	2.89	2.78	2.79
image set 2	**complex**	**handheld**	**lolm**	**losm**	**multiview**	**nrm**	**occlusion**	**solm**	**sosm**
dynamic range ($log10$)	2.43	2.42	2.56	2.52	2.49	2.46	2.89	2.54	2.42
image set 3	**complex**	**handheld**	**lolm**	**losm**	**multiview**	**nrm**	**occlusion**	**solm**	**sosm**
dynamic range ($log10$)	3.56	3.65	3.71	3.70	3.65	3.72	3.77	3.68	3.70
image set 4	**complex**	**handheld**	**lolm**	**losm**	**multiview**	**nrm**	**occlusion**	**solm**	**sosm**
dynamic range ($log10$)	2.72	2.85	2.84	2.89	3.26	2.95	2.90	2.70	2.72

tal number of comparison pairs was $36 \times \binom{6}{2} = 540$. The experiment was divided into 4 sessions, where each session contained 9 scenes from each image set (i.e. 135 comparison pairs) that were presented randomly for each participant. Each observer participated in all 4 sessions, where each session lasted maximum 30 minutes. Screen position of the images within each pair was also randomized

Table 3: Summary of noted comments from observers during subjective evaluation.

Algorithm	Motion	Dynamic range recovery	Noise	Color
No-degh.	Severe ghosting and blur artefacts.	No visible artefacts.	Noise in low luminance regions in some scenes.	Visible color artefacts in regions with very large motion displacement.
Hu et al.	Good in deghosting.	Possible dynamic range reduction in very high intensity regions of the scene.	Noise visible in low luminance regions in some scenes.	Visible color artefacts in high intensity regions.
Photomatix	Ghost and blur artefacts in regions with large motion displacement.	Reduced dynamic range in some scenes.	Noise in low luminance regions in some scenes.	No visible artefacts.
Photoshop	Generally good in deghosting.	Reduced dynamic range in some scenes.	Noise in low luminance regions in some scenes.	Visible color artefacts in high intensity regions in some scenes.
Sen et al.	Good in deghosting. Contains ghost artefacts in scenes that contain very large over saturated region.	No visible artefacts.	Noise visible in low luminance regions.	No visible artefacts.
Silk et al.	Visible blur. Produces large 'washed out' patches where there was motion.	Severe loss of dynamic range, especially in regions that contain motion.	No visible artefacts.	Dark images, loss of color.

(left/right). Each image pair was displayed side by side on two 21″ 1600×900 HP 2011x LCD monitors. Monitors were rotated 20° around the vertical axis (to be perpendicular to the viewing direction) and at an eye level of the participants, with a viewing distance of 70 cm. All experiments were performed in a dark room where the only light source was coming from a corridor light, which was constant throughout the experiments. Based on the most common artefacts that could be introduced by a deghosting process [24], the participants were asked to choose the preferred image based on the following criteria: firstly, select an image that has the least amount of motion artefacts. If it is not possible to make a difference between an image pair based on motion artefacts, (i.e. there was no visible difference in motion artefacts between image pair), select the image that has lower amount of any of the following three artefacts: loss of details in under-/over-exposed regions, an image with least amount of color artefacts, or an image with least amount of noise, if any. If there was no visible difference between images, participants still had to make a choice between the images. An observer was asked to note down some general comments about the presented images. Table 3 provides a summary

of noted comments. No time limit was imposed during a selection of the preferred image. Before the start of the experiment, a short briefing on possible multi-exposure HDR deghosting artefacts was presented to the participants. A pilot study was performed to evaluate the time required for participants to perform an experiment session, and the overall clarity of the experiment.

6. Objective Evaluation

Six objective metrics were tested whether they can predict deghosting artefacts. The input into full-reference objective metrics was two HDR images, a reference and a test image. Because each evaluated method produces slightly different HDR pixel values (in terms of both contrast and absolute values), test and reference HDR images were generated individually by each method. To minimize possible small pixel misalignments, each reference image was aligned to the test image by homographic transformation found from SURF key-point matching (*pfsalign* command from *pfstools*[46]) which implements an HDR alignment algorithm introduced by Tomaszewska et al. [47]. The input into reduced reference UDQM [18] metric is a stack of multi-exposures, acquisition settings (exposure time, ISO and f-number) and a deghosted image. Handheld images were also aligned by [47].

7. Results

7.1. Subjective experiment results

The results of the subjective experiments were analyzed by estimating which portion of the population would select one algorithm over another. To do this, pairwise comparison data was scaled in Just-Noticable-Difference (JND) units (Figure 2) under Thurstone Case V assumptions, where the difference in 1 JND unit corresponds to 75% of observers selecting one algorithm over another. To scale the pairwise comparison data in JND units, we applied Bayesian method of Silverstein and Farrel [48]. Briefly, the method scales the collected data by solving for maximum likelihood estimator explaining the experiment under the Thurstone Case V assumptions. Applied Bayesian method is robust to unanimous answers, which are common when a large number of methods are compared. For better visualization, JND value for 'non-deghosted' method is set to the baseline 1. In this manner, JND values greater than one indicate that the deghosting method generates an HDR image which reduces artefacts, and values less than one indicate that the deghosting method introduces different types of artefacts during the deghosting process. The error bars in Figure 2 denote 95% confidence intervals computed by bootstrapping JND values.

The results show that Sen et al.'s method outperforms other algorithms for almost all motion types in image sets 1, 2, and 3 (see Figure 2, 3, 4, 5, 6, 7, 8). In general, the method is good in deghosting all tested motion types, the only challenge for this algorithm are the scenes where the reference image is over saturated (e.g. scenes in image set 4), and the method produces visible artefacts (Figure 10).

The algorithm by Hu et al. is the second best performing algorithm in image sets 1 and 2. In image set 3, this algorithm produces color artefacts in over-saturated regions. These artefacts are mostly

visible in area close to the lamp's light bulb in most of the scenes in image set 3 (Figure 9). Similarly to Sen et al.'s method, Hu et al.'s method is generally good in deghosting for all motion types (including complex motion and non-rigid motion, Figure 5). Like Sen et al.'s method method, Hu et al.'s method also generates visible artefacts in regions where the reference image is over-saturated (scenes from image set 4, Figure 10).

In general, it was found that Photoshop outperforms Photomatix in almost all scenes. It has also been observed that the dynamic range of moving content is often reduced in images produced by Photoshop and Photomatix. Furthermore, for some scenes, Photoshop produces color artefacts in high intensity regions. Figure 9 shows such artefacts. Both methods struggle with non-rigid motion generating motion artefacts (Figure 5).

Subjective results show that for most scenes, Silk et al.'s algorithm has the lowest score from the evaluated algorithms (not considering the non-deghosted image). It has also been observed that Silk et al. algorithm produces images of reduced dynamic range and the black level is elevated (see Figure 5, 6, 7, 8).

The results of the subjective experiments were also used to compute statistical significance of the differences, between the algorithms by performing multiple comparison test. Figures 3 and 4 show ranking and rating of the evaluated methods using such analysis. Evaluated algorithms are ordered according to their ranking, with the most preferred algorithm on the right. The X-axis represents rating of each algorithm, expressed in JND units, where higher number of votes corresponds to higher quality. The algorithms connected by the continuous blue lines are statistically different at the significance level $\alpha = 0.05$. The red dashed lines indicate no statistical difference between the methods or that the difference cannot be measured with the collected data. These figures show that Sen et al.'s method was significantly better than Hu et al.'s method in 24 out of 36 scenes. The results also show that Photoshop's method was significantly better than Photomatix' method in 22 out of 36 scenes.

7.2. Objective metric results

To measure the success of objective quality metrics, the metric prediction error was determined by Spearman (ρ) and Pearson (r) correlation coefficients computed between subjective experiment results scaled in JND units and objective quality metrics' values.

To compute Spearman and Pearson correlation scores, for *each* scene category (see Table 1), we grouped a set of JND values (i.e. from 6 evaluated algorithms), across all four image sets and computed the correlation with each objective quality metric result, which were also grouped across image sets for each scene category. Thus, for each scene category, two vectors of size 1×24 (6 algorithms $\times 4$ image sets) were used to compute Spearman and Pearson correlation coefficients.

Since the relation between subjective JND values and objective metric prediction can be non-linear, a logistic function was fitted to map from the metric scores to JND values. This is a standard procedure when using Pearson correlation to evaluate metric performance [49].

The limitation of JND scaling is that the JND

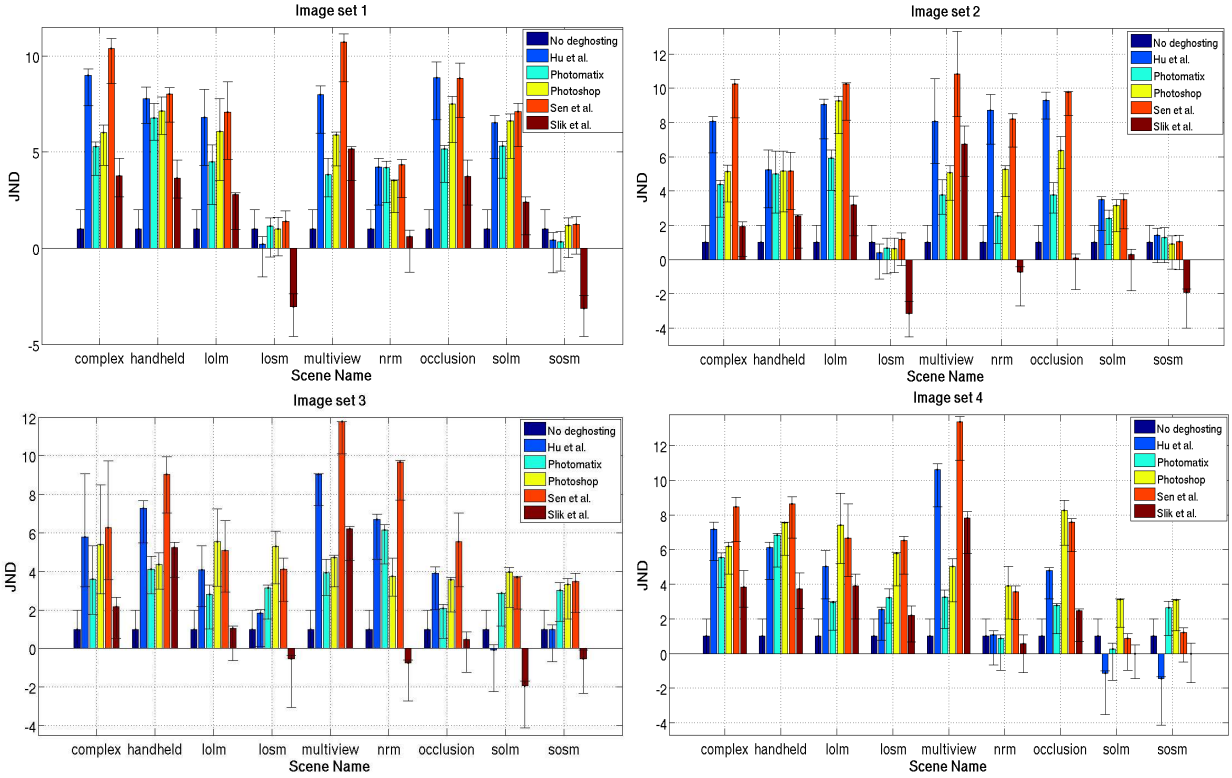


Figure 2: The results of the subjective experiment for all 4 image sets scaled in JND units (higher the values, the better) under Thurstone Case V assumptions, where the difference in 1 JND unit corresponds to 75% of observers selecting one algorithm over another. Absolute values are arbitrary and only the relative differences are relevant. The error bars denote 95% confidence intervals computed by bootstrapping.

values give only relative measure of quality across compared conditions, by themselves they cannot provide absolute measure of quality. As a consequence of that, the JND values measured for one scene are not comparable to JND values measured for another scene. Because of that, we were able to compute metric correlation values only separately for each scene in our prior conference paper [28].

To be able to compute a single correlation score across all scenes, in this work we adjust the JND values for each scene to so that they are comparable across all scenes. To do this, we introduce an offset o_k when fitting the logistic function:

$$f(x) = \frac{a_0}{1 + e^{-a_1 * (x - a_2)}} + o_k \tag{4}$$

where x are objective metric outputs, a_0, a_1, and a_2 are the logistic function parameters that determine its shape. o_k is a score offset value for scene k. When fitting a logistic function, an individual o_k value was found for each scene in our dataset, except the first scene, which served as a point of reference. The fitting was repeated for each objective metric.

An example of logistic function fitting is displayed in Figure 11. Values highlighted bold in Table 4 represents statistically significant Spear-

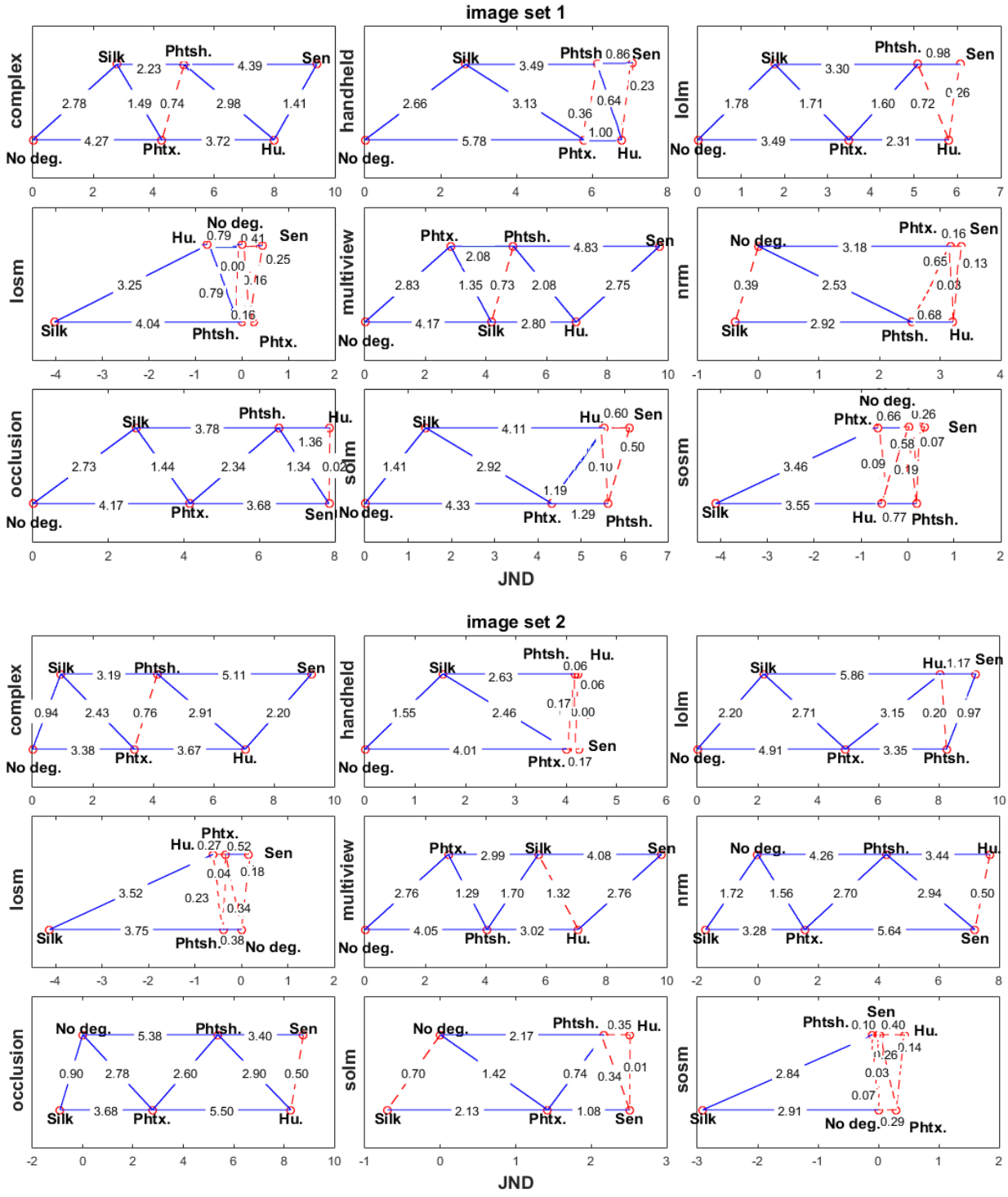


Figure 3: Ranking and rating of the evaluated algorithms for image set 1 and image set 2. Algorithms are ordered according to their ranking, with the most preferred algorithm on the right. The X-axis represents rating of each algorithm, expressed in JND units, where higher number of votes corresponds to higher quality. Blue continuous lines indicate statistical significance at $\alpha = 0.05$, and red dashed lines indicate lack of the statistical difference.

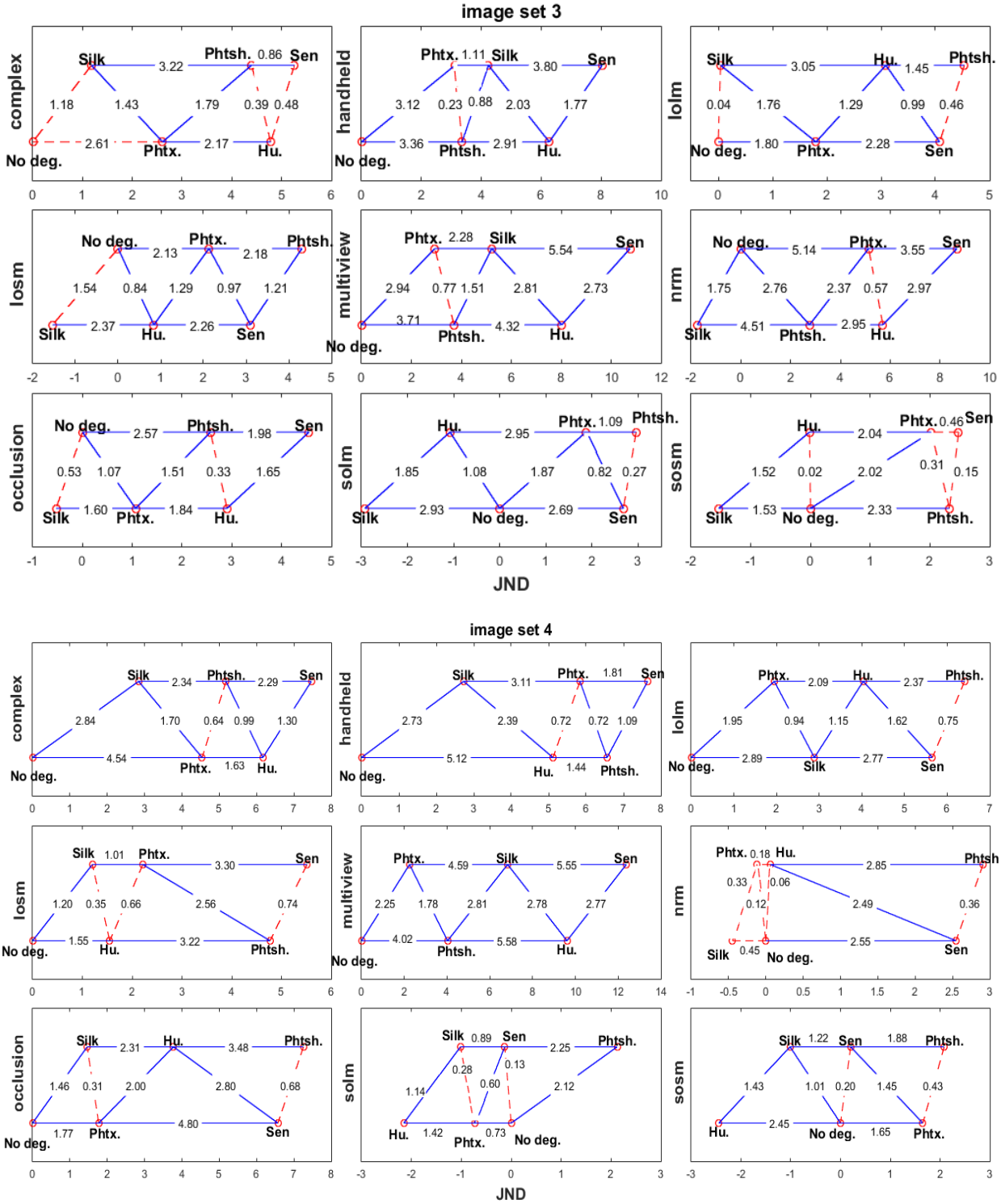


Figure 4: Ranking and rating of the evaluated algorithms for image set 3 and image set 4. Notation is the same as in Figure 3

Figure 5: Outputs of images generated by HDR deghosting algorithms (image set 3, *non-rigid motion* scene). Moving objects in the scene are white roses and small violet plant in the vase (i.e. non-rigid objects). (a) Silk et al.'s method generates motion artefacts in non-rigid regions of the scene. In addition, this method produces color artefacts, reconstructing an unnatural looking image resulting in worst ranked method in subjective experiment (Fig 2). Both (c) Photoshop and (d) Photomatix produce motion artefacts in non-rigid region of the scene (marked regions in (c) and (d)). (e) Hu et al. and (f) Sen et al. perform well in deghosting non-rigid objects. However, Sen et al. outperforms Hu et al., due to the color artefacts generated by Hu et al.'s method (marked region in (e)).

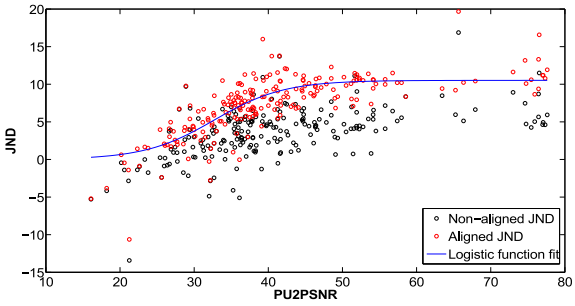


Figure 11: An example of logistic function fitting for PU2PSNR metric.

man and Pearson correlation scores at $\alpha = 0.05$ using a t-test distributed as Student's distribution with 18 degrees of freedom. For each objective metric, expected value for all 36 scenes was computed by bootstrapping JND values. Bootstrapping (i.e. random sampling with replacement of collected subjective data) was done to create multiple randomized samples of the same size so that confidence intervals and expected values could be computed. 500 bootstrap samples were generated and the data was scaled in JND units for each bootstrap sample.

Figure 13 shows Spearman and Pearson correlation scores where error bars denote 95% confidence intervals also computed by bootstrapping.

For completeness, we also computed *per scene* Spearman and Pearson correlation coefficients be-

Figure 6: Tone mapped outputs of images generated by HDR deghosting algorithms, (image set 1, *complex* scene). Marked regions in images (a), (b), (c) and (d) show artefacts generated by No deghosting, Silk et al., Photomatix and Photoshop methods respectively. Images generated by Hu et al. and Sen et al. methods do not produce any visible ghosting artefacts and are therefore not shown in this figure.

tween subjective results and objective scores for each image set (Tables 5 - 8). For Pearson correlation, fitting of the logistic function of the form in Equation 4 was again used, but without the offset vector o_k, because per scene rather than across scenes correlation was used. Table 9 shows the aggregate results for each scene category, where values are averaged across image sets using the computed per scene correlations displayed in Tables 5 - 8.

The results shown in all correlation tables (i.e. Tables 4 - 9) show that HDR-VDP-2 metric has the highest correlation scores for almost all scenes. One of the emerging patterns for full reference metrics is that in general, all metrics except the HDR-VDP-2, show weak correlation for the small-object-small-motion (*sosm*) scene. Even HDR-VDP-2 metric has the lowest correlation score for this scene in image set 1 (Table 5) and image set 2 (Table 6), when compared to HDR-VDP-2 scores of other scenes. In particular for complex motion scenes, as well as for scenes with large displacements of large objects, the correlation results of full reference metrics are higher than for small motion displacements and motion of small objects. This suggest that human eye may not be as sensitive to these small pixel changes as computational metrics.

Figure 12 shows the graph for HDR-VDP-2 'Q' results (higher the values, the better). These re-

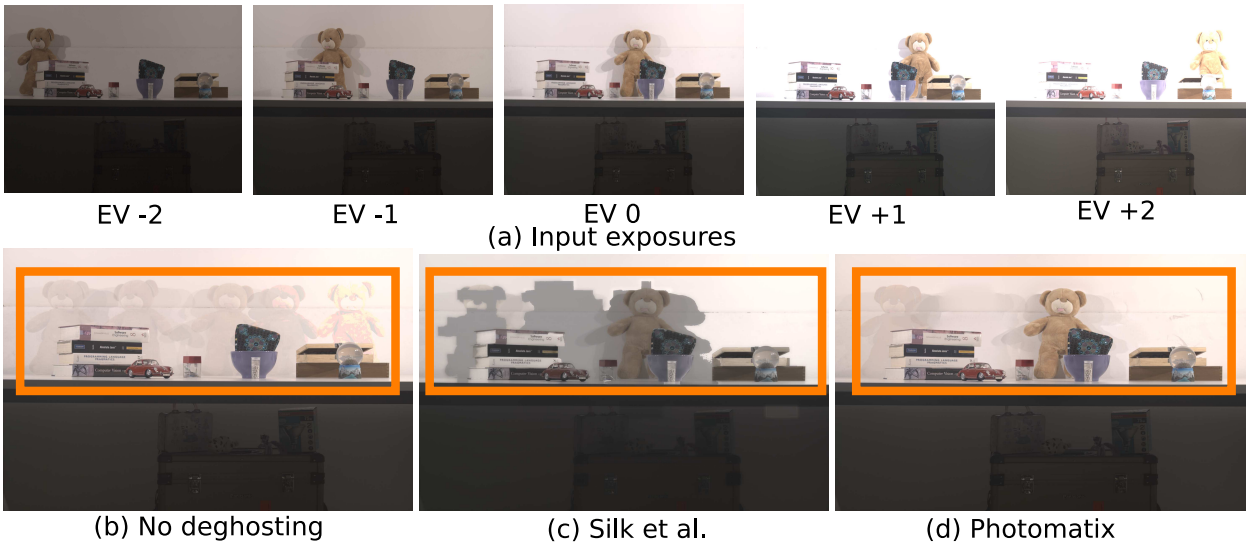


Figure 7: Outputs of images generated by HDR deghosting algorithms with 5 exposures, (image set 2, *large object large motion* scene). Marked regions in images (b), (c) and (d) indicate artefacts produced by No deghosting, Silk et al. and Photomatix methods respectively. Images generated by Photoshop, Hu et al.'s and Sen et al.'s methods do not produce any ghosting artefacts and are therefore not shown in this figure.

sults can be used to compare the performance of other existing and future deghosting methods against those already tested, by simply applying that method to benchmark dataset images by computing the 'Q' value HDR-VDP-2 metric.

Low correlation scores for Liu et al.'s metric imply that this metric is not suitable for evaluating HDR deghosting methods.

8. Discussions and Conclusions

This paper is an extended version of our conference paper [28], where subjective and objective assessment of five state-of-the-art HDR deghosting algorithms has been performed. The extensions include addition of one more objective quality metric (i.e. UDQM metric) in the evaluation; summary of comments for evaluated HDR deghosting algorithms obtained during the subjective experiments; statistical significance plots for JND units; computation of Spearman and Pearson correlation coefficients for each scene category by grouping both subjective and objective scores across image sets, as well as per scene correlations; expected values and confidence intervals of Spearman and Pearson correlation coefficients computed by bootstrapping JND scores. The paper also includes a more detailed insight of subjective experiment results that includes several additional figures which visualize artefacts generated by evaluated HDR deghosting algorithms.

We created a comprehensive dataset that can be used to evaluate multi-exposure HDR deghosting algorithms. Because algorithm performance may be scene dependent, we created a benchmark HDR dataset that consists of 36 scenes, each posing a different challenge to a deghosting algorithm. Other

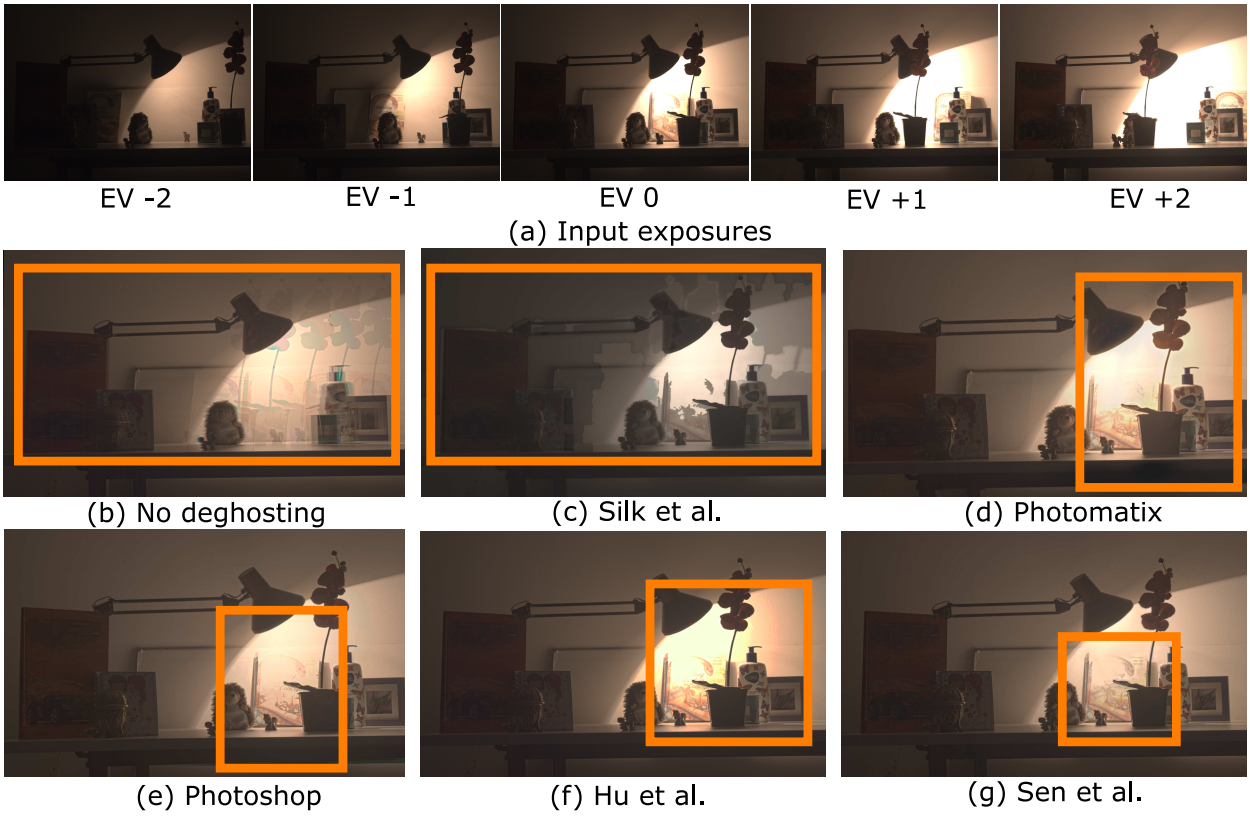


Figure 8: Outputs of images generated by HDR deghosting algorithms with 5 exposures, (image set 3, *complex* scene). All of the methods struggle with reconstructing the high-intensity region of the scene, with Sen et al.'s method producing the best results. Marked regions indicate various artefacts: in (b) No deghosting method produces ghosting and color artefacts, in (c) Silk et al. produces ghost, reduced dynamic range and color artefacts, in (d) Photomatix and (e) Photoshop generate images with reduced dynamic range (around the light bulb) and ghost artefacts (below the table), and in (f) Hu et al. and (g) Sen et al. methods produce an image with reduced dynamic range.

than the larger number of scenes, the main difference between the benchmark HDR dataset provided by Tursun et al. [17] and our dataset is that our dataset also contains multi-exposure sequence of reference (ground truth) images without any ghosting. This feature of the captured dataset makes it also suitable to be used with the full-reference quality metric, such as HDR-VPD-2, PU2PSNR, and PU2SSIM. Captured dataset is made publicly available and can therefore be used for future evaluations of existing and future HDR deghosting algorithms. Then, subjective experiments were performed based on the most common HDR deghosting algorithms' artefacts. Besides analyzing the subjective results, from the graph obtained by scaling the results of the subjective experiment in JND units, we also computed statistical significance of the differences between algorithms.

From the results based on the performed evaluation we make the following observations which provide strengths and weaknesses of evaluated methods:

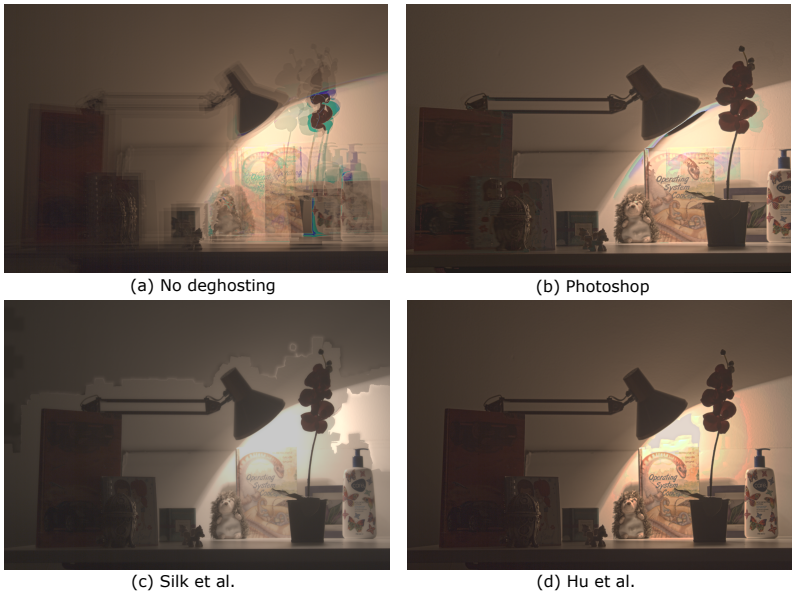


Figure 9: Visible color artefacts generated in over-saturated region of the scene (image set 3, *handheld* scene), generated by (a) No deghosting, (b) Photoshop, (c) Silk et al.'s and (d) Hu et al.'s methods.

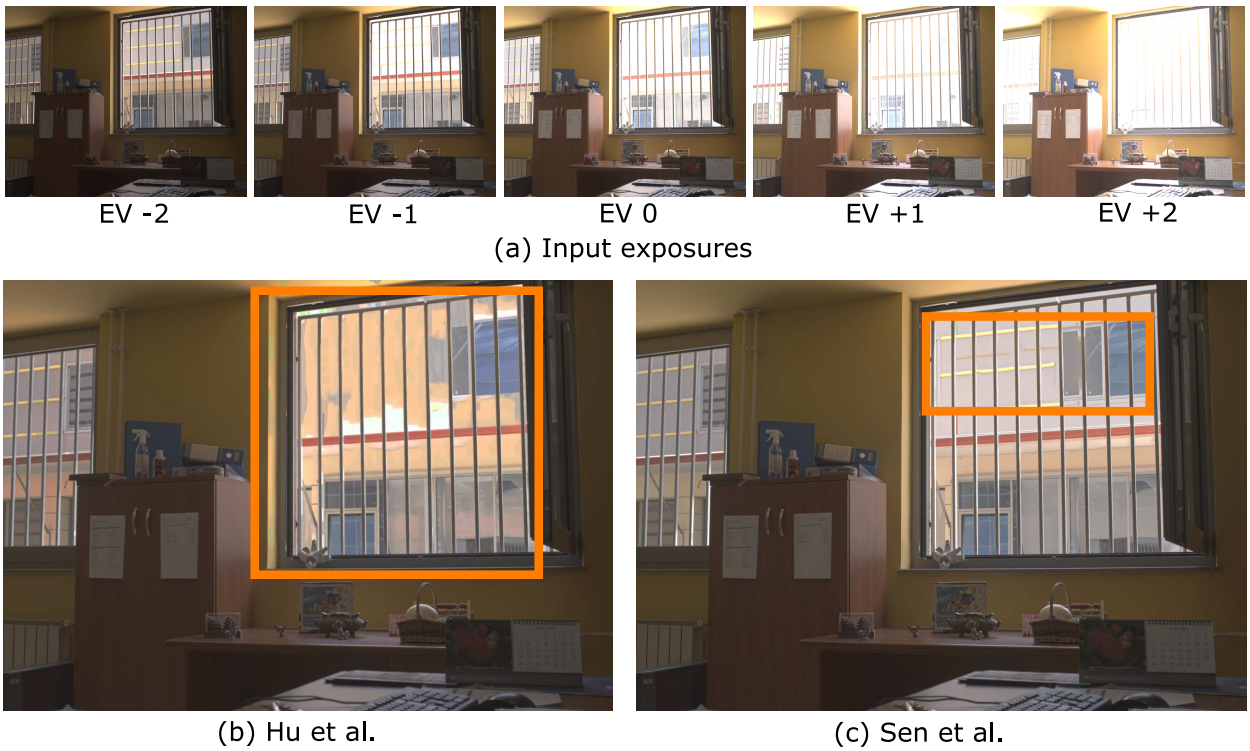


Figure 10: Outputs of images generated by (b) Hu et al.'s and (c) Sen et al.'s methods produce visible artefacts (image set 4, *small object large motion* scene). Marked region in image (b) shows reduced dynamic range produced by Hu et al.'s method in over-saturated region of the scene, and in image (c) ghost artefact produced by Sen et al.'s method.

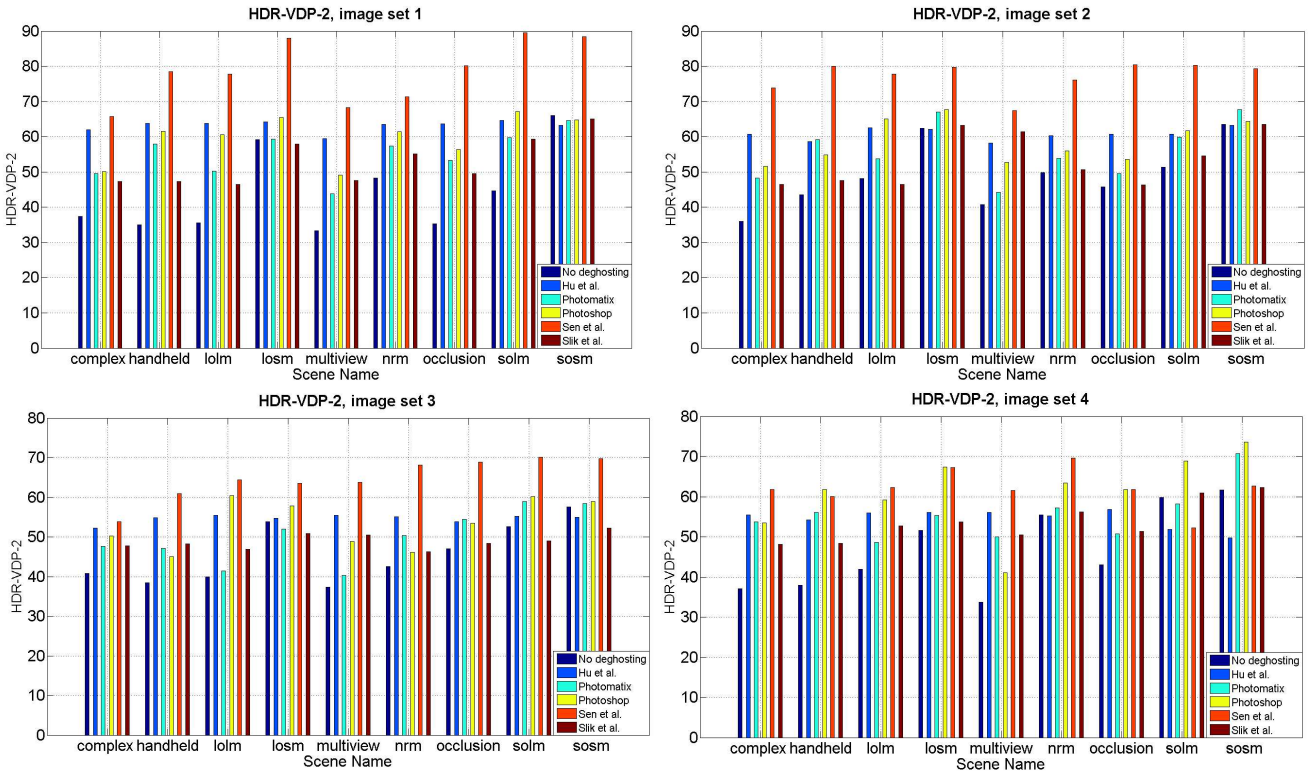


Figure 12: The results of the HDR-VDP-2 metrics for all 4 image sets (higher the values, better the result).

Table 4: Spearman's (ρ) and Pearson's (r) correlation coefficients for all image sets for relation between objective metric predictions and subjective evaluation scores. Correlation scores are computed for each scene category by grouping values across all four image sets. Bolded values represent statistically significant correlation scores at $\alpha = 0.05$. Expected values are computed by bootstrapping.

	PU2PSNR		PU2SSIM		HDRVDP2Q		WeberRMSE		UDQM		LR		LR with ref.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
complex	**0.86**	**0.76**	**0.82**	**0.79**	**0.95**	**0.90**	**0.88**	**0.87**	0.39	0.32	0.28	0.24	0.30	0.29
handheld	**0.88**	**0.75**	**0.90**	**0.90**	**0.92**	**0.87**	**0.88**	**0.87**	0.17	0.08	0.14	0.04	-0.02	0.07
lolm	**0.88**	**0.69**	**0.70**	**0.50**	**0.87**	**0.81**	**0.83**	**0.68**	0.16	0.28	0.05	-0.02	0.09	0.01
losm	**0.63**	**0.52**	**0.59**	**0.50**	**0.73**	**0.67**	**0.54**	**0.48**	**0.51**	**0.66**	0.02	-0.01	0.01	-0.06
multiview	**0.79**	**0.78**	**0.81**	**0.70**	**0.91**	**0.84**	**0.74**	**0.69**	0.35	0.34	-0.23	-0.21	0.31	0.28
nrm	**0.64**	**0.65**	**0.69**	**0.56**	**0.90**	**0.84**	**0.51**	**0.60**	0.42	**0.59**	-0.12	-0.02	-0.11	-0.07
occlusion	**0.74**	**0.68**	**0.68**	0.44	**0.94**	**0.81**	**0.69**	**0.57**	0.30	0.37	0.05	-0.03	-0.14	-0.08
solm	**0.57**	**0.58**	0.44	0.36	**0.85**	**0.76**	**0.45**	0.37	0.44	**0.52**	-0.02	-0.05	-0.03	-0.06
sosm	0.40	0.42	**0.48**	0.39	**0.71**	**0.67**	0.21	0.34	**0.45**	**0.68**	-0.06	0.00	0.04	0.00
Expected value	**0.71**	**0.62**	**0.66**	**0.56**	**0.85**	**0.77**	**0.62**	**0.59**	0.35	0.40	0.07	0.03	0.09	0.08

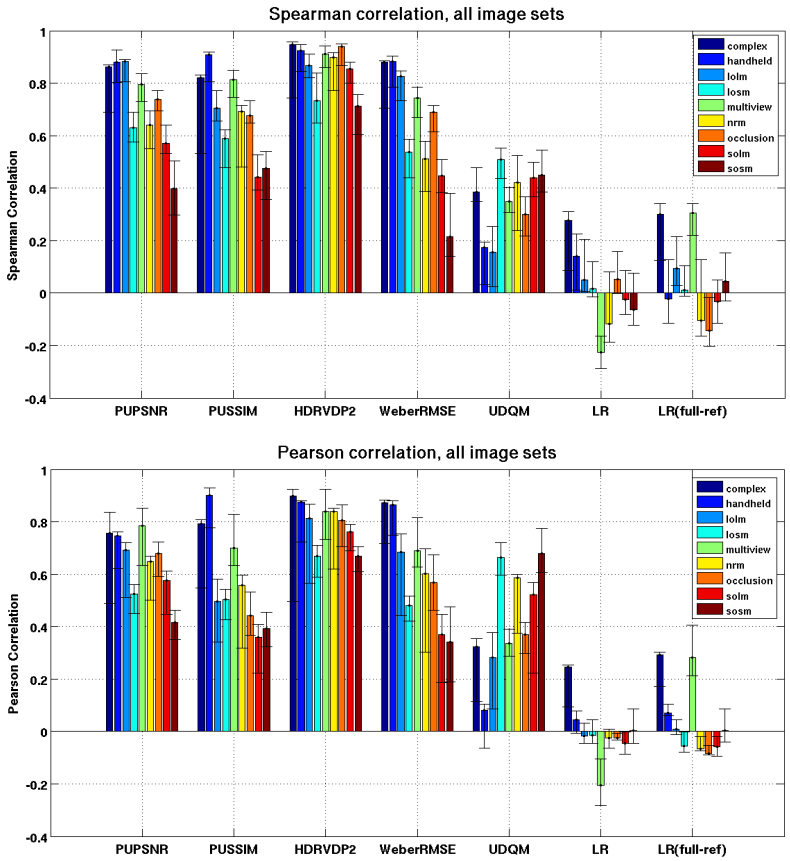


Figure 13: Spearman and Pearson correlation scores for all image sets. Correlation scores are computed for each scene category by grouping values across all four image sets. The error bars denote 95% confidence intervals computed by bootstrapping.

Table 5: Per scene Spearman's (ρ) and Pearson's (r) correlation coefficients for relation between objective metric predictions and subjective evaluation scores for *image set* 1. Bolded values represent statistically significant correlation scores at $\alpha = 0.05$.

	PU2PSNR		PU2SSIM		HDRVDP2Q		WeberRMSE		UDQM		LR		LR with ref.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
complex	**0.83**	**0.90**	**0.94**	**0.87**	**1.00**	**0.97**	**0.83**	**0.89**	**0.60**	0.30	0.37	0.44	-0.14	0.25
handheld	**0.77**	**0.92**	**1.00**	**0.96**	**1.00**	**0.98**	**0.77**	**0.97**	**0.60**	0.16	0.26	0.08	-0.14	0.26
lolm	**1.00**	**0.96**	**0.83**	**0.91**	**1.00**	**1.00**	**1.00**	**0.95**	**0.94**	0.38	0.26	0.33	0.20	-0.14
losm	**0.89**	**0.59**	**0.61**	0.35	**0.54**	**0.97**	**0.84**	**0.59**	0.37	**0.96**	-0.54	-0.07	-0.43	-0.39
multiview	**0.94**	**0.96**	**0.71**	**0.84**	**1.00**	**0.96**	**0.94**	**0.92**	**0.49**	0.28	0.09	0.06	**0.49**	-0.02
nrm	0.43	**0.49**	**0.54**	**0.68**	**0.89**	**0.97**	0.26	0.12	**0.89**	**0.68**	0.14	0.37	-0.83	-0.73
occlusion	**0.66**	0.34	**0.60**	**0.82**	**0.94**	**0.98**	**0.66**	**0.65**	**0.66**	**0.73**	**0.49**	0.34	-0.03	0.40
solm	**0.66**	0.44	**0.66**	0.28	**1.00**	**1.00**	**0.60**	0.19	-0.03	**0.45**	0.37	**0.48**	0.31	-0.11
sosm	0.37	-0.05	0.35	0.08	**0.49**	0.33	0.37	0.04	**0.71**	**0.98**	**0.77**	**0.98**	0.37	-0.52
Average	**0.73**	**0.62**	**0.69**	**0.64**	**0.87**	**0.91**	**0.70**	**0.59**	**0.58**	**0.55**	0.24	0.33	-0.02	-0.11
Std. dev.	0.22	0.35	0.20	0.32	0.21	0.22	0.25	0.38	0.29	0.30	0.36	0.31	0.42	0.38

Table 6: Per scene Spearman's (ρ) and Pearson's (r) correlation coefficients for relation between objective metric predictions and subjective evaluation scores for *image set* 2. Bolded values represent statistically significant correlation scores at $\alpha = 0.05$.

	PU2PSNR		PU2SSIM		HDRVDP2Q		WeberRMSE		UDQM		LR		LR with ref.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
complex	**0.94**	**0.91**	**0.77**	**0.76**	**1.00**	**0.98**	**0.94**	**0.94**	0.26	0.42	0.09	0.14	0.14	0.30
handheld	**0.60**	**0.84**	**0.94**	**0.98**	**0.71**	**1.00**	**0.60**	**0.98**	**0.60**	0.35	0.37	0.35	-0.54	- 0.15
lolm	**1.00**	**0.98**	**0.94**	**0.93**	**0.94**	**0.98**	**0.94**	**0.97**	-0.03	0.43	-0.26	-0.03	0.03	-0.27
losm	0.43	**-0.06**	**0.54**	0.30	**0.49**	0.41	**0.70**	0.03	**0.77**	**0.99**	0.31	**0.98**	-0.03	-0.28
multiview	**0.94**	**0.97**	**0.94**	**0.91**	**0.94**	**0.82**	**0.83**	**0.78**	**0.49**	**0.51**	-0.03	-0.09	**0.66**	0.12
nrm	**0.60**	**0.57**	**0.89**	**0.85**	**0.89**	**0.98**	**0.60**	**0.51**	0.37	**0.76**	0.37	**0.64**	0.14	0.06
occlusion	**0.77**	**0.88**	**0.89**	**0.85**	**0.94**	**0.94**	**0.83**	**0.71**	0.09	**0.54**	-0.20	-0.38	-0.77	-0.61
solm	0.09	0.39	0.09	0.35	**0.77**	**0.88**	0.09	0.18	0.31	**0.74**	0.20	-0.27	-0.14	0.05
sosm	0.09	-0.30	0.03	0.01	-0.09	0.32	0.00	0.27	**0.49**	**0.98**	0.37	0.33	0.03	-0.72
Average	**0.61**	**0.58**	**0.67**	**0.66**	**0.73**	**0.81**	**0.61**	**0.60**	0.37	**0.63**	0.14	0.19	-0.05	-0.17
Std. dev.	0.35	0.47	0.37	0.35	0.35	0.26	0.35	0.37	0.25	0.24	0.25	0.44	0.41	0.34

Table 7: Spearman's (ρ) and Pearson's (r) correlation coefficients for relation between objective metric predictions and subjective evaluation scores for *image set* 3. Bolded values represent statistically significant correlation scores at $\alpha = 0.05$.

	PU2PSNR		PU2SSIM		HDRVDP2Q		WeberRMSE		UDQM		LR		LR with ref.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
complex	**1.00**	**0.96**	**0.66**	**0.78**	**0.94**	**0.94**	**1.00**	**0.95**	0.26	0.34	-0.31	-0.34	-0.09	0.23
handheld	**0.77**	**0.94**	**0.89**	**0.95**	**0.94**	**0.96**	**0.83**	**0.84**	0.26	**0.71**	0.20	0.19	**0.54**	-0.08
lolm	**0.83**	**0.98**	**0.54**	**0.53**	**0.89**	**0.94**	**0.83**	**0.92**	0.14	**0.53**	-0.43	-0.54	-0.14	**0.56**
losm	**0.77**	0.38	**0.54**	**0.52**	**0.77**	**0.77**	**0.66**	**0.75**	**0.77**	**0.79**	-0.09	0.44	0.09	**0.76**
multiview	**0.71**	**0.93**	**0.89**	**0.87**	**1.00**	**0.94**	**0.71**	**0.82**	0.31	0.24	-0.31	-0.22	**0.49**	**0.75**
nrm	**0.94**	**0.90**	0.43	**0.69**	**0.83**	**0.89**	**0.89**	**0.98**	**0.71**	**0.72**	-1.00	-0.81	-0.43	-0.66
occlusion	**0.77**	**0.91**	0.26	**0.47**	**0.77**	**0.90**	**0.83**	**0.89**	**0.66**	**0.71**	0.09	**0.47**	-0.77	-0.71
solm	**0.83**	0.44	**0.26**	0.35	**0.89**	**0.93**	**0.66**	**0.82**	**0.89**	**0.80**	-0.89	-0.90	-0.26	-0.74
sosm	0.37	**0.51**	0.26	0.22	**1.00**	**0.92**	0.37	0.10	**1.00**	**0.75**	-0.94	-0.81	-0.26	0.09
Average	**0.78**	**0.77**	**0.52**	**0.60**	**0.89**	**0.91**	**0.75**	**0.79**	**0.56**	**0.62**	-0.41	-0.28	-0.09	0.02
Std. dev.	0.18	0.25	0.25	0.24	0.09	0.05	0.18	0.27	0.32	0.21	0.45	0.54	0.42	0.61

1. **Sen et al.:** The results show that Sen et al.'s method outperforms other algorithms for most scenes. In the benchmark dataset there are scenes for which this algorithm is outperformed by other methods. Such scenes can be found in image set 4 (i.e. *lolm, nrm, occlusion, solm, sosm*). By analysis, in all of these scenes, ghost artefacts are generated in regions where the reference image is over-saturated (Figure 10), regardless of the motion type. In fact, visible artefacts are produced when the reference image is over-saturated even if there is no motion in that region of the scene. In such scenes, even the shortest exposure contains a

Table 8: Spearman's (ρ) and Pearson's (r) correlation coefficients for relation between objective metric predictions and subjective evaluation scores for *image set* 4. Bolded values represent statistically significant correlation scores at $\alpha = 0.05$.

	PU2PSNR		PU2SSIM		HDRVDP2Q		WeberRMSE		UDQM		LR		LR with ref.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
complex	**0.77**	**0.93**	**0.94**	**0.92**	**0.94**	**0.98**	**0.83**	**0.92**	**0.60**	0.28	0.14	-0.04	0.03	-0.30
handheld	**1.00**	**0.99**	**0.94**	**0.97**	**0.94**	**0.98**	**1.00**	**0.98**	-0.14	0.15	0.37	0.32	0.26	-0.06
lolm	**0.71**	**0.63**	**0.89**	**0.86**	**0.94**	**0.97**	**0.54**	**0.50**	-0.49	0.12	0.37	0.01	-0.49	-0.71
losm	**0.49**	0.26	**0.54**	**0.74**	**0.89**	**0.77**	0.26	0.38	-0.03	0.30	0.26	-0.63	0.43	**0.46**
multiview	**0.66**	**0.73**	**0.71**	**0.72**	**0.94**	**0.90**	**0.60**	**0.59**	**0.83**	**0.75**	-0.54	-0.63	-0.09	0.36
nrm	0.14	**0.51**	0.29	**0.55**	0.43	**0.92**	0.14	0.41	0.43	0.44	0.37	0.33	0.31	-0.39
occlusion	**0.77**	**0.92**	**0.77**	**0.89**	**0.89**	**1.00**	**0.89**	**0.91**	0.31	0.36	-0.49	0.11	-0.26	0.05
solm	0.37	-0.25	0.20	**0.59**	**0.60**	**0.57**	0.37	**0.61**	-0.09	-0.22	0.43	**0.56**	0.26	0.06
sosm	0.14	**0.77**	0.32	**0.75**	**0.94**	**0.98**	0.14	**0.75**	-0.03	0.22	0.14	**0.77**	0.26	**0.77**
Avg.	**0.56**	**0.61**	**0.62**	**0.78**	**0.83**	**0.90**	**0.53**	**0.67**	0.16	0.27	0.12	0.09	0.08	0.03
Std. dev.	0.30	0.40	0.29	0.15	0.19	0.14	0.32	0.23	0.41	0.26	0.37	0.48	0.30	0.46

Table 9: Spearman's (ρ) and Pearson's (r) correlation coefficients for relation between objective metric predictions and subjective evaluation scores. Values averaged across image sets from the computed per scene correlation scores displayed in Tables 5 - 8. Bolded values represent statistically significant correlation scores at $\alpha = 0.05$.

	PU2PSNR		PU2SSIM		HDRVDP2Q		WeberRMSE		UDQM		LR		LR with ref.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
complex	**0.89**	**0.92**	**0.83**	**0.83**	**0.97**	**0.97**	**0.90**	**0.93**	0.43	0.33	0.07	0.05	-0.01	0.12
handheld	**0.79**	**0.92**	**0.94**	**0.97**	**0.90**	**0.98**	**0.80**	**0.94**	0.33	0.34	0.30	0.23	0.03	-0.01
lolm	**0.89**	**0.89**	**0.80**	**0.81**	**0.94**	**0.97**	**0.83**	**0.84**	0.14	0.36	-0.01	-0.06	-0.10	-0.14
losm	**0.64**	0.29	**0.56**	**0.48**	**0.67**	**0.73**	**0.61**	0.43	**0.47**	**0.76**	-0.01	0.18	0.01	0.14
multiview	**0.81**	**0.90**	**0.81**	**0.83**	**0.97**	**0.90**	**0.77**	**0.78**	**0.53**	**0.45**	-0.20	-0.22	0.39	0.30
nrm	**0.53**	**0.62**	**0.54**	**0.69**	**0.76**	**0.94**	**0.47**	**0.51**	**0.60**	**0.65**	-0.03	0.13	-0.20	-0.43
occlusion	**0.74**	**0.76**	**0.63**	**0.76**	**0.89**	**0.95**	**0.80**	**0.79**	0.43	**0.58**	-0.03	0.13	-0.46	-0.22
solm	**0.49**	0.26	0.30	0.39	**0.81**	**0.84**	**0.43**	**0.45**	0.27	0.44	0.03	-0.03	0.04	-0.19
sosm	0.24	0.23	0.24	0.27	**0.59**	**0.64**	0.22	0.29	**0.54**	**0.73**	0.09	0.32	0.10	-0.10
Avg.	**0.67**	**0.64**	**0.63**	**0.67**	**0.83**	**0.88**	**0.65**	**0.66**	0.42	**0.52**	0.02	0.08	-0.02	-0.06
Std. dev.	0.21	0.30	0.24	0.24	0.14	0.12	0.23	0.24	0.15	0.17	0.13	0.16	0.23	0.22

very large over-saturated region. The reason for visible artefacts is that during patch-match, the method fails to find the corresponding pixels between the reference image and other images in over-saturated regions of the reference image.

2. **Hu et al.:** Hu et al.'s method is the second best performing method for most of the tested scenes. For similar reasons as mentioned previously for Sen et al.'s method, this method also produces artefacts in the regions where the reference image is over-saturated. However, the

artefacts produced are more severe than those produced by Sen et al. Furthermore, Hu et al.'s method also generates color artefacts in over-saturated regions of the reference, producing a very unnatural looking image.

3. **Photoshop:** It was found that Photoshop outperforms Photomatix for almost all scenes. For complex scenes, where there is plenty of motion and motion occurs in large region of the scene, Photoshop produces ghost artefacts. However, when the motion does not occupy large region of the scene (regardless of the motion type), the method does not produce ghost artefacts. It was also observed that the dynamic range of the moving content is usually reduced.

4. **Photomatix:** It was observed that Photomatix produces ghost artefacts for complex scenes, in particular for scenes with large object displacement (both for small and large objects.). Similar to the Photoshop method, Photomatix does not recover the dynamic range of the moving content. Observed loss of dynamic range of moving content in images produced by Photoshop and Photomatix suggests that these methods may use subset of exposures to handle ghosting.

5. **Silk et al.:** Silk et al. was found to have the lowest score from the evaluated algorithms (not considering the non-deghosted image). Even when this algorithm performed well in deghosting, the low score can be mainly contributed due to the 'washed out' faded trail generated by the algorithm usually in the region where there was object movement.

As reported in [11], Granados et al.'s method performs better than Sen et al.'s method for cluttered scenes with large object displacements. Because Granados et al.'s method is based on modelling noise distribution of color values measured by the camera, it is also expected to perform very well in terms of noise reduction. Therefore, we would expect the method to produce images that contain less noise than Sen et al.'s method (which has been observed to contain noise in low-luminance regions in some scenes). Granados et al.'s method might also produce potentially better results for image set 4 scenes (especially those with large object displacements) where Sen et al.'s method generated visible artefacts. Because we did not have a chance to evaluate Granados et al.'s method, it should be part of future evaluations of deghosting algorithms.

After subjective evaluation, a set of 6 suitable objective metrics were evaluated to test whether they can be used to assess HDR deghosting algorithms. To measure the success of objective quality metric results, Spearman and Pearson correlation coefficients between subjective and objective scores were computed by bootstrapping.

We found that existing full-reference image quality metric correlate well with subjective assessment of deghosting artefacts. The best performing metric, HDR-VDP-2, resulted in Pearson and Spearman correlation values between 0.67 and 0.95. Simpler metrics, such as PU2PSNR and PU2SSIM resulted in the correlation values between 0.39 and 0.90. The good performance of HDR-VDP-2 for

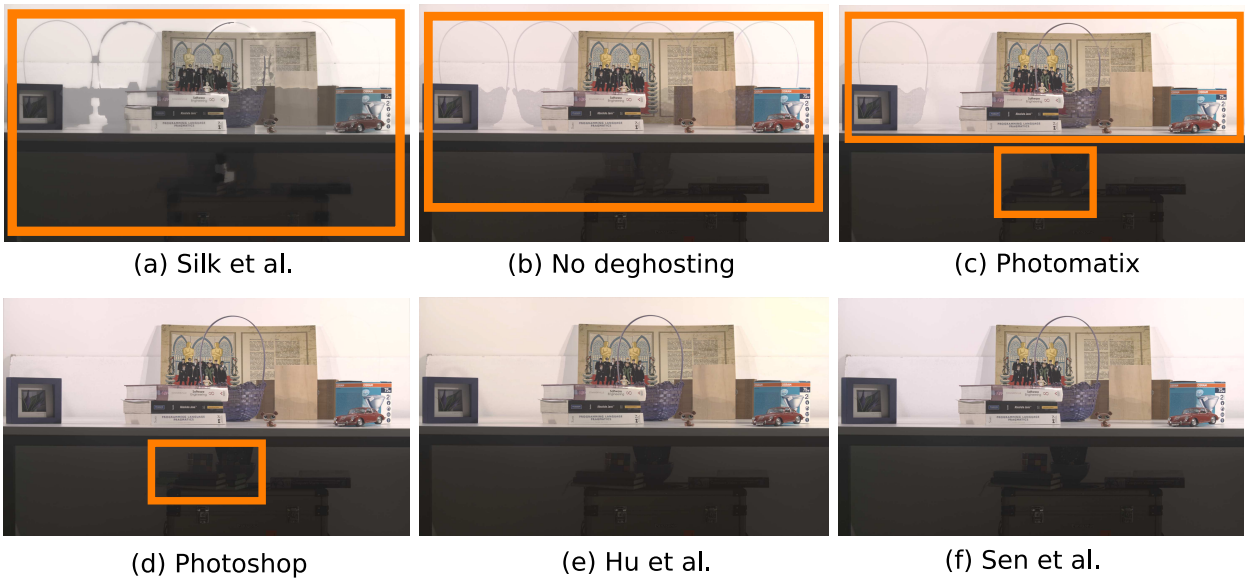


Figure 14: Outputs of images generated by (a) Silk et al.'s and (b) No deghosting (c) Photomatix (d) Photoshop (e) Sen et al.'s and (d) Hu et al.'s algorithms (image set 2, *occlusion* scene) ordered from worst to best performing method according to the subjective evaluation scores (2). UDQM metrics ranks 'No deghosting' method as best performing followed by Hu et al.'s, Photoshop, Photomatix, Sen et al.'s and Silk et al.'s methods. This contradicts the perceived subjective image quality.

our dataset can be attributed to the human visual system model around which the metric is built. The main limitation of the full-reference metrics is that they require reference ground truth images, which are typically not available for multi-exposure sequences. The non-reference metric, UDQM, is free from this limitation and it can be used in cases where full-reference metrics are not applicable, such as Khan et al.'s method, which removes moving objects. However, we found that UDQM correlates poorly with our subjective data. Figure 14 demonstrates an example where the results of UDQM have low correlation with perceived subjective image quality. One reason for low correlation of UDQM metric could be due to over-training and limited cross-validation used to validate this metric.

9. References

[1] Wen DD. High dynamic range charge-coupled device. US Patent 4873561; 1989.

[2] Street RA. High dynamic range segmented pixel sensor array. US Patent 5789737; 1998.

[3] Nayar S, Mitsunaga T. High dynamic range imaging: Spatially varying pixel exposures. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on; vol. 1. IEEE; 2000, p. 472–9.

[4] Zhao H, Shi B, Fernandez-Cull C, Yeung SK, Raskar R. Unbounded high dynamic range photography using a modulo camera. In: International Conference on Computational Photography (ICCP). 2015,.

[5] Manakov A, Restrepo JF, Klehm O, Hegedüs R, Eisemann E, Seidel HP, et al. A reconfigurable camera add-on for high dynamic range, multi-spectral, polarization, and light-field imaging. ACM Trans Graph (Proc SIGGRAPH 2013) 2013;32(4):47:1–47:14.

[6] Serrano A, Heide F, Gutierrez D, Wetzstein G, Masia B. Convolutional sparse coding for high dynamic range imaging. Computer Graphics Forum 2016;35(2).

[7] Hajisharif S, Unger J, Kronander J. HDR reconstruc-

tion for alternating gain (ISO) sensor readout. In: Proceedings of Eurographics Short Papers. 2014,.

[8] Debevec P, Malik J. Recovering high dynamic range radiance maps from photographs. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH '97). ACM Press/Addison-Wesley Publishing Co.; 1998, p. 369–78.

[9] Mitsunaga T, Nayar SK. Radiometric self calibration. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.; vol. 1. 1999, p. 380.

[10] Robertson MA, Borman S, Stevenson RL. Estimation-theoretic approach to dynamic range enhancement using multiple exposures. vol. 12. International Society for Optics and Photonics; 2003, p. 219–28.

[11] Granados M, Kim KI, Tompkin J, Theobalt C. Automatic noise modeling for ghost-free HDR reconstruction. ACM Transactions on Graphics (TOG) 2013;32(6):201:1–201:10. doi:10.1016/S0031-8914(53)80099-6.

[12] Jacobs K, Loscos C, Ward G. Automatic high-dynamic range image generation for dynamic scenes. IEEE Computer Graphics and Applications 2008;28(2):84–93.

[13] Heo YS, Lee KM, Lee SU, Moon Y, Cha J. Ghost-free high dynamic range imaging. In: ACCV (4)'10. 2010, p. 486–500.

[14] Sen P, Kalantari NK, Yaesoubi M, Darabi S, Goldman DB, Shechtman E. Robust patch-based HDR reconstruction of dynamic scenes. ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH Asia 2012) 2012;31(6):203:1–203:11.

[15] Hu J, Gallo O, Pulli K, Sun X. HDR deghosting: How to deal with saturation ? In: CVPR. 2013, p. 1063–170.

[16] Silk S, Lang J. Fast high dynamic range image deghosting for arbitrary scene motion. In: Proceedings of Graphics Interface 2012. Toronto, Ont., Canada, Canada: Canadian Information Processing Society. ISBN 978-1-4503-1420-6; 2012, p. 85–92.

[17] Tursun OT, Akyüz AO, Erdem A, Erdem E. The state of the art in HDR deghosting: A survey and evaluation. In: Computer Graphics Forum; vol. 34. Wiley Online Library; 2015, p. 683–707.

[18] Tursun OT, Akyüz AO, Erdem A, Erdem E. An objective deghosting quality metric for HDR images. In: Computer Graphics Forum; vol. 35. Wiley Online Library; 2016, p. 139–52.

[19] Hanhart P, Bernardo MV, Pereira M, Pinheiro AMG, Ebrahimi T. Benchmarking of objective quality metrics for HDR image quality assessment. EURASIP J Image and Video Processing 2015;2015:39.

[20] Korshunov P, Hanhart P, Richter T, Artusi A, Mantiuk R, Ebrahimi T. Subjective quality assessment database of HDR images compressed with JPEG XT. In: 7th International Workshop on Quality of Multimedia Experience (QoMEX). 2015,.

[21] Mantiuk R, Kim KJ, Rempel AG, Heidrich W. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans Graph (Proc SIGGRAPH) 2011;30(4):40.

[22] Narwaria M, Da Silva MP, Le Callet P. HDR-VQM: An objective quality measure for high dynamic range video. Signal Processing: Image Communication 2015;35:46–60.

[23] Karaduzovic K, Hasic J, Mantiuk R. Comparison of deghosting algorithms for multi-exposure high dynamic range imaging. In: Proceedings of the 29th Spring Conference on Computer Graphics. ACM; 2013, p. 21–8.

[24] Karaduzovic K, Hasic J, Mantiuk R. Expert evaluation of deghosting algorithms for multi-exposure high dynamic range imaging. In: In: Proc. of HDRi2014 - Second International Conference and SME Workshop on HDR imaging (2014). 2014,.

[25] Tursun OT, Akyüz AO, Erdem A, Erdem E. Evaluating deghosting algorithms for HDR images. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU). IEEE; 2014, p. 1275–8.

[26] Liu Y, Wang J, Cho S, Finkelstein A, Rusinkiewicz S. A no-reference metric for evaluating the quality of motion deblurring. ACM Trans Graph 2013;32(6):175:1–175:12.

[27] Aydin TO, Mantiuk R, Myszkowski K, Seidel HP. Dynamic range independent image quality assessment. ACM Trans Graph 2008;27(3):69:1–69:10.

[28] Karaduzovic-Hadziabdic K, Hasic J, Mantiuk R. Subjective and objective evaluation of multi-exposure high dynamic range image deghosting methods. In: EG

2016 - Short Papers. The Eurographics Association; 2016,doi:10.2312/egsh.20161007.

[29] Aydın TO, Mantiuk R, Seidel HP. Extending quality metrics to full luminance range images. In: Proceedings of SPIE, the International Society for Optical Engineering. Society of Photo-Optical Instrumentation Engineers; 2008, p. 68060B–10.

[30] Jacobs K, Loscos C, Ward G. Automatic high-dynamic range image generation for dynamic scenes. j-IEEE-CGA 2008;28(2):84–93.

[31] Pece F, jan Kautz . Bitmap movement detection: HDR for dynamic scenes. In: Visual Media Production (CVMP), 2010. Proceedings. IEEE Conference on; vol. 28. IEEE; 2010, p. 1–8.

[32] Gallo O, Gelfand N, Chen W, Tico M, Pulli K. Artifact-free high dynamic range imaging. IEEE International Conference on Computational Photography (ICCP) 2009;.

[33] Oh TH, Lee JY, Tai YW, Kweon IS. Robust high dynamic range imaging by rank minimization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 2015;37(6):1219–32.

[34] Bogoni L. Extending dynamic range of monochrome and color images through fusion. In: Proceedings of the 15th International Conference on Pattern Recognition, Volume 3. IEEE Computer Society; 2000, p. 3007–16.

[35] Zimmer H, Bruhn A, Weickert J. Comput Graph Forum 2011;(2):405–14.

[36] Grosch T. Fast and robust high dynamic range image generation with camera and object movement. In: Vision, Modeling and Visualization, RWTH Aachen. 2006, p. 277–84.

[37] Khan EA, Akyüz AO, Reinhard E. Ghost removal in high dynamic range images. In: Proceedings of the International Conference on Image Processing, ICIP 2006, October 8-11, Atlanta, Georgia, USA. 2006, p. 2005–8. doi:10.1109/ICIP.2006.312892.

[38] Robertson M, Borman S, Stevenson R. In: Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on; vol. 3. IEEE; 1999, p. 159–63.

[39] Simakov D, Caspi Y, Shectman E, Irani M. Summarizing visual data using bidirectional similarity. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. 2008, p. 1–8.

[40] Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 1981;24(6):381–95.

[41] Bay H, Ess A, Tuytelaars T, Gool LV. Speeded-up robust features (SURF). Computer Vision and Image Understanding 2008,;110(3):346–59.

[42] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 2004;13(4):600–12.

[43] Baker S, Scharstein D, Lewis J, Roth S, Black MJ, Szeliski R. A database and evaluation methodology for optical flow. International Journal of Computer Vision 2011;92:1–31. doi:10.1007/s11263-010-0390-2.

[44] https://doi.org/10.17863/cam.6881. 2016.

[45] Durand F, Dorsey J. Fast bilateral filtering for the display of high-dynamic-range images. ACM Trans Graph 2002;21(3):257–66.

[46] Mantiuk R, Krawczyk G, Mantiuk R, Seidel Hp. High dynamic range imaging pipeline: Perception-motivated representation of visual content. In: Electronic Imaging 2007. International Society for Optics and Photonics; 2007, p. 649212–.

[47] Tomaszewska A, Mantiuk R. Image registration for multi-exposure high dynamic range image acquisition. In: In: Proc. of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. 2007,.

[48] Silverstein D, Farrell J. Efficient method for paired comparison. Journal of Electronic Imaging 2001;10(2):394–8.

[49] Sheikh H, Sabir M, Bovik A. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. IEEE Transactions on Image Processing 2006;15(11):3440–51. doi:10.1109/TIP.2006.881959.