Comparison of single image HDR reconstruction methods — the caveats of quality assessment

Param Hanji
param.hanji@cl.cam.ac.uk
University of Cambridge
United Kingdom

Rafał K. Mantiuk
University of Cambridge
United Kingdom
rafal.mantiuk@cl.cam.ac.uk

Gabriel Eilertsen
Linköping University
Sweden
gabriel.eilertsen@liu.se

Saghi Hajisharif
Linköping University
Sweden
saghi.hajisharif@liu.se

Jonas Unger
Linköping University
Sweden
jonas.unger@liu.se

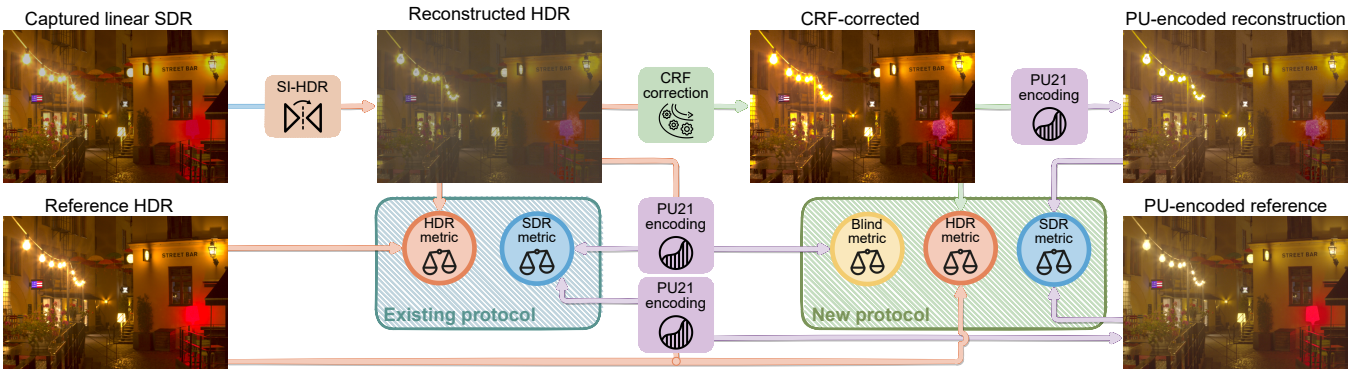

Figure 1: Existing protocols for evaluating single-image HDR reconstruction methods directly compare the reconstructed HDR images with the reference, as depicted by the blue shaded rectangle. This is unreliable due to large tone and color differences between the reference and reconstructed HDR images. We demonstrate that the accuracy of metrics can be much improved if we correct for camera-response-curve inversion errors before computing image quality using existing full-reference metrics as shown in the green shaded rectangle. Still, the metrics can detect only very large image differences in this task and conducting a controlled experiment is the recommended option.

ABSTRACT

As the problem of reconstructing high dynamic range (HDR) images from a single exposure has attracted much research effort, it is essential to provide a robust protocol and clear guidelines on how to evaluate and compare new methods. In this work, we compared six recent single image HDR reconstruction (SI-HDR) methods in a subjective image quality experiment on an HDR display. We found that only two methods produced results that are, on average, more preferred than the unprocessed single exposure images. When the same methods are evaluated using image quality metrics, as typically done in papers, the metric predictions correlate poorly with subjective quality scores. The main reason is a significant tone and color difference between the reference and reconstructed HDR images. To improve the predictions of image quality metrics, we propose correcting for the inaccuracies of the estimated camera response curve before computing quality values. We further analyze the sources of prediction noise when evaluating SI-HDR methods and demonstrate that existing metrics can reliably predict only large quality differences.

CCS CONCEPTS

• **General and reference** → **Reliability**; *Metrics*; *Evaluation*; • **Mathematics of computing** → *Bootstrapping*.

KEYWORDS

High dynamic range, inverse problems, image quality metrics

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08.
https://doi.org/10.1145/3528233.3530729

ACM Reference Format:
Param Hanji, Rafał K. Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. 2022. Comparison of single image HDR reconstruction methods — the caveats of quality assessment. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3528233.3530729

1 INTRODUCTION

The remarkable improvement in deep learning for image reconstruction problems over the last few years has produced numerous CNN-architectures that try to restore high dynamic range (HDR) information from a single standard dynamic range (SDR) image. However, unlike related problems such as super-resolution and denoising, the objectives of single image HDR reconstruction (SI-HDR) are not well defined. While most SI-HDR methods are trained to recover the physical HDR radiance values, we found that they perform poorly in this task. Despite this, many SI-HDR methods provide appreciable image enhancement that improves the quality of resulting images. These two tasks have radically different goals [Didyk et al. 2008], making it difficult to automatically compare competing methods with image quality metrics.

Another issue is that most methods are evaluated on datasets with similar distributions to training images, which provides an unfair advantage. To ensure fair evaluation, we collected a new SI-HDR dataset — a diverse set of HDR images with SDR counterparts, which were generated using a physically accurate camera model.

With a pairwise comparison experiment on a subset of the SI-HDR dataset, we demonstrate that some SI-HDR methods show promising image enhancement results, even though the reconstructions are far from HDR references. Surprisingly, two out of the six tested methods are more likely to decrease quality than improve it. The existing protocol for automatic evaluation conceals such results because most image quality metrics are more suitable for the assessment of the restoration task rather than the enhancement task. When comparing a reconstructed image with the reference HDR, metrics tend to be highly sensitive to shifts in tone and color due to incorrect camera response function (CRF) estimation [Eilertsen et al. 2021]. As a result, artifacts objectionable to humans, particularly in saturated regions of the input image, have less influence on the metric scores than slight changes in tone and color.

To remedy this, we propose an improved SI-HDR evaluation protocol that includes CRF correction, which significantly boosts the predictions of most metrics. We then compare an extensive list of quality metrics and identify those that highly correlate with the results of our experiment. The selected metrics reproduced the subjective rankings within predicted error bounds when run on the larger evaluation dataset. To estimate these bounds, we identified various sources of noise in the predictions and, for selected metrics, estimated the difference in metric scores necessary to claim an improvement above the noise level. Such differences are substantial for most metrics, suggesting that only large quality differences provide a reliable claim in the assessment of SI-HDR methods.

The main contributions of this work are: (a) new dataset for evaluation of SI-HDR reconstruction methods (Section 3); (b) report on the performance of those methods (Section 6); (c) better protocol for their assessment (Section 7.2); and (d) analysis of metric prediction error and caveats of using quality metrics for SI-HDR (Section 7.4). Code and data for this paper can be found at the project web page[1].

[1]Project web page: https://www.cl.cam.ac.uk/research/rainbow/projects/sihdr_benchmark

2 RELATED WORK

In this section, we discuss evaluation protocols to validate SI-HDR methods. For brevity, we do not review each method and refer to [Wang and Yoon 2021] for an overview of deep HDR imaging.

SI-HDR methods can be most reliably evaluated in a large-scale perceptual study on an HDR display. However, this is a tedious task, and most works instead rely on objective metrics to show the improvement of a newly proposed method. To measure the difference between the output of an SI-HDR method and the reference, metrics are usually applied directly or after a linear transformation, such as a global scaling with the image median or some other percentile. The comparison could use a dedicated HDR metric, such as HDR-VDP-2 or 3 [Mantiuk et al. 2011], or SDR metrics, such as PSNR, SSIM [Wang et al. 2004] or LPIPS [Zhang et al. 2018], on tone-mapped HDR values. Instead of tone mapping, which may strongly reduce contrast, it has been suggested to use the PU-transform [Mantiuk and Azimi 2021] or μ-law transform [Kalantari and Ramamoorthi 2017]. Some existing works directly apply SDR metrics on linear HDR images. This should be avoided since linear radiance values are perceptually non-uniform and incorrectly represent perceived differences [Mantiuk 2016].

There are several potential problems with existing protocols. First, test data, camera simulation, and comparison methods differ substantially between evaluations, making it impossible to compare results between papers. Second, as demonstrated in [Eilertsen et al. 2021], results are easily dominated by the ability of a particular method to invert the CRF (g in Eq. 1). This affects all pixels in an image and obscures the evaluation of recovered HDR information in, e.g., saturated regions of the image. Third, new SI-HDR reconstruction methods are often evaluated on images generated by a similar camera simulation as training images, which is likely to introduce a bias towards the proposed method since it could be better at inverting the CRF. Finally, comparisons are often performed using HDR metrics without properly calibrating the compared images.

A recent attempt at an independent and standardized evaluation was presented as a part of the HDR single and multi-frame imaging challenge in the New Trends in Image Restoration and Enhancement (NTIRE) CVPR workshop [Pérez-Pellitero et al. 2021]. The methods were evaluated on a new HDR dataset, using PSNR on linear and tone-mapped images. Their work overcomes some of the issues with diverse evaluation conditions, but problems around CRF dominance and metric calibration are yet to be addressed. In this work, we aim to address the remaining issues by (1) introducing an explicit CRF correction step to prevent CRF dominance, (2) identifying SI-HDR metrics based on a thorough subjective experiment.

3 SI-HDR DATASET

As most existing HDR datasets were used to train SI-HDR methods, we create a new dataset for the purpose of this evaluation. The new dataset needs to be sufficiently large and diverse in scene content and dynamic range. Most importantly, we need to guarantee that the data was not used to train any of the methods we compare. In this section, we describe our HDR data collection and SDR simulation, as well as the SI-HDR methods used in comparisons.

3.1 Multi-exposure stacks

183 HDR images were captured using a Canon 5D Mark III full-format DSLR camera. The scenes were selected to cover a wide range of image content and lighting conditions. Each scene is composed of up to 7 RAW exposures and merged into an HDR image using the estimator that accounted the photon noise [Hanji et al. 2020]. Next, a simple color correction was applied using a reference white point and all images were resized to 1920×1280 pixels.

3.2 Camera simulation

From the reference HDR scenes H represented in a linear RGB color space, we generated SDR images L using the camera simulation

$$L_i = q\left(\min\{1, g(e\,H_i + \eta(H_i))\}\right), \tag{1}$$

where i is the pixel index, e is the exposure, g is the CRF, η is camera noise, and q denotes quantization to the desired bit-depth. We used the popular normal approximation of the camera noise model [Foi et al. 2008; Hanji et al. 2020; Hasinoff et al. 2010],

$$\eta \sim \mathcal{N}(0, \alpha H_i + \beta), \tag{2}$$

where α and β are camera and ISO specific noise parameters that represent signal-dependent photon noise and signal-independent static noise respectively. For all images, we simulated the Canon EOS-1Ds with $e = 1/30$ sec and ISO 800.

We used the measured CRFs from the dataset in [Grossberg and Nayar 2003]. This collection contains a wide variety of CRFs, including measurements from analogue slide film which are not representative of modern digital cameras. Thus, we performed k-means clustering of CRFs with 5 clusters, and only used the cluster with a mean closest to the overall mean. This cluster contains 85 different CRFs with shapes that are representative of modern cameras. For each SDR simulation, we randomly picked a CRF from this set and individually selected a scene-specific exposure so that either 3% or 5% of pixels were saturated.

3.3 SI-HDR methods

We tested six deep-learning SI-HDR methods from the literature, DrTMO [Endo et al. 2017], HDR-CNN [Eilertsen et al. 2017], ExpandNet [Marnerides et al. 2018], HDR-GAN [Lee et al. 2018], SingleHDR [Liu et al. 2020], and Mask-HDR [Santos et al. 2020]. The selection includes frequently occurring methods in SI-HDR evaluations (DrTMO, HDR-CNN, and ExpandNet), as well as more recent ones (SingleHDR and Mask-HDR), including an adversarially trained network (HDR-GAN). While most methods directly predict linear HDR images as the output of the respective network, DrTMO and HDR-GAN predict exposure stacks.

Ideally, all methods would be retrained on a standardized dataset but unfortunately, such dataset is not available. Moreover, retraining existing methods typically results in worse performance due to differences in data preparation, choice of camera simulation, augmentation, and hyper-parameter tuning. Instead, we test the methods on a new unseen dataset using the models that have been trained by the respective authors. We include an interactive viewer in the supplementary to compare the quality of reconstructions produced by all methods over a variety of scenes.

4 CRF CORRECTION

As noted in [Eilertsen et al. 2021], SI-HDR methods typically fail to invert the CRF, making the resulting colors and tones different from the reference HDR images. The inaccuracies in CRF reconstruction are often so large, that the images are too different to be reliably compared in a subjective image quality experiment. We later show that such differences are also a major reason for the failure of image quality metrics. Hence, we developed a simple method that improves CRF inversion so that our evaluation can focus on the reconstruction of saturated pixels.

The goal of this step is to find a global smooth color mapping function, $\mathbb{R}^3 \rightarrow \mathbb{R}^3$, that corrects for the inaccurate CRF reconstruction, without interfering with the reconstruction of saturated pixels. To achieve this, we fit the coefficients of a polynomial basis function.

Performing the correction on log RGB values works well in most cases. However, in some situations this produced visible artifacts, especially for pixels with highly saturated colors. Instead, we first perform the optimization on the luma values P encoded using the PQ transfer function [Miller et al. 2013], $P_i = PQ(Y_i)$, where Y_i is the luminance of a pixel. As PQ encoding expects absolute luminance values, we normalize by the median and scale with a factor 500 prior to encoding. The value of 500 was found empirically to produce good results. Next, to find the luminance mapping between the reference HDR image P and the SI-HDR output $\hat{P}$, we solve:

$$\underset{w_1,\ldots,w_4}{\arg\min} \left\| \begin{bmatrix} \hat{P}_1^3 & \hat{P}_1^2 & \hat{P}_1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \hat{P}_N^3 & \hat{P}_N^2 & \hat{P}_N & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} P_1 \\ \vdots \\ P_N \end{bmatrix} \right\|, \tag{3}$$

where subscripts $1, \ldots, N$ denote pixel indices, and the four coefficients w_k describe a third-degree polynomial. Then, we optimize the mapping of the CIE chromatic coordinates $u'v'$:

$$\underset{w_{1,1},\ldots,w_{8,2}}{\arg\min} \left\| \begin{bmatrix} \hat{u}_1'^3 & \hat{v}_1'^3 & \hat{u}_1'^2 & \hat{v}_1'^2 & \hat{u}_1'\hat{v}_1' & \hat{u}_1' & \hat{v}_1' & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{u}_N'^3 & \hat{v}_N'^3 & \hat{u}_N'^2 & \hat{v}_N'^2 & \hat{u}_N'\hat{v}_N' & \hat{u}_N' & \hat{v}_N' & 1 \end{bmatrix} \begin{bmatrix} w_{1,1} & w_{1,2} \\ \vdots & \vdots \\ w_{8,1} & w_{8,2} \end{bmatrix} - \begin{bmatrix} u_1' & v_1' \\ \vdots & \vdots \\ u_N' & v_N' \end{bmatrix} \right\|, \tag{4}$$

where $\hat{u}_i'$ and $\hat{v}_i'$ are the chromaticity coordinates of the SI-HDR reconstruction for pixel i, while u_i and v_i are those of the reference HDR image. The 16 coefficients $w_{k,c}$ describe third-degree polynomials for each channel, with cross-channel dependencies.

In practice, we solve both optimization problems by minimizing the squared error,

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\arg\min} \, ||\mathbf{X}\mathbf{W} - \mathbf{Y}||_2^2 + \lambda ||\mathbf{W} - \mathbf{W}_0||_2^2, \tag{5}$$

where the first term corresponds to Eq. (3) or Eq. (4), i.e. where $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Y}$ are the pixel values of the reconstructed image, the polynomial weights, and the pixel values of the reference image, respectively. The formulation includes a Tikhonov regularization term, with strength λ. This is applied to penalize extreme values in the coefficient matrix $\hat{\mathbf{W}}$, which cause color distortions in the corrected image. The weight penalization is performed against the point $\mathbf{W}_0$, which is set with the weights that result in identity mapping of color/luminance values (all coefficients are 0 except for the linear term which is set to 1, i.e. $w_3 = 1$ in Eq. 3 and $w_{6,1} =$

1, $w_{7,2} = 1$ in Eq. 4). The solution of the minimization problem is:

$$\hat{\mathbf{W}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{Y} + \lambda\mathbf{W}_0), \tag{6}$$

where $\mathbf{I}$ is the identity matrix. Given that the matrix $\mathbf{X}$ is of size $N \times K$, we use $\lambda = 0.01N/K$ for the chrominance correction (Eq. 4), which makes the regularization invariant to image resolution and the degree of the polynomial. We use no regularization ($\lambda = 0$) for the luminance correction (Eq. 3).

5 SUBJECTIVE EVALUATION

The goal of the experiment was to obtain a possibly accurate measure of the gain (or loss) of visual quality that can be achieved by each SI-HDR method. We used the method of paired comparisons, as it was shown to provide higher sensitivity than direct rating methods [Perez-Ortiz et al. 2020].

Display. The images were shown on a 32" Asus ProArt PA32UCX 4k, HDR monitor. The monitor was set to use the PQ1000 response curve and was calibrated with a spectroradiometer (Specbos 1221). We noted that the displayed values had luminance 30% lower than expected (mapped PQ values vs. measurements). This was compensated for with our colorimetric calibration. The viewing distance was restricted to approximately 80 cm, resulting in an effective display resolution of 77 pixels per degree.

Images. We selected a subset of 27 HDR images from the SI-HDR dataset. The images, shown in Figure 2, were selected for a wide variety of content: portraits, nature, cities, indoor and outdoor, daylight and night time scenes. While we attempted to run the experiment directly on the results of SI-HDR methods, we found the images differ too much in tone and color to be compared. Instead, we used the images with the corrected CRF (Section 4) so that the participants judged the ability of the methods to reconstruct saturated pixels rather than their subjective opinion about tone and color of each image. The images were further cropped to a resolution of 1888×1280 so that two images could be shown side-by-side on our 4K monitor. Their exposure was adjusted so that the saturation point mapped to 100 cd/m^2 on the display. The range of luminance from 100 to 1000 cd/m^2 was used to reproduce the reconstructed pixel values. In total, the quality was assessed for 27 images (contents) × 2 exposures × (6 methods + HDR reference + SDR input) = 432 conditions.

Experimental procedure. The participants were shown a pair of images side by side and were asked to select "the image of higher quality — the one that better resembles a natural scene and contains fewer distortions" (the wording from the briefing form). The images were compared within blocks showing the same content and generated using different SI-HDR methods. Both SDR input and HDR reference images were included in the comparison. An active sampling method, ASAP [Mikhailiuk et al. 2021], was used to maximize the accuracy of the collected data. ASAP determines a batch of comparisons that maximizes the information gain and was shown to outperform heuristics, such as the Swiss chess system.

Participants. 14 volunteers participated in the experiment, each completing a full batch of comparisons scheduled by ASAP. Because ASAP ensures that each condition is compared at least once in each batch (builds a minimum-spanning tree), it means that each tested condition was compared at least 14 times with another condition. For reference, this is higher than 9 comparisons collected for TID2013 dataset [Ponomarenko et al. 2015] or 2-5 comparisons collected for BAAPS [Zhang et al. 2018].

Scaling and outlier rejection. The results of pairwise comparison was scaled under Thurstone's case V model into Just-Objectionable-Difference (JODs) using the *pwcmp* software [Perez-Ortiz and Mantiuk 2017]. A difference of 1 JOD unit means that 75% of observers select one condition over another. We used the same software to identify the potential outliers and removed the data for one observer. To account for measurement error, we report bootstrapped confidence intervals in all plots.

6 RESULTS: SI-HDR BENCHMARK

The results aggregated across all contents and exposures are shown in the left of Figure 3. For individual results, we refer to Figure 3 in the supplementary. The results are shifted with respect to the input SDR image, so that positive JOD values indicate improvement in quality and negative values indicate degradation of quality. The right plot indicates the percentage of images that were assessed to be better, same or worse than the SDR input image. Per image JOD of more than 0.5 or less than -0.5 was used to distinguish between better and worse. The results show the rather disappointing performance of SI-HDR methods — many methods more often degrade the quality of an input image rather than improve it. This level of performance shows the difficulty of the task. It is worth noting that high fail rate is not uncommon for similar challenging tasks, such as image inpainting, and such techniques are still useful, but they require manual screening of the results.

We encourage the reader to inspect individual images included in the project web page. As an example of successful reconstruction, we highlight image 177. We show the reconstructed portion of the image for all the methods in the top row of Figure 4. Even though the reconstructed colors are quite different from the HDR reference, all the reconstructions improve on the SDR input image with clipped colors. In the middle row of Figure 4 we show image 052, for which ExpandNet, HDR-CNN and MaskHDR show moderate improvement by boosting the saturated region in a convincing manner, but DrTMO and HDR-GAN introduce objectionable artifacts. Finally, in the bottom row of Figure 4 we show image 123, for which all methods failed to reconstruct large saturated regions. The JOD scores in Figure 1 in the supplementary reflect the plausibility of hallucinated regions.

7 QUALITY METRICS FOR SI-HDR

To identify an automatic evaluation procedure that correlates with our results from Section 6, we start by looking at the reliability of existing SI-HDR evaluation methods. Then, we propose an improved protocol, which accounts for the random nature of quality assessment and the inaccuracies of quality metrics.

We identified a range of metrics, both full-reference, and no-reference, listed in Table 1. All the metrics either supported direct comparison of HDR images (HDR-VDP-2 and -3, FovVideoVDP) or were suitably adopted. For this, we used the PU21 transform [Mantiuk and Azimi 2021], μ-transform [Kalantari and Ramamoorthi 2017] or a global tone-mapping operator [Reinhard et al. 2002].

Figure 2: The subset of HDR images from SI-HDR dataset used for the subjective evaluation, tone mapped with a global operator [Mantiuk et al. 2008] for visualization. We selected a wide variety of content covering nature, portraits, cities, indoor and outdoor, dalight and night scenes.

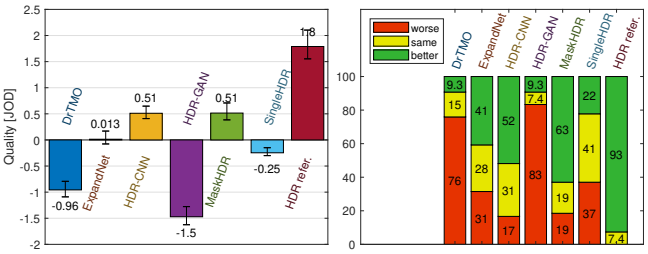

Figure 3: Preference of the SI-HDR method results. Left: The bars indicate the preference in JOD units, relative to the source SDR image. Negative values indicate that on average the method produced less preferable result than the non-processed source image. Right: The bars indicate the percentage of images in which the method produced better, same or worse image than the input SDR image.

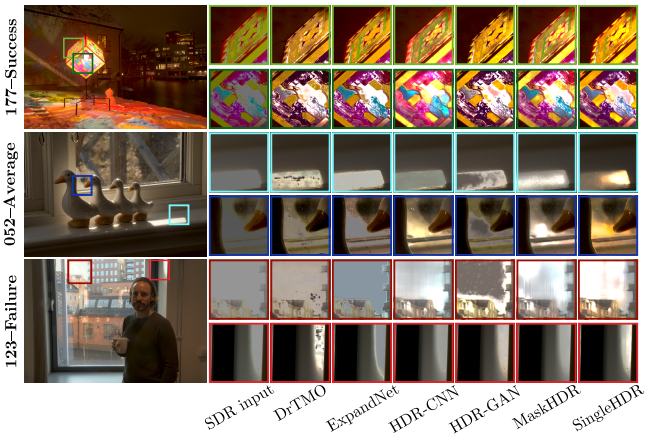

Figure 4: Reconstructions by various methods of selected scenes from the subjective data. Please check the interactive HTML viewer in the supplementary to assess the quality of reconstructions for other scenes.

Table 1: List of quality metrics used in our evaluation.

Metric	Reference required	HDR metric	Details
PSNR	✓	✗	Widely used ratio to measure noise relative to the signal in log units
SSIM [Wang et al. 2004]	✓	✗	Popular quality measure that relies on block-wise correlations
MS-SSIM [Wang et al. 2003]	✓	✗	Multi-scale version of SSIM
VIF [Sheikh and Bovik 2006]	✓	✗	Employs natural scene statistics (NSS) models to measure enhancement
FSIM [Zhang et al. 2011]	✓	✗	Low-level feature similarity index based on the visual system
VSI [Zhang et al. 2014]	✓	✗	Weighted average of local quality maps guided by visual saliency
LPIPS [Zhang et al. 2018]	✓	✗	Perceptual similarity metric across CNN architectures
HDR-VDP-2 and 3 [Mantiuk et al. 2011]	✓	✓	Low-level vision model, works on HDR images
FovVideoVDP [Mantiuk et al. 2021]	✓	✓	Low-level vision model for images, video and foveation
BRISQUE [Mittal et al. 2012]	✗	✗	Support vector regression trained on IQA dataset
NIQE [Mittal et al. 2013]	✗	✗	Distance between NSS-based features to those from a database
PIQE [Venkatanath N et al. 2015]	✗	✗	Averaged block-wise distortion estimation
NIMA [Talebi and Milanfar 2018]	✗	✗	Object recognition CNNs re-purposed as a blind metric

PU21 maps linear RGB color values into approximately perceptually uniform units (accounting for glare and contrast sensitivity). The μ-transform offers a logarithmic scaling with empirically selected parameters. The images passed to all metrics were scaled in absolute units representing physical color emitted from our display. For metrics that required display geometry (HDR-VDP-2, HDR-VDP-3, FovVideoVDP), we matched the configuration of our experiment (77 ppd). Finally, we clamped values above 1000 cd/m^2 as they exceed the peak luminance of our display.

Quality can be assessed both with respect to the HDR reference image (reconstruction task) and relative to the input SDR image (enhancement task). To test whether the latter is more suitable, we included VIF, a metric capable of assessing image enhancement. We append -SDR to the metric name when SDR input is used as a reference and -HDR otherwise.

7.1 Existing validation protocols

To test the suitability of popular metrics, we followed the standard protocol used to evaluate image quality metrics [Ponomarenko et al. 2015], and computed the rank-order (Spearman) correlations between metric predictions and JOD values for individual conditions. This gave us very low correlation values, especially for no-reference metrics. The highest correlation was 0.47 for PU21-PSNR. When we computed the correlations for the images after CRF correction (Section 4), the highest correlation increased to 0.55 for HDR-VDP-3, which is still too low to provide meaningful predictions. Our initial assessment hints that none of the existing metrics is suitable for predicting the quality of images reconstructed by SI-HDR methods. Please check Table 1 in the supplementary for the complete list of correlations.

However, this evaluation protocol does not reflect the standard quantitative evaluation for SI-HDR methods. SI-HDR methods are typically compared by averaging metric scores across multiple images and comparing the means. When we compute the correlations of averaged metric scores with subjective JOD values for all the scenes (plotted in Figure 3), they are much higher. We suspect that this is due to the poor performance of quality metrics in assessing the absolute level of impairment. Each metric introduces per-content bias, which reduces correlation with the subjective data. However, such biases cancel out when the quality values are averaged across content, and the resulting scores correlate much better with the subjective data.

To ensure a fair assessment for each metric given our data, we generated 2000 bootstrap samples for each estimated correlation by randomizing (sampling with replacement) both the participants and the selection of images [Mooney and Robert D. Duval 1993]. Each sample involved independently scaling the JOD values using a subset of data. Our bootstrapping simulated 2000 outcomes of the experiment to capture the variance we can expect due to measurement noise. To compute the average correlations, we use the unbiased estimator given by [Olkin and Pratt 1958].

The correlations for averaged metric scores, shown in Figure 5a, are higher especially for the no-reference PU21-PIQE metric (ρ = 0.83) and PU21-VSI (ρ = 0.78). However, these are not metrics typically used to evaluate SI-HDR methods. Interestingly, the correlations for metrics used in SI-HDR papers (PU21-PSNR, PU21-SSIM, PU21-LPIPS, and HDR-VDP-2, and 3) are all below 0.64. We can thus conclude that the metrics used in the SI-HDR papers were not sensitive enough to differentiate the compared methods and, therefore, could not provide evidence for the improvement of the proposed methods. This observation explains why our participants disagreed with quantitative results in recent SI-HDR papers.

7.2 Improved protocol for SI-HDR evaluation

First, we confirmed the observation from [Eilertsen et al. 2021] that the inaccuracy in the CRF reconstruction is the major reason for the poor metric predictions. When metric scores are computed on images with the CRF correction, we see a major gain in metric performance (compare Figure 5a and Figure 5b). This is because even small differences in the CRF, sometimes unnoticeable to the human eye, result in large metric errors as it affects all pixels in the image. With the CRF correction, the correlation of PU21-PSNR

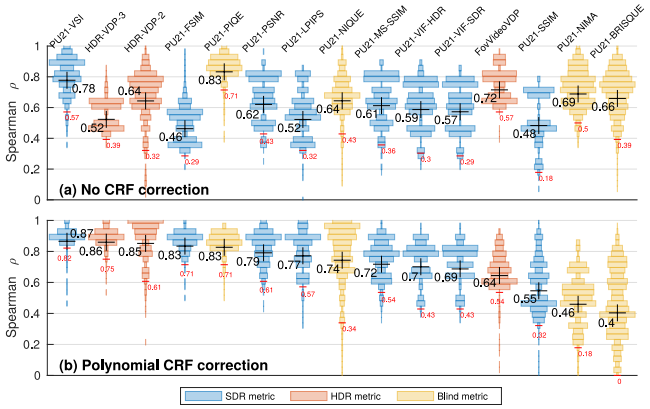

Figure 5: Bootstrapped distributions of correlation coefficients for all metrics compared (a) before and (b) after polynomial correction. "+" denotes unbiased mean correlation [Olkin and Pratt 1958] and "–" (in red) denotes the 5th percentile (an estimate of the bad-case performance).

jumps from 0.62 to 0.79 and a similar trend is seen for other metrics. We also attempted computing metric values only on saturated areas by copying non-saturated pixels from the reference to the test image, but we did not notice any improvement in metric predictions. The existing metrics do not seem to cope well with localized distortions. We intentionally show the distributions of correlation coefficients in Figure 5 to stress out that any metric evaluation study comes with a high degree of uncertainty.

We also compared three adaptations of the SDR metrics to HDR images: PU21 transform [Mantiuk and Azimi 2021], μ-transform [Kalantari and Ramamoorthi 2017] and Reinhard et al. global tone-mapping operator [Reinhard et al. 2002]. We did not find evidence for significant difference between those methods in our application. The full analysis can be found in the supplementary. In the following analysis, we will use PU21 due to its stronger perceptual basis.

Our recommendation for the new protocol is to report the results for PU21-PSNR, because of its simplicity and good performance, for PU-VSI and HDR-VDP-3, since they are likely to perform well, and for PU21-PIQE as this is a well performing no-reference metric. CRF correction must be performed for all metrics except PU21-PIQE, which performs well without such correction. Our results motivate the need for an HDR no-reference metric invariant to the type of distortion. A candidate, proposed by [Banterle et al. 2020], is restricted to a single kind of distortion. We do not recommend using PU21-SSIM and PU21-NIMA.

7.3 Validation of quality metrics

Here we investigate whether the selected metrics used with the new, improved protocol with CRF correction yield good enough predictions to measure the performance improvement of SI-HDR methods. To evaluate the methods, we used the remaining 156 images from our dataset (excluding images depicted in Figure 2 since they were part of the experiment). If the metrics were accurate, the results shown in Figure 6 should match the subjective results from Figure 3. While most metrics capture the trend, many incorrectly

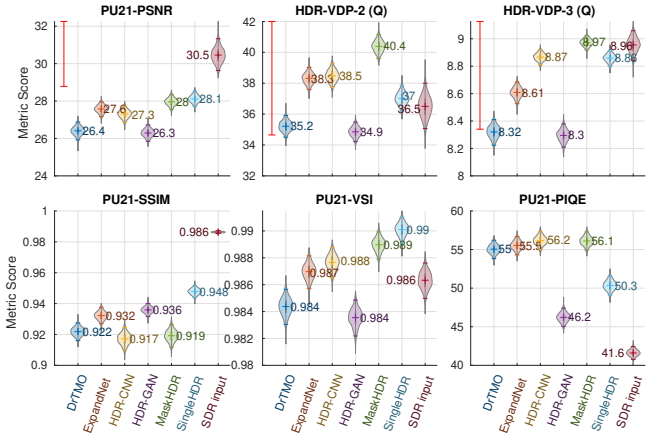

Figure 6: Ranking bootstrapped distributions for SI-HDR methods on the validation dataset. For each distribution, small dashes denote 95% confidence intervals, while the red errors bars show the minimum measurable increment in quality for selected metrics (described in Section 7.4).

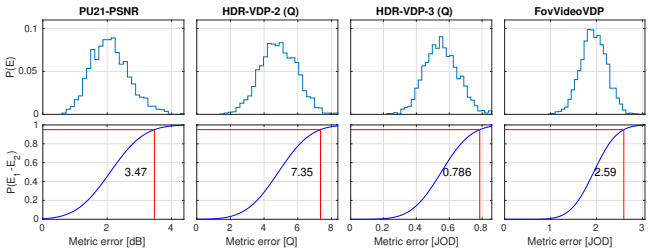

Figure 7: Estimated prediction error for 4 selected metrics. Similar to Figure 5, the distributions were obtained by bootstrapping the experiment results over the subset of 27 images. The improvement in quality metric values reported in many papers falls below the expected accuracy of each metric.

predicted the quality of the input SDR image. Additionally, contrary to the subjective data, a few metrics assigned higher quality to the SingleHDR method. Relying on metrics would thus result in an incorrect ranking of methods with a high likelihood.

7.4 Minimum measurable increment of quality

Correlation coefficients are difficult to interpret. Even for the metric with the highest expected correlation (PU21-VSI $\rho = 0.87$), it is impossible to say whether this value is good enough to evaluate SI-HDR methods. Instead, we want to determine the smallest difference in metric scores that can tell us with confidence (at $\alpha = 0.05$) that one method is better than the other. To do that, we need to account for all sources of measurement error: (a) due to the selection of images; (b) due to the measurement error in subjective experiment results (error bars in Figure 3); and (c) due to the inaccuracy of a metric. Hidden in the metric error (c) is an inherent difficulty related to the ill-posed nature of SI-HDR task. In saturated regions, hallucinated details may be different to the reference, yet perfectly plausible. To obtain a comprehensive measure, we bootstrapped the RMSE values. For each bootstrap sample, we found a linear mapping from JOD values to metric predictions and then computed RMSE in terms of the metric error.

Such an analysis can be applied only to the metrics that produce quality values that are (approximately) linearly related to the perceived magnitude of quality (MOS or JOD). For example, the strongly non-linear relation between SSIM values and MOS means that a small difference in SSIM values has a different impact on perceived quality depending on the absolute SSIM values. Therefore, for this analysis, we selected only PU21-PSNR, HDR-VDP-2, HDR-VDP-3, and FovVideoVDP, which are all designed to be well correlated with the perceived magnitude of distortion.

The distributions of the prediction errors are shown in the top row of Figure 7. These distributions could be interpreted as an expected metric error, E, with respect to the subjective scores. To use a metric to compare two methods, we are interested in the error for the difference between two quality metric scores. Since the distributions are normal (Kolmogorov-Smirnov test, at $\alpha = 0.05$ significance level), we estimated the mean and standard deviation for the bootstrapped samples. We plotted the cumulative distribution for the difference in the error estimates (also normal with the same mean and standard deviation multiplied by $\sqrt{2}$) in the bottom of Figure 7. The plots show that we need a PU21-PSNR difference of at least 3.5 dB to be confident that the method with higher PSNR is on average better (at 5% chance of making an error). The required minimum differences are also high for other metrics. Since the improvement in quality reported in most papers falls below these amounts, it casts doubts on the reliability of the evaluation performed solely with objective quality metrics.

8 CONCLUSIONS

The evidence from our subjective quality assessment experiment indicates that the overall progress in single-image HDR reconstruction is less impressive than reported in the papers. The results indicate little improvement in the quality of SI-HDR results over the last five years. In fact, a few of the methods were more likely to degrade image quality than improve it.

We believe that a critical reason for the lack of visible progress is the wrong use of quality metrics and insufficient subjective evaluation. We found that metrics commonly used to evaluate SI-HDR methods have low correlation with subjective scores ($\rho = 0.62$ for PU21-PSNR). Instead, we propose to compute metric scores on images that have been corrected for errors in the CRF inversion. We found PU21-VSI and HDR-VDP-3 to be the best performing metrics, but also recommend PU21-PSNR and the no-reference metric PU21-PIQE, which does not require CRF correction.

Finally, we demonstrate that even the best metrics computed on CRF-corrected images introduce a substantial prediction error. Using non-parametric statistics and our subjective results, we estimated the minimum improvement in quality scores required to find the difference between two SI-HDR methods at $\alpha = 0.05$ to be 3.5 dB for PU21-PSNR. Since most of the papers report much smaller improvement in quality, there is a high risk that the reported improvement is due to random errors rather than actual gains in method performance. While we still recommend computing metric scores, we also suggest running a subjective quality assessment on

20-30 randomly selected images from an independent dataset. Our findings are specific for SI-HDR methods but we believe that our analysis can be extended to other reconstruction problems, such as super-resolution or denoising.

ACKNOWLEDGMENTS

We would like to thank anonymous reviewers and Aamir Mustafa for their comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement N° 725253–EyeCode).

REFERENCES

Francesco Banterle, Alessandro Artusi, Alejandro Moreo, and Fabio Carrara. 2020. NoR-VDPNet: A No-Reference High Dynamic Range Quality Metric Trained on HDR-VDP 2. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 126–130. https://doi.org/10.1109/ICIP40778.2020.9191202

P Didyk, R Mantiuk, M Hein, and H.P. Seidel. 2008. Enhancement of Bright Video Features for HDR Displays. *Computer Graphics Forum* 27, 4 (jun 2008), 1265–1274. https://doi.org/10.1111/j.1467-8659.2008.01265.x

Gabriel Eilertsen, Saghi Hajisharif, Param Hanji, Apostolia Tsirikoglou, Rafal K Mantiuk, and Jonas Unger. 2021. How to cheat with metrics in single-image HDR reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 3998–4007.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM transactions on graphics (TOG)* 36, 6 (2017), 1–15.

Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. 2017. Deep reverse tone mapping. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–10.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. 2008. Practical Poissonian-Gaussian Noise Modeling and Fitting for Single-Image Raw-Data. *IEEE Transactions on Image Processing* 17, 10 (2008), 1737–1754. https://doi.org/10.1109/TIP.2008.2001399

Michael D Grossberg and Shree K Nayar. 2003. What is the space of camera response functions?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, Vol. 2. IEEE, II–602.

Param Hanji, Fangcheng Zhong, and Rafał K. Mantiuk. 2020. Noise-Aware Merging of High Dynamic Range Image Stacks without Camera Calibration. In *Advances in Image Manipulation (ECCV workshop)*. Springer, 376–391.

S. W. Hasinoff, F. Durand, and W. T. Freeman. 2010. Noise-Optimal Capture for High Dynamic Range Photography. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. 553–560.

Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* 36, 4, Article 144 (jul 2017), 12 pages. https://doi.org/10.1145/3072959.3073609

Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. 2018. Deep recursive HDRi: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 596–611.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2020. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 1651–1660.

Rafał Mantiuk, Scott Daly, and Louis Kerofsky. 2008. Display Adaptive Tone Mapping. In *ACM SIGGRAPH 2008 Papers* (Los Angeles, California) *(SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, Article 68, 10 pages. https://doi.org/10.1145/1399504.1360667

Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)* 30, 4 (2011), 1–14.

Rafal K. Mantiuk. 2016. Practicalities of predicting quality of high dynamic range images and video. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 904–908. https://doi.org/10.1109/ICIP.2016.7532488

Rafał K. Mantiuk and Maryam Azimi. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Picture Coding Symposium*. 1–5.

Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP: A Visible Difference Predictor for Wide Field-of-View Video. *ACM Trans. Graph.* 40, 4, Article 49 (jul 2021), 19 pages. https://doi.org/10.1145/3450626.3459831

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. 2018. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 37–49.

Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K. Mantiuk. 2021. Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 2559–2566. https://doi.org/10.1109/ICPR48806.2021.9412676

Scott Miller, Mahdi Nezamabadi, and Scott Daly. 2013. Perceptual signal coding for more efficient usage of bit codes. *SMPTE Motion Imaging Journal* 122, 4 (2013), 52–59.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Christopher Z. Mooney and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage.

Ingram Olkin and John W. Pratt. 1958. Unbiased Estimation of Certain Correlation Coefficients. *The Annals of Mathematical Statistics* 29, 1 (1958), 201 – 211. https://doi.org/10.1214/aoms/1177706717

Maria Perez-Ortiz and Rafal K. Mantiuk. 2017. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint* (dec 2017). arXiv:1712.03686 http://arxiv.org/abs/1712.03686

Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafal K. Mantiuk. 2020. From Pairwise Comparisons and Rating to a Unified Quality Scale. *IEEE Transactions on Image Processing* 29 (2020), 1139–1151. https://doi.org/10.1109/TIP.2019.2936103

Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. 2021. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 691–700.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30 (jan 2015), 57–77. https://doi.org/10.1016/j.image.2014.10.009

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. 2002. Photographic Tone Reproduction for Digital Images. *ACM Trans. Graph.* 21, 3 (jul 2002), 267–276. https://doi.org/10.1145/566654.566575

Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. 2020. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 80–1. https://doi.org/10.1145/3386569.3392403

H.R. Sheikh and A.C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (2006), 430–444. https://doi.org/10.1109/TIP.2005.859378

Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899

Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. 2015. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*. 1–6. https://doi.org/10.1109/NCC.2015.7084843

Lin Wang and Kuk-Jin Yoon. 2021. Deep Learning for HDR Imaging: State-of-the-Art and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

Z. Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Vol. 2. 1398–1402 Vol.2. https://doi.org/10.1109/ACSSC.2003.1292216

Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing* 23, 10 (2014), 4270–4281. https://doi.org/10.1109/TIP.2014.2346028

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386. https://doi.org/10.1109/TIP.2011.2109730

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. 586–595. https://doi.org/10.1109/CVPR.2018.00068