Volume xx (200y), Number z, pp. 1–9

Reconfigurable Snapshot HDR Imaging Using Coded Masks and Inception Network

M. Alghamdi, Q. Fu, A. Thabet & W. Heidrich

King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

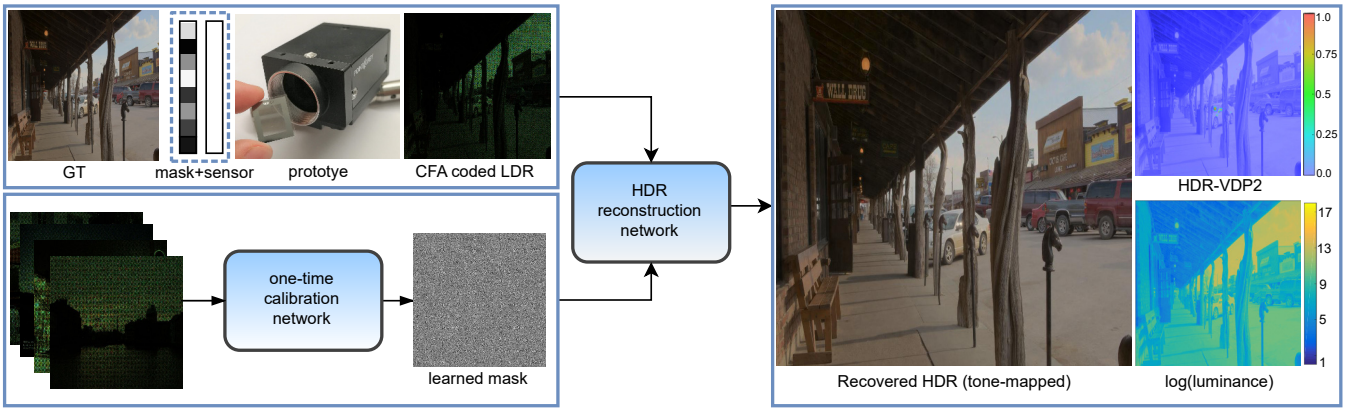

Figure 1: *Overview of the proposed HDR imaging system. In hardware a binary optical mask is placed near the image sensor to achieve spatially varying exposures (top left). In reconstruction we first devise a casual one-time calibration network to accurately estimate the mask (bottom left), and then input the raw noisy color filter array (CFA) image and estimated mask to our HDR reconstruction network to obtain the final HDR image. The HDR-VDP-2 visibility probability maps for our result (right-top), blue indicates unperceivable differences, which means that our system can recover a high quality HDR image from a single raw LDR image.*

Abstract

High Dynamic Range (HDR) image acquisition from a single image capture, also known as snapshot HDR imaging, is challenging because the bit depths of camera sensors are far from sufficient to cover the full dynamic range of the scene. Existing HDR techniques focus either on algorithmic reconstruction or hardware modification to extend the dynamic range. In this paper we propose a joint design for snapshot HDR imaging by devising a spatially-varying modulation mask in the hardware as well as building an inception network to reconstruct the HDR image. We achieve a reconfigurable HDR camera design that does not require custom sensors, and instead can be reconfigured between HDR and conventional mode with very simple calibration steps. We demonstrate that the proposed hardware-software solution offers a flexible yet robust way to modulating per-pixel exposures, and the network requires little knowledge of the hardware to faithfully reconstruct the HDR image. Comparison results show that our method outperforms state of the art in terms of visual perception quality.

CCS Concepts

• ***Computing methodologies*** → ***Computational photography;***

1. Introduction

The human visual system can sense up to 20 F-stops of luminance contrast with minimal eye adaption [BADC17]. Modern image sensor technology is incapable of matching this performance and reproducing the full dynamic range of natural scenes within a single exposure. The challenge of single shot (or snapshot) High Dynamic Range (HDR) imaging arises from the tremendous gap between

the huge intensity range in natural scenes and the very limited bit depths that modern camera sensors can offer.

To overcome this limitation, many computational imaging techniques have been developed via co-designing the sensor architecture and post-processing algorithms for HDR image acquisition [RHD*10]. These methods can be categorized into three distinct approaches. The most common way is to capture a sequence of low dynamic range (LDR) images with different exposures and fuse them into an HDR image [DM97, MPMP95]. Modern cameras and mobile devices can easily afford successive image capture, making this method capable of producing decent HDR images for static scenes. However, when either the scene is dynamic or the camera shakes during capture, the resulting images can suffer from ghosting artifacts. The second approach is to utilize multiple sensors to simultaneously capture differently exposed LDR images by, for example, splitting the light to multiple sensors with a beam-splitter [MMP*07, TKTS11, KGBU13]. This sophisticated approach is expensive and needs additional rigorous calibration. The third approach is to capture a single LDR image with a per-pixel or per-scanline coded exposure. Reconstruction algorithms are applied later to create HDR images [NM00, NB03, SHG*16]. This type of computational camera can be achieved by using a per-pixel coded exposures in the sensor architecture [KKM14] or by mounting an optical mask onto an off-the-shelf camera sensor.

In this work, we propose a computational imaging solution to single shot HDR imaging by minimal modifications of the camera to implement per-pixel exposure, as well as a deep learning algorithm based on the inception network to reconstruct HDR images, as shown in Figure 1. Specifically we explore a variant of the spatially modulated HDR camera design that does not require a custom sensor, and can be incorporated into any existing camera, be it a smartphone, a machine vision camera, or a digital SLR. We envision in particular a scenario where the camera can be reconfigured on the fly into an HDR mode with the introduction of an optical element into the optical stack of the camera. To realize these desired properties, we propose a mask that is not attached directly to the surface of the image sensor as in the case of Assorted Pixels [NM00, NB03], but is instead placed at a small standoff distance in front of the sensor. To support dynamic hardware reconfiguration, we explore rapid calibration of the mask, as well as snapshot HDR image reconstruction.

The main technical contributions of the paper are (1) an easy-to-implement modulation method that requires minimum hardware modification and a simple self-calibration technique; (2) a new HDR reconstruction algorithm built upon inception network that decodes decent HDR images from the raw Bayer data. We demonstrate both in simulation and by a prototype that the combination of hardware encoding and software decoding leads to a simple yet efficient HDR image acquisition system.

2. Related Work

2.1. Inverse Tone Mapping

An HDR image can be created from an LDR image by an Inverse Tone Mapping Operator (iTMO), which can be achieved by approximately inverting the tone mapping algorithms [RSSF02, BLD*07, BLDC08]. To enhance the LDR image brightness light source density estimation is used to generate an expand map. Rempel et al. [RTS*07] proposes an online iTMO for video streams that can be integrated directly into HDR displays. They start by contrast stretching of the input image, and then enhancing the brightness in the saturated regions by blurring the image with a Gaussian kernel while preserving strong edges using edge stopping functions. Kovaleski et al. [KO14] speed up this method by applying a faster edge preserving filter. Others [HYB13, HYDB14, MSG17] achieve iTMO by employing more heuristics in the process. Because iTMO in general is an ill-posed problem, existing methods are not able to properly compensate missing information due to overexposure, underexposure and quantization, especially in large saturated regions.

2.2. Coded Exposure

Another category of HDR image acquisition is to encode per-pixel exposures. Nayar et al. [NM00, NN02] initialize the idea of per-pixel exposures by assigning different (but fixed) exposures to neighboring pixels on the sensor. This can be done by placing an optical mask with spatially varying transmittance in the optical path, or by a spatial light modulator to control the exposure of each pixel over space and time [NB03]. In HDR reconstruction, Nayar and Mitsunaga [NM00] apply aggregation and interpolation algorithms to reconstruct HDR from simulated coded LDR using a fixed four-exposure mask. The mask results in aliasing artifacts and spatial resolution loss in the reconstruction. To overcome the aliasing problem, Schöberl et al. [SBS*12] introduce a binary random mask instead of a fixed one, and a frequency selective extrapolation iterative algorithm in reconstruction. Aguerrebere et al. [AAG*14] propose an iterative patch-based algorithm for jointly denoising and interpolation, but artifacts arise in very high dynamic range scenes. Serrano et al. [SHG*16] recently present a global iterative solution using convolutional sparse coding (CSC) to reconstruct HDR images from a coded LDR image to show superior performance.

2.3. HDR with Deep Learning

The past few years have seen the rapid development of deep learning methods applied in HDR image reconstruction. Such methods either learn to blend several LDR images optimally, or directly estimate HDR from a single LDR image. Frame merging methods learn, through a CNN, how to blend LDR images with different exposures into a single HDR; these methods are specially useful under dynamic scene conditions. Kalantari and Ramamoorthi [KR17] propose a blending deep learning algorithm, where the input images are first aligned using optical flow. In contrast, Wu et al. [WXTT18] present an encoder-decoder network that learns image alignment as well as blending. More recently deep learning has boosted interest on iTMO algorithms. Endo et al. [EKM17] take a single LDR image, and learn to create a series of up-exposed and under-exposed images, by learning 2 encoder-decoder networks. Similar work is proposed by [LAK18a]. The main drawback of these methods is the requirement of further networks for every image generated. This problem is addressed by Lee et al. [LAK18b], where the authors train two Generative Adversarial Networks (GANs), each of which generates new images with a

relative exposure (either lower or higher) to the input image. Eilertsen et al. [EKD*17] present an encoder decoder network to directly reconstruct an HDR image from a single LDR image. The authors assume the input to have small amounts of saturated areas. Like other iTMO methods, this work is not able to properly compensate missing information due to saturation, under exposure and quantization, and should be understood as detail "hallucination" instead of reconstruction for input images with large saturated regions.

3. Methodology

3.1. Problem Definition

In our HDR system, we propose to place an optical mask into the optical path in close proximity to the image sensor. The propagation of light from the mask to the sensor leads to a grayscale modulation pattern on the captured image. In a color camera, a Bayer Color Filter Array (CFA) samples the radiance into three color channels. The camera sensor then converts the photons impinging on the image plane over a specific exposure time into electrons, and quantizes the voltage values into digital numbers (DNs). This process can be expressed as

$$\mathbf{y} = g\left(f\left(\mathbf{B}\Phi\mathbf{x}\Delta t\right)\right), \tag{1}$$

$\mathbf{y} \in \mathbb{R}^M$ is the captured raw LDR image with a total of M pixels , $\mathbf{x} \in \mathbb{R}^M$ is the irradiance on the sensor , $\Phi \in \mathbb{R}^{M\times M}$ is a diagonal matrix for the per-pixel spatially-varying modulation mask, $\mathbf{B} \in \mathbb{R}^{M\times M}$ is a diagonal matrix that represents the Bayer filter, and Δt is the exposure time. f is a nonlinear function that includes the camera response and quantization. g is a noise function that accounts for different noise types in modern CCD and CMOS sensors, including photon shot noise, dark current, fixed pattern noise, quantisation noise and other nonlinearities [KW14]. Note that different from the image formation model in Serrano et al.'s work [SHG*16], we employ a more realistic noise model than the simple additive Gaussian noise. This model is more suitable when the scene spans a wide dynamic range.

The goal of our method is to invert the image formation in Eq. (1) from a coded LDR image to recover the HDR information in the scene. To solve the inverse problem, we devise a deep learning mechanism based on the inception network. We reconstruct decent HDR images in a 2-step process. First we calibrate our system once by directly estimating the mask Φ from a series of casually captured images. We then feed both the mask and coded LDR image to our deep network. The network jointly performs denoising, demosaicking, and HDR reconstruction, and outputs the final HDR result.

With the help of the modulation mask, the dynamic range can be improved compared with the conventional LDR camera. As indicated in [NM00, SHG*16], the dynamic range of a digital camera is $20\log\left(I_{\max}/I_{\min}\right)$, where $I_{\max}$ and $I_{\min}$ are the maximum and minimum DNs respectively. Suppose the maximum and minimum transmittance of the mask in Φ are $e_{\max}$ and $e_{\min}$, the dynamic range is consequently expressed as $20\log\left(\left(I_{\max}\cdot e_{\max}\right)/\left(I_{\min}\cdot e_{\min}\right)\right)$. For a typical 8-bit sensor, $I_{\max} = 255$ and $I_{\min} = 1$, so the dynamic range is $48dB$. If the transmittance is quantized such that $e_{\max} = 64e_{\min}$, the dynamic range becomes $84dB$, which is a significant improvement.

3.2. Coded Mask

For easy implementation of a grayscale mask, we choose to place a random binary optical mask at a short distance (typically 1-2mm) in front of the sensor. Note that we don't optimize the distance, but just mount our mask on the cover glass that is usually present in front of the sensor. Light propagation from the mask to the sensor results in a blurred version of the binary mask. The actual statistics depends on both the mask and propagation distance.

Similar ideas of inserting a mask in the optical path have been expored in the light field literature [VRA*07]. The optical mask is essentially a modulation in the 4D ray space, and the captured image is then an intensity image of the modulated rays integrated over the angular dimensions. With significant depth variations in the scene, the modulation would vary with distances. In our case, as well as in similar mask based works [NM00, SBS*12, AAG*14, SHG*16], the depth of field is assumed to be small, so the modulation is constant and becomes an element-wise product with a flat image. In practical implementations as shown later, the distance between the optical mask and the sensor is typically 1-2 mm, while the amount of defocus is usually in the range of micrometers. Therefore, this effect is negligible.

Prior works have explored a host of different fixed mask types, e.g. fixed four-exposure, non-regular, Gaussian, uniform, etc. [NM00, SBS*12, AAG*14, SHG*16] for spatially-varying exposure. Therefore, we first investigate the effects of different types of masks. Considering the mask setup in our configuration, we investigate four types of mask statistics: binary, continuous uniform, high frequency (HF) Gaussian and low frequency (LF) Gaussian, which correspond to the distances from close to far away from the sensor respectively. In Figure 2 we show the effects of these masks on reconstruction results.

With a binary mask, if the value of $e_{\min}$ is small, the captured LDR image suffers from noise in under-exposed areas. In the very left column, we show an example mask, which has $e_{\max}/e_{\min} = 8$. Little information could be retrieved particularly in the saturated regions. With a HF Gaussian mask, we have more information modulated. The mask throughput can also be higher than the binary mask. The reconstruction is better as well. However, there might be relatively large overexposed or underexposed areas where little HDR information can be retrieved. For the continuous uniform mask we can have both a good throughput and high dynamic range extension if we set, e.g. $e_{\max}/e_{\min} = 64$. We can guarantee that each pixel is surrounded by a variety of different exposures, which minimizes the probability of having large areas of over-exposure or under-exposure. So it is preferable to use such a mask. Compared with conventional masks with fixed numbers of exposures (e.g. four-exposure in [NM00]), it is also more practical to mount a uniform mask, because a slight misalignment would lead to a significant deviation for the four-exposure mask, but negligible change for the uniform mask. A LF Gaussian mask is applicable in cases where the distance between the mask and the sensor plane is relatively large and is not controllable by the users. The performance may drop compared with the other masks, but considerate gain in dynamic range can be still be achieved.

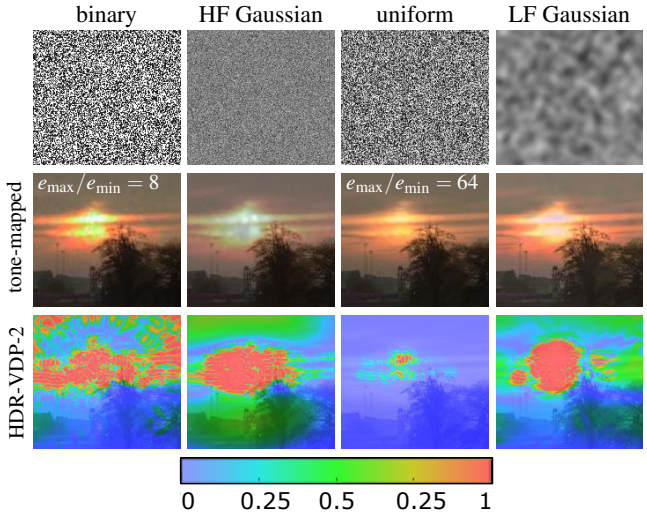

Figure 2: *Coded exposure mask comparison. The uniform mask performs better than the rest in terms of visual perception. HDR-VDP-2 [MKRH11] estimates the visual difference between the reconstructed HDR and the ground truth. Smaller values indicate higher reconstruction quality.*

3.3. HDR Reconstruction

Throughout our work, we take a modified inception block by removing pooling layers as suggested by Kim et al. [KHC17b] as the building block of our networks as shown in Figure 3. This modified block is able to extract the scale invariant features by using multi-size kernels. The utilization of 1×1 convolution leads to less number of parameters than other general CNNs with the same depth. This block has been successfully applied in different image restoration and labeling tasks like compression artifacts reduction, semantic segmentation and skin detection [KHC17a, KHC17b].

We now present our HDR pipeline and details of the calibration and HDR reconstruction networks as follows.

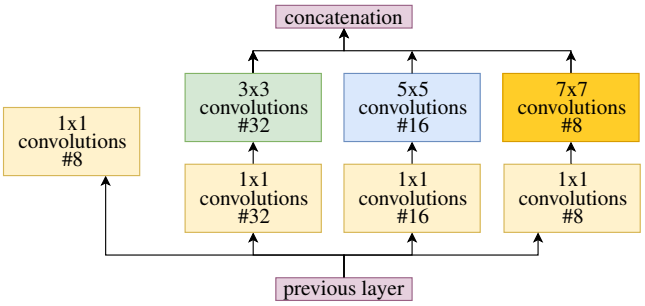

Figure 3: *Modified inception block. The number of filters is indicated in each convolution layer in our implementation.*

3.3.1. Calibration

Since one can design the optical mask, a natural way is to do a rigorous calibration in the laboratory for later HDR reconstruction. However, we opt not to include such a calibration step in our system because such a process is quite tedious and sensible to environmental changes. If the mask is detached and mounted again, another rigorous calibration has to be repeated. To avoid such problems, we instead propose a deep learning approach for casual self-calibration. Our deep network just takes a few coded images of arbitrary natural scenes and outputs an accurate estimate of the mask. As with our training process, we simulate coded HDR images used for calibration, using the camera models detailed in section 3.4. Refer to section 2 in the Supplementary Material for detailed calibratrion information. It is particularly more interesting for our goal of reconfigurable HDR imaging where the mask can be mounted to any sensors for multiple times without rigorous re-calibration in the lab.

Given N raw CFA coded exposure images with size of $H \times W \times 3N$ (3 channels for each CFA coded image), our network outputs an estimate of the mask $\hat{\Phi}$ of size $H \times W \times 1$ as well as the demosaicked green channels $\hat{G}$ of size $H \times W \times N$. Here we include also the demosaicked green channel of the images to guide the calibration network to converge to better results. This is because cameras usually have higher responses on their green channel and thus making it a great candidate as a regularizer for our network to learn the mask. We found that adding such a regularization speeds up our network convergence and improves the mask estimation accuracy significantly.Our network consists of 2 convolutional layers followed by 8 inception blocks and followed by a convolutional layer, the network pipeline is shown in fig. 1 in the supplementary.

We jointly minimize the ℓ_2-loss of both $\hat{\Phi}$ and $\hat{G}$ defined in Eq.(2). We add a regularizing parameter μ to control the effect of our green channel demosaicking. Since we simulate the calibration images, both Φ and G are available for supervision.

$$\ell\left(\Phi,\hat{\Phi},G,\hat{G}\right) = \frac{1}{WH}\sum_{j=1}^{H}\sum_{i=1}^{W}\left|\Phi(i,j)-\hat{\Phi}(i,j)\right|^2 + \mu\left|G(i,j)-\hat{G}(i,j)\right|^2. \tag{2}$$

During the training phase we found that larger N produce a lower loss, however require a longer training time for the same number of epochs due to the higher number of network parameters and increased memory usage. We chose N=15, as it results in a low calibration error. Also we found during the training phase that fixing the exposure time for most of the images as a baseline such that most of the images are properly exposed (the number of saturated and under-exposed pixels are minimized). While 3 or 4 of which are randomly chosen to have 1 or 2 stops more exposure, produced a lower training loss specifically for the regions with low values in the mask. Please refer to the supplementary material for more information about the network architecture, required number of images, testing scenarios and results.

3.3.2. HDR Reconstruction Network

We use the calibrated mask and the raw LDR image as inputs to our HDR reconstruction network to generate the final HDR image.

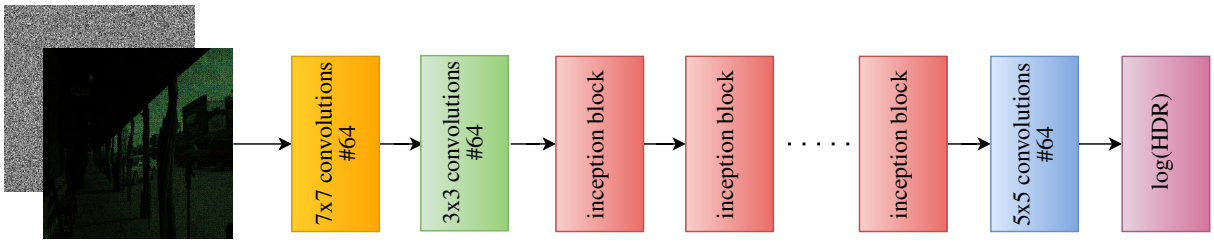

Figure 4: *Our HDR reconstruction network consists of 2 convolutional layers followed by 6 inception blocks and 1 output layer.*

The pipeline is shown in Figure 4. The goal of the network is to demosaick, denoise, and linearize the LDR image, as well as to recover information lost in the quantization process. Our network consists of two inputs, two convolution layers, followed by six inception modules, and one convolutional output layer. We follow the method proposed in [EKD*17] and estimate the logarithm of HDR values instead of linear ones; high luminance linear HDR values will bias the loss function and cause deterioration in low luminance estimates.

We want here to give more attention to the loss layer of the deep network, instead of using the default ℓ_2 norm option. The ℓ_2 norm has many appealing characteristics, for example, it is a convex and differentiable function which is useful optimization criteria, and it is the maximum likelihood estimator for independent and identically distributed Gaussian noise. It is, however, generally admitted that ℓ_2 and, as a result, the PSNR, do not align well with human image perception [ZZMZ12], ℓ_2 does not detect the sophisticated features of the human visual system.

Recently, several works show how internal activations of deep convolutional networks, trained on classical image classification problems, are remarkably effective as a feature space for a significantly more extensive range of tasks. For instance, features from the VGG network [SZ14] were used for image regression problems such as [JAFF16, DB16, CK17, GEB16, JAFF16]. Such image restoration solutions used the quantified distance in the VGG representation space as a perceptual loss [DB16, JAFF16]. We present a similar approach here. Specifically, given a network V (e.g. VGG), we compute the loss between each learned image $\hat{x}$ and its corresponding ground truth x as follows. First, we utilize V as a feature extractor, by selecting the output of N layers. Finally, we define our new loss as in Eq.(3), where x is the logarithm of the ground truth HDR image, $\hat{x}$ is the output of the last layer of our HDR reconstruction network , x^i and $\hat{x}^i \in \mathbb{R}^{H^i \times W^i \times C^i}$ are outputs from the i-th layer of network V.

$$L_{VGG}(x,\hat{x}) = \sum_{i=1}^{i=N} \frac{1}{H^i \times W^i \times C^i} \omega_i \left\| x^i - \hat{x}^i \right\|_1, \qquad (3)$$

where ω_i is used for weighting the contribution of features extracted from layer i to the loss function. Our final loss function can be defined as

$$L_{HDR}(x,\hat{x}) = \lambda L_1(x,\hat{x}) + (1-\lambda) L_{VGG}(x,\hat{x}), \qquad (4)$$

where $L_1(x,\hat{x})$ is the ℓ_1 norm, and λ is a trade-off coefficient.

3.4. Training

To train our pipeline, we combine HDR images from several datasets: DML-HDR [BDAPN14], HDR Photographic Survey [Fai07], Funt-HDR [FS10], Natural Scenes HDR [XDCW02], LiU HDRv [KGBU13], sIBL Archive [Con12]. In total, we have 1170 HDR images, with their LDR counterparts. We refer to this dataset as HDR Core. Images in HDR Core are not coded with our mask, thus, we also propose a method to integrate a coded pattern to LDR images in order to obtain a coded HDR image. Additionally, we pre-train our network with synthesized HDR data from MIT Places dataset [ZLX*14], and 19K HD 16 bit images collected from 200 HD YouTube videos, where we select a frame every 5 seconds. Refer to Supplementary Material for more details about training.

To construct a more realistic simulation for the coded raw image, we collected parameters of 67 different Sony CMOS and CCD cameras. Specifically we looked for sensor type (CMOS/CCD), pixel size, quantum efficiency of each channel, bit depth, full well capacity and dark noise. To synthesize a coded raw image we randomly select one set of camera parameters to set the corresponding simulation parameters, then modulate the HDR image (or simulated HDR) with a random coded optical mask. Next, the input is transformed, in sequence, into the number of photons, the number of electrons, voltages, and lastly to a digital signal using the method proposed in [KW14].

We train our reconstruction network on three stages, one for each of our datasets. We provide full training details of each stage in the Supplementary Material. We use Keras API on top of TensorFlow [C*15]. For learning we use the ADAM optimizer [KB14]. We pre-train on synthesized samples of the places dataset for 3M back-propagation steps, then we fine-tune on synthesized images from YouTube HD dataset for another 2M back-propagation steps. Finally we fine-tune on HDR-Core dataset for 2.5M steps of back-propagation, with a mini-batch size of 4, taking about 6 days on two GeForce GTX 1080 GPUs. To reconstruct an HDR image from a 1000×1500 raw LDR image, it takes $438ms$ on one same GPU.

4. Results

We present the performance of the proposed method with both simulation and experimental results in this section. We compare our results with existing algorithms in terms of the perceptual visual metric HDR-VDP-2 [MKRH11]. A prototype has been built to demonstrate the capabilities of our method with real world images.

4.1. Simulation Results

To demonstrate the overall performance, we compare the performance of our network in the test image set with eight iTMO algorithms [AFR*07, HYB13, HYDB14, KTC12, KO14, MSG17, EKD*17] as well as two representative coded exposure techniques [NM00, SHG*16]. We report HDR-VDP-2 [MKRH11] as an evaluation metric. Rather than mathematical differences, HDR-VDP-2 (version 2.2.1) computes the visual difference based on human perception, which is more appropriate for HDR images. For all the algorithms we calculate the average Qscore of the HDR-VDP-2 results as follows,

$$Q = \frac{1}{F \cdot O} \sum_{f=1}^{F} \sum_{o=1}^{O} w_f \log \left(\frac{1}{I} \sum_{i=1}^{I} D^2 [f,o](i) + \varepsilon \right), \qquad (5)$$

where i is the pixel index, D is the noise-normalized difference between the f-th spatial frequency and the o-th orientation of the steerable pyramid for both reference and test images, w_f is the vector of per-band pooling weights, which is determined by maximizing correlations with subjective opinion scores. I, F and O are the total numbers of pixels,frequencies and orientations respectively, ε is a small constant to avoid singularities when D is close to 0. The results are in Table 1.

Table 1: *Comparison of average HDR-VDP2 Qscore.*

	Method	Qscore
1	Akyuz et al. [AFR*07]	33.49 ± 12.11
2	Huo et al. [HYB13]	20.84 ± 7.26
3	Huo et al. [HYDB14]	32.13 ± 10.01
4	Kovaleski et al. [KO14]	32.16 ± 8.42
5	Kuo et al. [KTC12]	30.28 ± 6.84
6	Landis et al. [Lan02]	20.15 ± 3.99
7	Masia et al. [MSG17]	32.71 ± 10.65
8	Eilertsen et al. [EKD*17]	41.17 ± 13.29
9	Nayar et al. [NM00]	43.54 ± 11.08
10	Serrano et al. [SHG*16]	35.46 ± 15.24
11	**Ours**	**50.17 ± 12.42**

Figure 5 shows 2 arbitrary examples of the HDR reconstructed results along with the HDR-VDP2 visual differences between our reconstruction method and the ground-truth HDR images. Clearly our method can successfully construct a reliable HDR image. The inset images show the details in both the reconstructed images and ground-truth images. $\log_2$(luminance) shows the large dynamic range that our method can retrieve. As the results show, our algorithm outperforms the existing methods. In addition, Figure 6 shows 3 examples of the HDR-VDP-2 results for 8 iTMO methods compared with ours. Because of the extra information from the mask, our results outperform others both visually and qualitatively. Our mask-based approach aims to reconstruct the real scene being imaged. Conversely, all other methods are actually hallucinating missing details in large saturated areas.

One main challenge for our approach is scenes with a very large dynamic range see the example in fig. 7 where the dynamic range of the scene is more than 30 stop, when we have either large saturated or under-exposed regions where no information captured by the camera our network start to hallucinate the missing information zoom-in inside the fire region of our reconstruction to see this effect.

4.2. Experimental Results

We build a prototype HDR camera based on the proposed spatially-varying exposure strategy. We attach a binary optical mask onto a Point Grey GS3-U3-14S5C-C color camera. The mask is designed with random uniform distribution of either 0 or 1, and fabricated on a fused silica wafer by depositing a textured thin Chromium (Cr) film on one face. The feature size of the mask element is $12.9\mu m$, which is twice the size of the camera pixel. We place the mask onto the protective glass of the sensor, with the mask surface facing the sensor plane. This results in a distance of approximately $1.8mm$ between the mask and the sensor, which introduces the required per-pixel exposure modulation for the encoded LDR image. In all the results shown below, we use a $50mm$ Canon lens EF50mm f/1.8 II operating at F8 to capture the coded LDR images.

5. Conclusion

Due to the tremendous gap between limited camera bit depths and broad dynamic range in real world, high quality HDR image acquisition from a single shot requires taking into account both per-pixel exposure modulation and robust image reconstruction. We have presented a joint design strategy that adds a spatially-varying modulation mask onto the camera sensor and a corresponding algorithm based on inception network to faithfully reconstruct high quality HDR images. The advantage of the proposed method is that only a minimum modification to the camera is required. For any camera sensor that is accessible, a binary mask works as an easy-to-mount add-on element. Rigorous laboratory calibration is removed from the pipeline. Instead, a casual self-calibration process that takes a series of arbitrary natural images with different exposures as inputs to our proposed calibration network to learn the encoding mask. Our HDR reconstruction network subsequently outputs high quality HDR images from the encoded LDR image and the learned mask. The visual perception quality of the results has been demonstrated to outperform state-of-the-art methods.

However, there are some limitations in the current work. First, our method does not yet allow for arbitrary encoding mask. The statistics of the mask does have a major impact on the final reconstruction quality. When a low spatial frequency mask is used, large blobs (either bright or dark) could degrade the performance of the result. An optimized mask will be studied in the future. Second, although simplified, the calibration step is still present in the pipeline. A network that can learn both the mask and the decoded HDR image from a single encoded LDR image will be carried out in future work to completely eliminate the calibration step.

References

[AAG*14] AGUERREBERE C., ALMANSA A., GOUSSEAU Y., DELON J., MUSE P.: Single shot high dynamic range imaging using piecewise linear estimators. In *IEEE International Conference on Computational Photography (ICCP)* (2014), IEEE, pp. 1–10. 2, 3

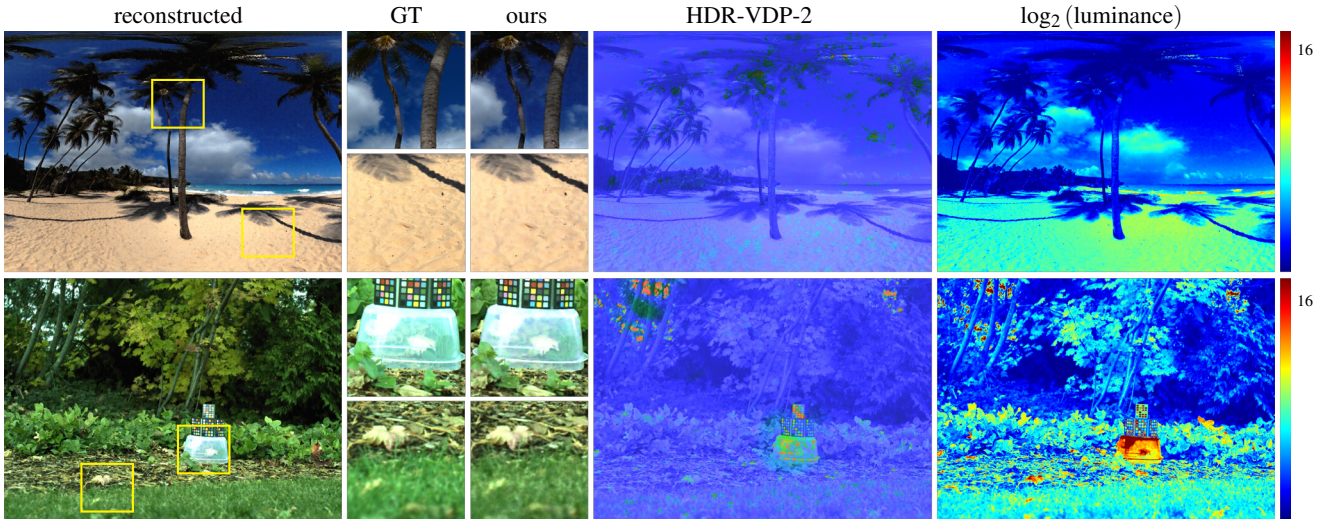

Figure 5: *Simulation results for two scenes. Left is the tone-mapped HDR images for our results. Zoom-in images show the comparison with the ground truth images. The HDR-VDP2 results show that the overall visual differences are suppressed. The $\log_2$ (luminance) maps indicate our method can achieve more than 16 stops in dynamic range.*

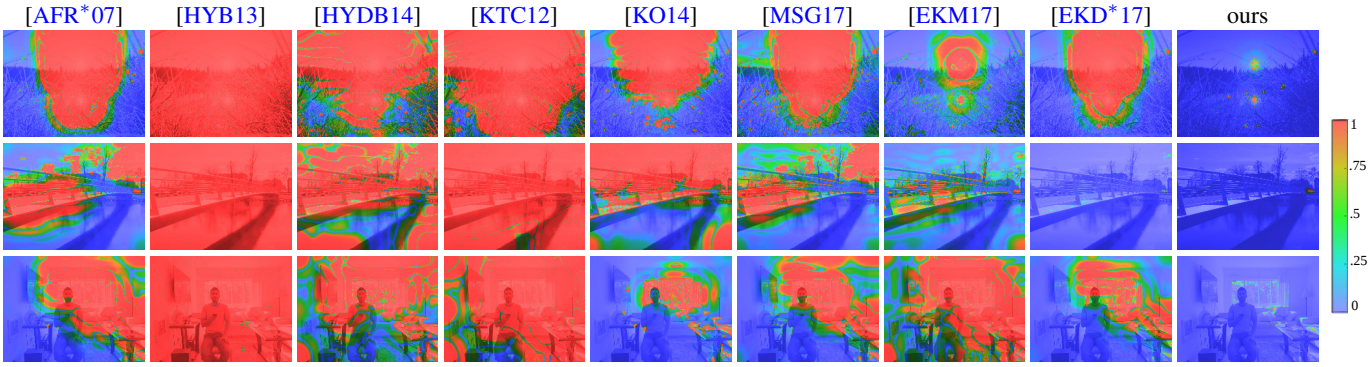

Figure 6: *Simulation HDR reconstruction comparison with conventional iTMO algorithm using HDR-VDP-2. Note how our approach reconstructs real scene information. Conversely, iTMO algorithms are "inventing" information in saturated and underexposed regions instead of reconstructing what was actually there. Smaller HDR-VDP-2 [MKRH11] values indicate higher reconstruction quality.*

Figure 7: *An example of challenging simulated scene with a very large dynamic range (over 30 stops) zoom in for better view. Left: demosaicked LDR image for illustration purpose note the complete saturation in the fire region, Middle: tone-mapped HDR images for our results note the artifact in the saturated region. right: tone-mapped ground truth HDR image.*

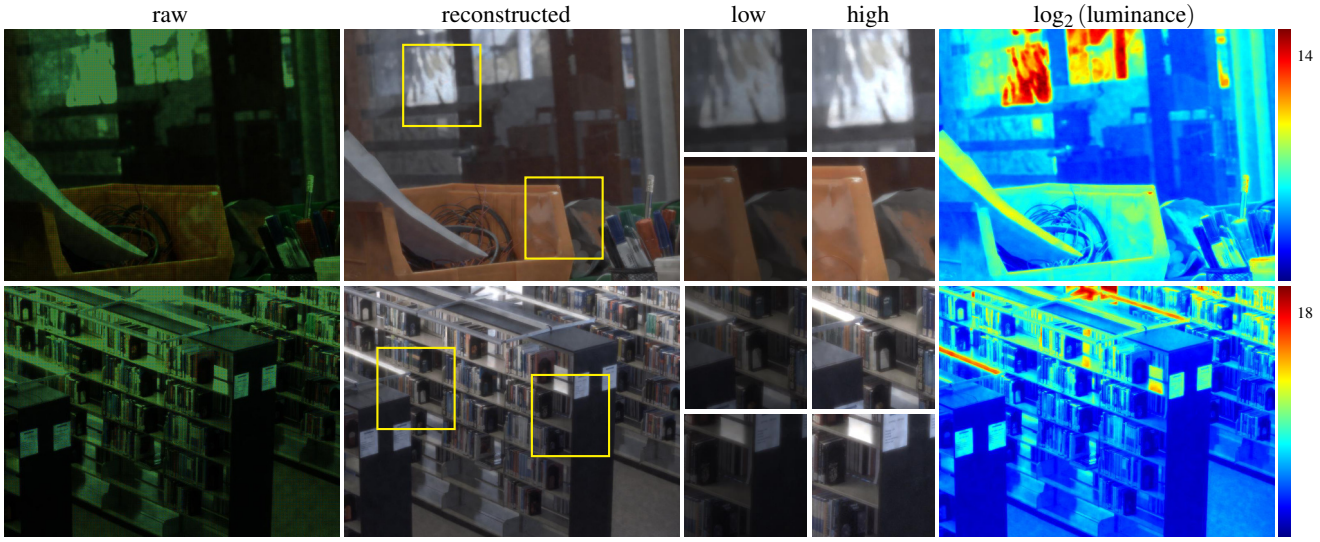

Figure 8: *Experimental results with our prototype. Left are raw Bayer images from the camera. The reconstructed HDR images (tone-mapped) are shown with zoom-in details for both low and high exposures indicating the dynamic range. Right are the $\log_2$ (luminance) maps showing the stop scales.*

[AFR*07] AKYÜZ A., FLEMING R., RIECKE B. E., REINHARD E., BÜLTHOFF H. H.: Do HDR displays support LDR content? A psychophysical evaluation. *ACM Trans. Graph. 26*, 3 (2007), 38. 6, 7

[BADC17] BANTERLE F., ARTUSI A., DEBATTISTA K., CHALMERS A.: *Advanced high dynamic range imaging*. AK Peters/CRC Press, New York, 2017. doi:10.1201/9781315119526. 1

[BDAPN14] BANITALEBI-DEHKORDI A., AZIMI M., POURAZAD M. T., NASIOPOULOS P.: Compression of high dynamic range video using the HEVC and H.264/AVC standards. In *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness* (2014), IEEE, pp. 8–12. 5

[BLD*07] BANTERLE F., LEDDA P., DEBATTISTA K., CHALMERS A., BLOJ M.: A framework for inverse tone mapping. *The Visual Computer 23*, 7 (2007), 467–478. 2

[BLDC08] BANTERLE F., LEDDA P., DEBATTISTA K., CHALMERS A.: Expanding low dynamic range videos for high dynamic range applications. In *Proceedings of the 24th Spring Conference on Computer Graphics* (2008), ACM, pp. 33–41. 2

[C*15] CHOLLET F., ET AL.: Keras: Deep learning library for theano and tensorflow., 2015. URL: https://keras.io. 5

[CK17] CHEN Q., KOLTUN V.: Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 1511–1520. 5

[Con12] CONTRIBUTORS S.: sIBL Archive Free HDRI sets for smart image-based lighting. http://www.hdrlabs.com/sibl/archive.html, 2012. 5

[DB16] DOSOVITSKIY A., BROX T.: Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems* (2016), pp. 658–666. 5

[DM97] DEBEVEC P. E., MALIK J.: Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), ACM Press/Addison-Wesley Publishing Co., pp. 369–378. 2

[EKD*17] EILERTSEN G., KRONANDER J., DENES G., MANTIUK R. K., UNGER J.: HDR image reconstruction from a single exposure using deep CNNs. *ACM Trans. Graph. 36*, 6 (2017), 178. 3, 5, 6, 7

[EKM17] ENDO Y., KANAMORI Y., MITANI J.: Deep reverse tone mapping. *ACM Trans. Graph. 36*, 6 (2017), 177:1–177:10. 2, 7

[Fai07] FAIRCHILD M. D.: The HDR photographic survey. In *Color and Imaging Conference* (2007), Society for Imaging Science and Technology, pp. 233–238. 5

[FS10] FUNT B., SHI L.: The rehabilitation of MaxRGB. In *Color and imaging conference* (2010), Society for Imaging Science and Technology, pp. 256–259. 5

[GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2414–2423. 5

[HYB13] HUO Y., YANG F., BROST V.: Dodging and burning inspired inverse tone mapping algorithm. *Journal of Computational Information Systems 9*, 9 (2013), 3461–3468. 2, 6, 7

[HYDB14] HUO Y., YANG F., DONG L., BROST V.: Physiological inverse tone mapping based on retina response. *The Visual Computer 30*, 5 (2014), 507–517. 2, 6, 7

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)* (2016), Springer, pp. 694–711. 5

[KB14] KINGMA D. P., BA J.: ADAM: A method for stochastic optimization. *arXiv preprint* (2014). arXiv:1412.6980. 5

[KGBU13] KRONANDER J., GUSTAVSON S., BONNET G., UNGER J.: Unified HDR reconstruction from raw CFA data. In *IEEE International Conference on Computational Photography (ICCP) 2013* (2013), IEEE, pp. 1–9. 2, 5

[KHC17a] KIM Y., HWANG I., CHO N. I.: Convolutional neural networks and training strategies for skin detection. In *IEEE International Conference on Image Processing (ICIP)* (2017), IEEE, pp. 3919–3923. 4

[KHC17b] KIM Y., HWANG I., CHO N. I.: A new convolutional network-in-network structure and its applications in skin detection, semantic segmentation, and artifact reduction. *arXiv preprint* (2017). arXiv:1701.06190. 4

[KKM14] KENSEI J., KAIZU S., MITSUNAGA T.: Image processing including image correction, Sept. 30 2014. US Patent 8,848,063. 2

[KO14] KOVALESKI R. P., OLIVEIRA M. M.: High-quality reverse tone mapping for a wide range of exposures. In *27th SIBGRAPI Conference on Graphics, Patterns and Images* (2014), IEEE, pp. 49–56. 2, 6, 7

[KR17] KALANTARI N. K., RAMAMOORTHI R.: Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph. 36*, 4 (2017), 144:1–144:12. 2

[KTC12] KUO P.-H., TANG C.-S., CHIEN S.-Y.: Content-adaptive inverse tone mapping. In *Visual Communications and Image Processing (VCIP), 2012 IEEE* (2012), IEEE, pp. 1–6. 6, 7

[KW14] KONNIK M., WELSH J.: High-level numerical simulations of noise in CCD and CMOS photosensors: review and tutorial. *arXiv preprint* (2014). arXiv:1412.4031. 3, 5

[LAK18a] LEE S., AN G. H., KANG S.-J.: Deep chain HDRI: Reconstructing a high dynamic range image from a single low dynamic range image. *arXiv preprint* (2018). arXiv:1801.06277. 2

[LAK18b] LEE S., AN G. H., KANG S.-J.: Deep recursive HDRI: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 596–611. 2

[Lan02] LANDIS H.: Production-ready global illumination. *Siggraph course notes 16*, 2002 (2002), 11. 6

[MKRH11] MANTIUK R., KIM K. J., REMPEL A. G., HEIDRICH W.: HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Trans. Graph.* (2011), vol. 30, ACM, p. 40. 4, 5, 6, 7

[MMP*07] MCGUIRE M., MATUSIK W., PFISTER H., CHEN B., HUGHES J. F., NAYAR S. K.: Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications 27*, 2 (2007). 2

[MPMP95] MANN, PICARD, MANN S., PICARD R. W.: On being "undigital" with digital cameras: extending dynamic range by combining differently combining differently exposed pictures. In *Proceedings of IS&T 1995* (1995), pp. 442–448. 2

[MSG17] MASIA B., SERRANO A., GUTIERREZ D.: Dynamic range expansion based on image statistics. *Multimedia Tools and Applications 76*, 1 (2017), 631–648. 2, 6, 7

[NB03] NAYAR S. K., BRANZOI V.: Adaptive dynamic range imaging: Optical control of pixel exposures over space and time. In *IEEE International Conference on Computer Vision (ICCV)* (2003), vol. 2, IEEE, pp. 1168–1175. 2

[NM00] NAYAR S. K., MITSUNAGA T.: High dynamic range imaging: Spatially varying pixel exposures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2000), vol. 1, IEEE, pp. 472–479. 2, 3, 6

[NN02] NAYAR S. K., NARASIMHAN S. G.: Assorted pixels: Multi-sampled imaging with structural models. In *European Conference on Computer Vision (ECCV)* (2002), pp. 636–652. 2

[RHD*10] REINHARD E., HEIDRICH W., DEBEVEC P., PATTANAIK S., WARD G., MYSZKOWSKI K.: *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 2

[RSSF02] REINHARD E., STARK M., SHIRLEY P., FERWERDA J.: Photographic tone reproduction for digital images. *ACM Trans. Graph. 21*, 3 (2002), 267–276. 2

[RTS*07] REMPEL A. G., TRENTACOSTE M., SEETZEN H., YOUNG H. D., HEIDRICH W., WHITEHEAD L., WARD G.: LDR2HDR: on-the-fly reverse tone mapping of legacy video and photographs. In *ACM Trans. Graph.* (2007), vol. 26, ACM, p. 39. 2

[SBS*12] SCHÖBERL M., BELZ A., SEILER J., FOESSEL S., KAUP A.: High dynamic range video by spatially non-regular optical filtering. In *19th IEEE International Conference on Image Processing (ICIP), 2012* (2012), IEEE, pp. 2757–2760. 2, 3

[SHG*16] SERRANO A., HEIDE F., GUTIERREZ D., WETZSTEIN G., MASIA B.: Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 153–163. 2, 3, 6

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2014). arXiv:1409.1556. 5

[TKTS11] TOCCI M. D., KISER C., TOCCI N., SEN P.: A versatile HDR video production system. In *ACM SIGGRAPH 2011* (New York, NY, USA, 2011), ACM, pp. 41:1–41:10. 2

[VRA*07] VEERARAGHAVAN A., RASKAR R., AGRAWAL A., MOHAN A., TUMBLIN J.: Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *ACM Trans. Graph.* (2007), vol. 26, ACM, p. 69. 3

[WXTT18] WU S., XU J., TAI Y.-W., TANG C.-K.: Deep high dynamic range imaging with large foreground motions. In *European Conference on Computer Vision (ECCV)* (2018), pp. 117–132. 2

[XDCW02] XIAO F., DICARLO J. M., CATRYSSE P. B., WANDELL B. A.: High dynamic range imaging of natural scenes. In *Color and Imaging Conference* (2002), Society for Imaging Science and Technology, pp. 337–342. 5

[ZLX*14] ZHOU B., LAPEDRIZA A., XIAO J., TORRALBA A., OLIVA A.: Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems* (2014), pp. 487–495. 5

[ZZMZ12] ZHANG L., ZHANG L., MOU X., ZHANG D.: A comprehensive evaluation of full reference image quality assessment algorithms. In *19th IEEE International Conference on Image Processing (ICIP)* (2012), IEEE, pp. 1477–1480. 5