MantissaCam: Learning Snapshot High-dynamic-range Imaging with Perceptually-based In-pixel Irradiance Encoding

Haley M. So[1] Julien N.P. Martel[1] Piotr Dudek[2] Gordon Wetzstein[1]
[1]Stanford University [2]The University of Manchester
computationalimaging.org/publications/mantissacam/

Abstract

The ability to image high-dynamic-range (HDR) scenes is crucial in many computer vision applications. The dynamic range of conventional sensors, however, is fundamentally limited by their well capacity, resulting in saturation of bright scene parts. To overcome this limitation, emerging sensors offer in-pixel processing capabilities to encode the incident irradiance. Among the most promising encoding schemes is modulo wrapping, which results in a computational photography problem where the HDR scene is computed by an irradiance unwrapping algorithm from the wrapped low-dynamic-range (LDR) sensor image. Here, we design a neural network–based algorithm that outperforms previous irradiance unwrapping methods and we design a perceptually inspired "mantissa" encoding scheme that more efficiently wraps an HDR scene into an LDR sensor. Combined with our reconstruction framework, MantissaCam achieves state-of-the-art results among modulo-type snapshot HDR imaging approaches. We demonstrate the efficacy of our method in simulation and show benefits of our algorithm on modulo images captured with a prototype implemented with a programmable sensor.

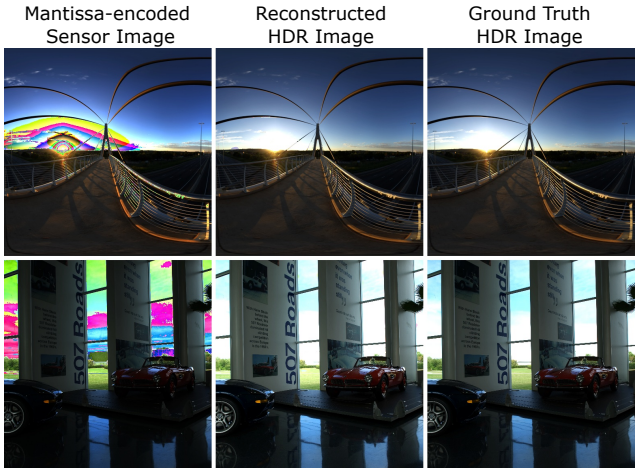

Figure 1: MantissaCam electronically encodes the irradiance incident on the sensor into an LDR image by wrapping the intensity in a perceptually inspired manner (left). The proposed reconstruction algorithm estimates the HDR scene from this LDR image (center) and achieves accurate reconstructions compared to the ground truth (right).

1. Introduction

High Dynamic Range (HDR) imaging is crucial for a vast range of applications, including automotive vision systems [27], HDR display [54], and image processing [48, 5]. When capturing natural scenes, which can have an extreme high dynamic range [48], the level of detail is limited by the full well capacity and the quantization precision of the sensor. Unfortunately, the dynamic range offered by modern sensors is far smaller than that encountered in the wild [46], making specialized sensors or computational photography approaches to HDR imaging necessary.

Among the many HDR imaging techniques proposed in the literature, exposure bracketing [32, 11, 39, 21, 19, 22] and temporally varying exposures [26, 55, 25] can be successful, but fast motion introduces ghosting. Multi-sensor approaches [3, 38, 60] can overcome this limitation, but are expensive, bulky, and difficult to calibrate. Existing snapshot HDR imaging approaches hallucinate saturated image detail using neural networks [34, 13, 14, 29, 53], use spatially varying pixel exposures which trade spatial resolution for dynamic range [45, 43, 44, 64, 20, 56, 4, 35], or use optical encoding approaches that blur the sensor image [52, 40, 59]. Specialized sensors, for example recording logarithmic irradiance [30] or floating point extended dynamic range values [**?**] have also been proposed, but these either trade extended dynamic range for precision or require additional bandwidth.

Our work (Fig. 1) is inspired by the idea of electronically applying a modulo encoding of the irradiance on the sensor followed by an intensity unwrapping algorithm [68, 69].

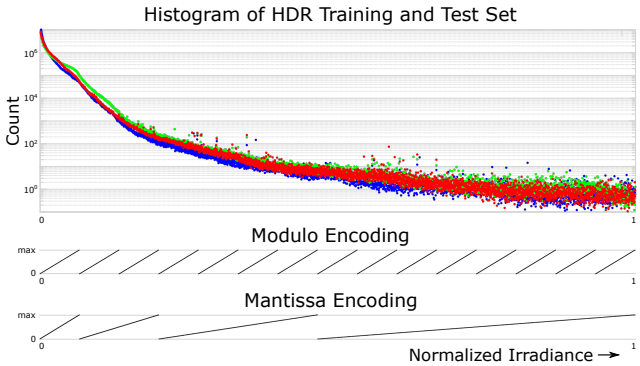

Figure 2: Log histogram of normalized irradiance values of all pixels in our training and test sets of HDR images for all color channels (top). This histogram is highly biased towards low-intensity values, indicating that irradiance values of natural images are not uniformly distributed. Yet, the modulo encoding subdivides this intensity range uniformly and wraps each of these areas into the available dynamic range of the sensor, as shown for a 1D ramp (center). The proposed mantissa encoding wraps the same 1D ramp in a perceptually more uniform manner in log space, which is observed as non-uniform wrapping in irradiance space (bottom).

This idea is beneficial over other snapshot approaches, because it does not degrade a low-dynamic-range (LDR) image, as optical encoding approaches do, it does not hallucinate detail but recovers them, it does not decrease image resolution, or increase the required bandwidth. As we show in this paper, there are several downsides to the modulo camera, as proposed in prior work. Specifically, modulo wrapping is done directly in irradiance space, which allocates precision and number of wraps linearly in this domain. However, the human visual system is perceptually approximately linear in the log-domain, so a conventional modulo encoding wastes precision for detail that we do not perceive. Moreover, the irradiance distribution of natural scenes is heavily skewed towards darker values (see log-histograms in Fig. 2), so it makes sense to nonlinearly distribute the irradiance wraps in order to minimize their number, because they have to be computationally unwrapped again.

We address these challenges by proposing a perceptually inspired modulo-type wrapping scheme that operates in the log-irradiance domain. This idea intuitively combines the principles of operation of both log [30] and modulo [68] cameras. Indeed, the signal we propose to measure is essentially a generalization of the mantissa used by the IEEE Standard for Floating-Point Arithmetic [23], or the log base 2 of the intensity modulo the well capacity. We demonstrate that such a log-modulo or *mantissa* camera allocates precision in a perceptually meaningful manner and it non-

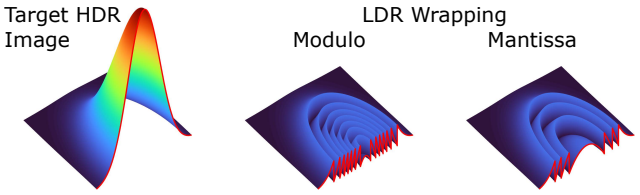

Figure 3: Example showing an HDR Gaussian function wrapped using the modulo and mantissa encoding in an LDR image. For this example, the modulo encoding requires more wraps than the mantissa encoding, which makes its reconstruction via computational unwrapping more challenging.

linearly distributes the wraps in irradiance space to better match the distribution of irradiance values in natural scenes (see Fig. 2, top). This directly leads to fewer wraps of natural scenes (see Figs. 2, center and bottom, and 3), which make the inverse problem of 2D irradiance unwrapping easier to solve. To solve the unwrapping problem, we introduce a neural network architecture that is more robust than prior work using graph cut algorithms [68] or other network architectures [69]. Finally, we prototype a modulo camera using a SCAMP-5 programmable sensor [10] which allows flexible re-configuration of the in-pixel irradiance encoding in software. These types of programmable sensors are expected to be widely available in the near future.

Specifically, we make the following contributions

- We introduce MantissaCam as a new snapshot approach to HDR imaging, combining perceptually motivated irradiance encoding and decoding.
- We develop a neural network architecture that outperforms existing unwrapping methods for modulo cameras and that demonstrates state-of-the-art performance with our mantissa encoding.
- We build a prototype modulo camera and show improved results over previous methods.

Overview of Limitations.

The SCAMP sensor we have does not include the log circuitry needed for capturing mantissa images, but we still demonstrate the benefits of the proposed reconstruction algorithm on captured modulo images.

2. Related Work

HDR Imaging. The limited dynamic range of conventional camera sensors has been addressed by a number of computational imaging techniques. Exposure bracketing, for example, fuses several low-dynamic-range (LDR) photographs into a single HDR image [32, 11, 39, 21, 19, 22].

Temporally varying exposures can also be processed to obtain HDR videos [26, 55, 25]. Yet, slight movements in the scene will create ghosting artifacts, which are challenging to be removed [61]. Another class of approaches involves multiple sensors to capture these LDR images simultaneously [3, 38, 60]. Although successful, these systems are expensive, bulky, and often difficult to calibrate.

Several approaches have been developed to estimate an HDR image from a single input image. Reverse tone mapping approaches aim at inverting a tone mapping operator [6, 41, 49], which is an ill-posed inverse problem. Convolutional neural networks can also be directly applied to an LDR image to hallucinate the HDR image [34, 13, 14, 29, 53]. Neither of these approaches, however, has the capability to recover true image details. Bright highlights can also be optically encoded in an LDR image [52, 40, 59], but this approach relies on the required deconvolution to clean up even an LDR scene perfectly to compete with the quality of conventional sensors, which is challenging. Spatially varying pixel exposures are a promising direction but, similar to color filter arrays, they trade spatial resolution for dynamic range [45, 43, 44, 64, 20, 56, 4, 35].

Among these, our approach to snapshot HDR imaging is most closely related to the modulo camera [68], which combines a modulo-type encoding of the irradiance on the sensor combined with a reconstruction algorithm that solves a 2D unwrapping problem. A conventional modulo operation, however, makes it difficult to distinguish between wrapping boundaries and high-frequency image detail. We introduce a perceptually motivated intensity wrapping technique for this class of computational cameras, which better preserves high-frequency image detail and dynamic range, and we also improve upon existing 2D upwrapping algorithms developed for related tasks.

Unwrapping Algorithms. Phase unwrapping is a problem often encountered in optical interferrometry, where the surface profile of some optical element or scene can be indirectly imaged as the wrapped phase of a coherent reference beam. A number of algorithms to unwrap these interferrograms has been developed, as surveyed in [18]. When working with wrapped intensities of natural images, instead of optical phase values, the complex interplay of high spatial frequencies and drastically varying light intensity has to be accounted for. Unwrapping techniques for natural images has been analyzed [8] and tailored algorithms developed [28, 57, 58], but these require multiple input images. Most recently, the UnModNet network architecture was introduced to unwrap a single intensity image with state-of-the-art quality [69]. Our network architecture improves upon this method for HDR imaging for modulo cameras but shows best results when used with the proposed mantissa encoding scheme.

Floating point sensors from the early 2000s allow for capturing high dynamic range with multiple sampling [2], [65] and variations with overlapping integration intervals[1], or choosing optimum integration time [50]. Floating point sensors have great potential, however they require additional bandwidth. Our work reconstructs an HDR image from a captured image of the same bit depth as a conventional LDR sensor, utilizing the programmability of new sensors for in-pixel irradiance encoding together computational post-processing of that data.

Exotic Sensors for HDR Imaging. Specialized sensor circuits have been developed to support spatially varying pixel and adaptive exposures [37, 65, 12, 63, 36] as well as logarithmic [30] or modulo [62, 9, 68] irradiance encoding. Emerging photon-counting sensors can facilitate HDR imaging but require high-speed readout circuitry and are best suited for low-light applications [17] or observe response functions that are similar to logarithmic sensors [24]. All of these systems are inflexible, because they are not programmable. Near-focal-plane sensor–processors [66] include some amount of computing capabilities in the sensor and related systems have become programmable [42, 31, 51, 67, 16, 10]. In this work, we use one of these platforms, SCAMP-5 [10], to prototype modulo encoding and the proposed neural network–based HDR reconstruction algorithm experimentally.

3. Perceptually-based HDR Imaging

The MantissaCam framework comprises an electronic in-pixel irradiance encoding scheme and a neural network–based decoding algorithm, which solves the 2D unwrapping problem to reconstruct the irradiance incident on the sensor. We discuss these aspects next.

3.1. In-pixel Irradiance Encoding

The image formation model of the MantissaCam is

$$I_{\text{sensor}}(x,y) = q\left(\text{mod}\left(\log_{\alpha}\left(I(x,y)\right), I_{\text{max}}\right)\right) + \eta, \quad (1)$$

where I describes the spatially varying irradiance (i.e., the target HDR image) on the sensor, I_{sensor} is the measured LDR sensor image, and η is zero-mean additive Gaussian noise. The parameter α models a family of logarithmic irradiance response functions. For example, the special case $\alpha = 2$ of our encoding scheme is similar to the mantissa encoding of the IEEE 754 standard for floating point arithmetic. Sensor quantization is modeled by the function $q(\cdot)$. I_{max} is the maximum allowed irradiance value before the intensity wraps. This could be the well capacity of a pixel or a user-defined value that is slightly lower than that.

3.2. Irradiance Decoding

The proposed decoding scheme is implemented by two neural networks. The first takes the wrapped sensor image

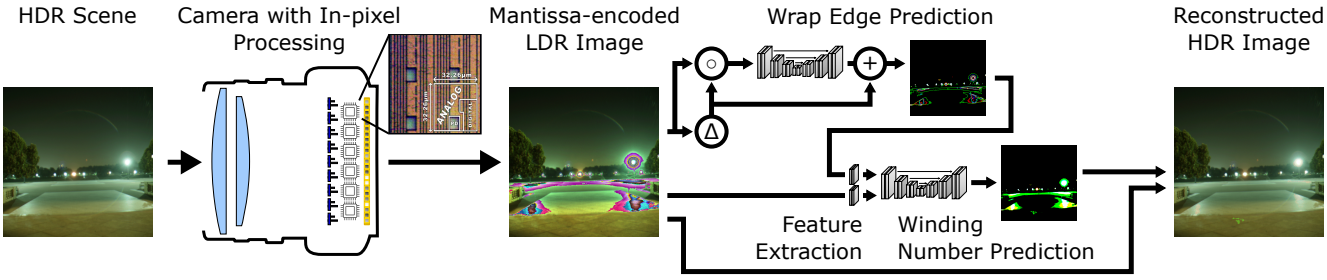

Figure 4: MantissaCam pipeline. An HDR scene is imaged by a camera with in-pixel processing capabilities, implementing the proposed irradiance encoding scheme (left). The resulting LDR sensor image encodes lower irradiance values similar to a conventional camera, but bright image regions, including the lamp and the reflections on the ground, are wrapped rather than saturated (center). The mantissa-encoded image is first processed by a network that predicts the wrap edges and then by another network that predicts the winding number (center right). The per-pixel winding number, together with the mantissa-encoded image, are used to reconstruct the HDR image (right). The symbols Δ, $\circ$, and $+$ denote channel-wise Laplacian operators, channel concatenation, and addition, respectively.

as input and predicts the wrap edges, effectively separating them from the texture edges. The second network predicts the winding number (i.e., the number of times intensity has wrapped) of each pixel from these wrap edges.

To predict either modulo or mantissa wrap edges from a sensor image, we directly use the "modulo edge separator" proposed as part of the UnModNet architecture [69]. This edge separator is a residual-type convolutional neural network (CNN) that takes as input a concatenation of the LDR sensor image and a Laplacian-filtered copy of the same. We illustrate our network in Figure 4 and refer the interested reader to [69] for additional details.

Our second network predicts the winding number for each pixel, $W(x, y)$, given the wrap edges and the sensor image as input. For this purpose, features are extracted from both input images using the lightweight CNN-based feature extraction layers from [69]. These are fed into an attention UNet [47] with four downsampling and four upsampling blocks, with each downsampling block using a strided convolutional layer and a residual bottleneck block, and each upsampling block mirroring it but with the addition of attention gates. This is a standard neural network architecture, but its application to directly predicting the winding number of irradiance-wrapped images is new. Note that this part of our algorithm is substantially different from the iterative, graph-cuts inspired unwrapping procedure proposed in [69]. Their method aims at unwrapping the HDR image layer by layer, which is prone to propagating errors, whereas our approach directly predicts the number of wraps, i.e., the winding number, using a single pass through the UNet. We discuss additional details of this network architecture in the supplement and outline the training procedure of both networks in Section 3.4 and the supplement.

Given the predicted winding number for each pixel as well as the raw sensor I_{sensor}, we formulate the reconstruction of the HDR image $\widetilde{I}$ as

$$\widetilde{I}(x, y) = \alpha^{I_{\text{sensor}} + W(x,y) \cdot I_{\max}}. \tag{2}$$

In our implementation, we choose $\alpha = 2$.

3.3. Understanding the Relation between Resolution and Dynamic Range

The theory addressing the ability to perfectly reconstruct a signal with MantissaCam falls within the framework of unlimited sampling recently developed in [7, 8]. Here, rather than formally treating the reconstruction problem, we attempt to highlight the advantages of a mantissa over a modulo encoding and develop an understanding of the tradeoffs between those.

Let us consider the 1D band-limited irradiance function $I(x)$, with maximal frequency $f_{\max}$. The irradiance is encoded on the sensor by the wrapping function $\mathcal{W}$ of the imaging model:

$$\mathcal{W} : I \in \mathbb{R}_+ \mapsto \mathcal{W}(I) \in [0, I_{\max}]. \tag{3}$$

In particular, we consider the two wrapping functions:

$$\mathcal{W}_{\text{mod}}(I) = I - W\big(I(x)\big) \cdot I_{\max}, \tag{4}$$

and

$$\mathcal{W}_{\text{mant}}(I) = \log_\alpha(I) - W\big(\log_\alpha I(x)\big) \cdot I_{\max}, \tag{5}$$

with $W(\,\cdot\,) = \left\lfloor \frac{\cdot}{I_{\max}} \right\rfloor$ and $\lfloor \cdot \rfloor$ being the floor function.

In order to avoid aliasing on our discrete sensor array, we assume the sampling of I respects the Nyquist sampling criterion $f_s > 2 \cdot f_{\max}$, with the sampling frequency f_s

related to the inverse pixel pitch $T_s = \frac{1}{f_s}$ (i.e. the resolution or pixel density, for instance expressed in line pairs per millimeter) of the sensor array.

Recoverability of irradiance from modulo and mantissa encodings To get an intuition about the irradiance fields $\mathcal{W}(I)$ that can be perfectly reconstructed, let us consider the discretized irradiance $I[n] = I(n \cdot T_s)$ as seen by a pixel n.

If a wrap of $\mathcal{W}(I)$ occurs within a pixel, information is lost and it is impossible to reconstruct the incident irradiance field. Therefore, a set of conditions to recover the field is:

$$\begin{cases} |\mathcal{W}(I[n+1]) - \mathcal{W}(I[n])| \leq I_{\max}, \\ |W(I[n+1]) - W(I[n])| \leq 1, \end{cases} \tag{6}$$

where the first condition derives from the Euclidean Division Theorem and makes sure we cannot wrap "within" a pixel, the second condition allows at most one wrap between two pixels.

For the modulo encoding those conditions translate into

$$\big|I[n+1] - I[n]\big| \leq I_{\max}, \tag{7}$$

and for the mantissa encoding we have that

$$\big|\log_\alpha(I[n+1]) - \log_\alpha(I[n])\big| \leq I_{\max}. \tag{8}$$

This shows that while the modulo encoding can reconstruct any irradiance with arithmetic growth of $I_{\max}$, a mantissa encoding can reconstruct a larger class of functions with geometric growth of $I_{\max}$.

Dynamic range. For both types of encoding, these results imply an interesting tradeoff between the dynamic range of the sensor and its spatial resolution. With two sensors of the same size, using different pixel pitches T_s and T_s' such that $T_s' > T_s$, the sensor with a smaller pixel pitch T_s (i.e., of higher resolution) can reconstruct faster spatial variations of irradiance ($\frac{I_{\max}}{T_s} > \frac{I_{\max}}{T_s'}$ in the modulo case). Therefore, there is a relationship between the maximum dynamic range recoverable for a sensor given its resolution. For two sensors of fixed size with N pixels, the maximum recoverable irradiance is a ramp starting at pixel $n = 0$ and ending at pixel $n = N - 1$. In this setting, the sensor with modulo encoding can reconstruct a maximum dynamic range of $\mathrm{DR} \approx 10 \log(N \cdot I_{\max})$dB while the one with a mantissa encoding can recover a much wider dynamic range of $\mathrm{DR} \approx 10 \cdot N \log(I_{\max})$dB.

Quantization. The ultra-high dynamic range of the mantissa encoding comes at the expense of loss of precision. In practice, no sensor has infinite bit depth but is quantized to 8–12 bits. As shown in the bottom graphs of Figure 2, the same number of levels are distibuted on a much wider range as the winding number W increases. This means a MantissaCam cannot resolve irradiance with the same precision ModuloCam can at high irradiance levels—the quantization error is higher for our encoding. Yet, early psychophysics studies [15] noted that perceived light intensity is proportional to the logarithm of the light intensity. Known as Fechner-Weber law, this implies that the coarser quantization of MantissaCam at high irradiance levels might not be perceptually important.

3.4. Dataset and Implementation Details

For a fair comparison, the dataset used to train and evaluate our model was the same dataset created by UnModNet [69]. We randomly split the images into 400 training images and 193 testing images. We used the same process to augment the training dataset, over-exposing and cropping images to yield a total of 5,945 training images.

We train our networks in three stages. First, we train the wrap edge prediction network by itself for 400 epochs, taking simulated sensor images as input, using a binary cross entropy loss with the ground truth wrap edge images obtained via simulation. Second, we train the winding number prediction network by itself for 200 epochs, taking simulated sensor images and ground truth wrap edges as input, using a mean-squared error (MSE) loss on the ground truth winding number. Third, we train both networks end-to-end for another 200 epochs using the same MSE loss on ground truth winding number. Additional implementation details are found in the supplement.

Encoder	Modulo			Mantissa	None
Decoder	Graph Cuts [68]	UnModNet [69]	Ours	Ours	CNN [13]
PSNR (↑)	21.4	29.5	32.2	**37.4**	22.7*
Q Score (↑)	48.0	59.1	57.1	**60.9**	47.7*
SSIM (↑)	0.80	0.79	0.84	**0.97**	0.72*
MSSIM (↑)	0.82	0.91	0.93	**0.99**	0.76*
LPIPS (↓)	0.29	0.12	0.10	**0.03**	—

Table 1: Quantitative evaluation of modulo and mantissa in-pixel encoding combined with various reconstruction algorithms for simulated data. Our irradiance unwrapping network performs better than existing algorithms on the modulo encoding, as evaluated by several metrics. Combined with the proposed mantissa encoding, our approach achieves state-of-the-art results. We also show the quality of a CNN working with conventional LDR images using the same dataset. Values marked with * are reproduced from [69].

Figure 5: Evaluation of encoding and decoding schemes in simulation. A conventional modulo encoding wraps the irradiance of a scene into an LDR sensor image (column 1). A graph cuts–based reconstruction algorithm [68] usually performs poorly (column 2) whereas the recently proposed UnModNet architecture [69] often estimates reasonable HDR images (column 3). Yet, the proposed reconstruction framework works best among these methods (column 4). Moreover, the proposed mantissa encoding scheme (column 5) induces fewer irradiance wraps making it easier to reconstruct the HDR image using our framework (column 6). Our approach achieves reconstructions closest to the ground truth (column 7). 'P', 'S', and 'Q' indicate the PSNR, SSIM and Q-score for each reconstruction method.

4. Experiments

4.1. Evaluation on Synthetic Data

Figure 5 qualitatively and quantitatively compares modulo and mantissa encoding schemes combined with different reconstruction algorithms. Using a single modulo-wrapped image as input, graph cuts perform poorly [68]. The UnModNet network [69] does reasonably well in some cases, but struggles to reconstruct the large bright parts of the first example scene and the lights on the bridge of the third scene. Their iterative unwrapping procedure sometimes fails in stopping to unwrap, which results in extremely high irradiance values lowering their PSNR and obscuring fine image detail. Our algorithm achieves a better quality than these methods on the same modulo-encoded images, as evaluated by the peak signal-to-noise (PSNR or P),

structural similarity (SSIM or S), and Q-score of the perceptual HDR-VDP-2 [33] metrics. Moreover, when combined with the proposed mantissa irradiance encoding scheme, our framework achieves the best results among all of these methods.

Table 1 also quantitatively compares all of these approaches using several different metrics on the test set of the dataset described in Sec. 3.4. In addition to the above methods, we also include a comparison to a CNN operating directly on a conventional LDR sensor image to hallucinate the HDR scene [13]. Not shown are the results from the combination of the UnModNet architecture with the mantissa encoding. The average PSNR was less than 10 dB due to UnModNet's iterative unwinding. It is prone to propagating errors and with the mantissa encoding, the errors are "exponentially" propagated. As shown in Table 1, the proposed mantissa encoding scheme combined with our reconstruction framework achieves the best results using all metrics, outperforming the state of the art, i.e., UnModNet, by almost 8 dB of PSNR.

All simulations with synthetic data are run on noise-free images to study the upper bound of all of these algorithms. However, we do include results of simulations with simulated sensor noise in the supplement and also evaluate the best-performing algorithms on noisy captured data in the following.

4.2. Prototyping a Modulo Camera using SCAMP-5

We build a physical prototype using an example of an emerging class of sensors, dubbed focal-plane sensor–processors [66], that embed small processing circuits inside each pixel. We use SCAMP-5 [10], whose processing elements (PE) are programmable in a single instruction multiple data (SIMD) fashion, similar to a GPU where the same instruction is performed for all processing elements simultaneously on some local piece of data. Specifically, a PE is equipped with a few analog and digital memories. Instructions can be performed as light is being collected by the pixel's photo-sensitive element, thus enabling to change the way integration is performed, as required for our implementation. In other SCAMP versions, there is log circuitry that would allow us to take mantissa images, however, our version does not have this capability. We are still able to implement the modulo camera and show the benefits of our reconstruction method over previous state-of-the-art methods.

4.3. Experimental Results

We use SCAMP-5 to prototype a modulo camera and capture HDR scenes outside (see Fig. 6). This sensor records grayscale images with a resolution of 256×256 pixels. For this experiment, we retrained both UnModNet and our network on modulo images using the same training

Figure 6: Prototype camera capturing an outdoor HDR scene.

procedure described in Section 3.4, but on grayscale images captured with SCAMP-5. For this purpose, we collected a dataset of 14,810 modulo and corresponding reference HDR images using the SCAMP-5 prototype. We split this dataset into 13,329 training images and 1,481 test images. No artificial data augmentation was performed. We trained a single edge predictor network that we used for UnModNet's iterative unwrapping approach and also as part of our own pipeline. This network was trained for 150 epochs using the experimentally captured dataset.

Figure 7 shows captured modulo images, the tonemapped reconstructions, and a tonemapped reference HDR image. The captured images include sensor noise, which is especially noticeable around the irradiance wraps. The graph cuts and UnModNet algorithms usually fail to estimate reasonable HDR images, likely due to the noise in the sensor images. For more recognizable results, we limited the number of unwrappings for UnModNet to a maximum of five iterations. Otherwise, the reconstructions end up completely white. The dynamic range of this scene is far greater than that of the sensor, yet our method is able to reconstruct HDR images with high quality.

Table 2 shows the comparison of graph cuts, UnModNet, and our method averaged over the test set captured with the SCAMP-5. We compare PSNR, Q score, SSIM, MSSIM, and LPIPS scores. Across all metrics, ours outperforms previous methods by a large margin.

With our single-shot HDR image unwrapping method, we can also capture short HDR video clips, which would have been difficult to do with conventional HDR methods like bracketed exposures. In Figure 8, we show a sequence of modulo-encoded frames that we captured while moving the camera. We also show tonemapped reconstructions using UnModNet and our network. Our method unwraps the

Figure 7: Experimental results. Using a programmable sensor, SCAMP-5, we capture (noisy) modulo images (left) and process them using graph cuts, UnModNet, and our network applied to the captured modulo data. Tonemapped results using all these reconstruction methods as well as a reference HDR image are shown for several different scenes.

modulo video sequence with high temporal consistency and good quality, while lots of flickering and poor image quality are observed for UnModNet. Video clips of these and other example scenes are included in the supplemental material.

5. Discussion

Motivated by the emerging class of programmable sensors, we demonstrate new capabilities they could enable for the long-standing challenge of snapshot HDR imaging. For this purpose, we develop a reconstruction algorithm for the modulo camera that is more robust and achieves better results than the current state of the art. Moreover, we introduce the mantissa encoding scheme that is inspired by the human visual system and achieves a favorable tradeoff between dynamic range, spatial frequency, and precision when encoding HDR scenes compared to the modulo camera. We evaluate our system in simulation but also show preliminary results captured with a prototype SCAMP-5 programmable sensor, demonstrating the effectiveness of our reconstruc-

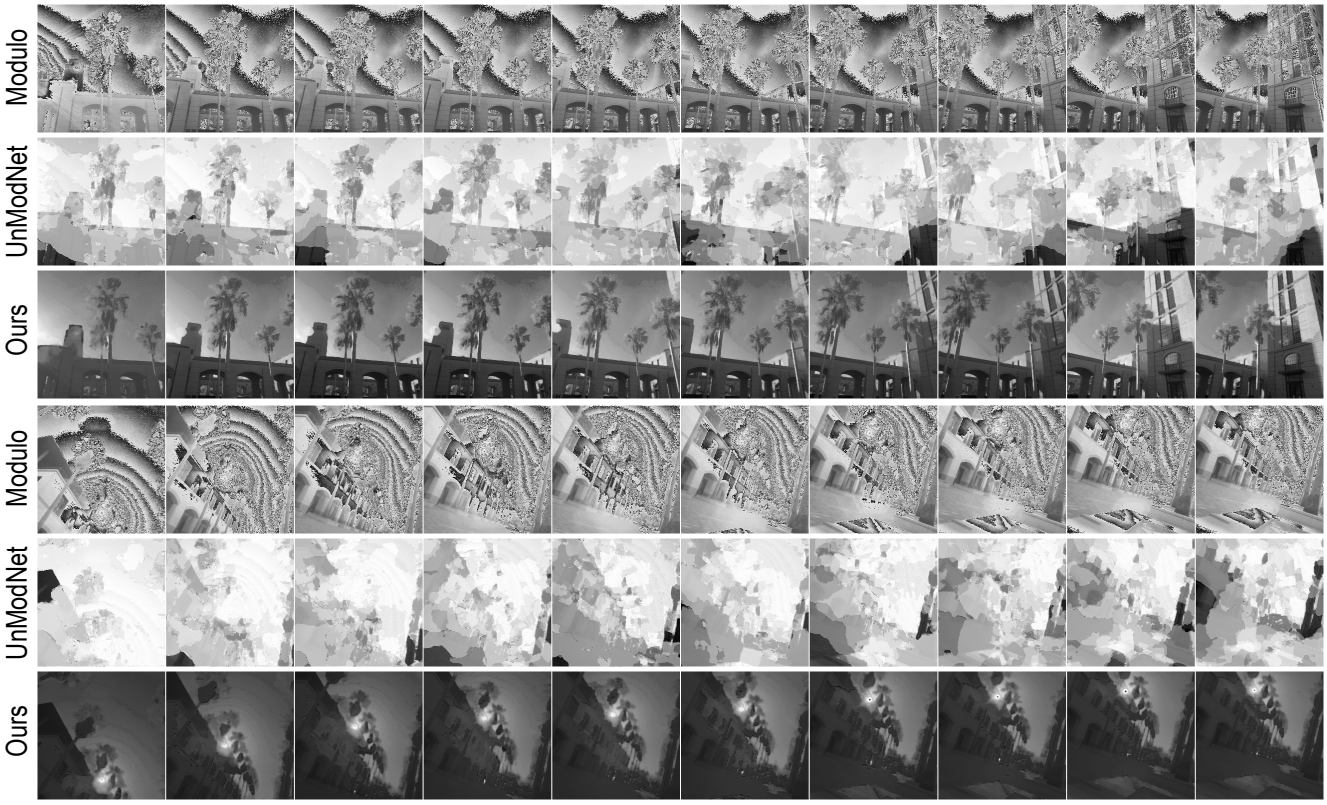

Figure 8: HDR video experimental results. We show 10 frames of two captured modulo video sequences, UnModNet's reconstruction, and our reconstruction. Our reconstruction shows temporal consistency and good image quality whereas UnModNet typically fails to estimate reasonable results.

Encoder	Modulo		
Decoder	Graph Cuts [68]	UnModNet [69]	Ours
PSNR (↑)	20.3	15.2	**33.7**
Q Score (↑)	43.7	45.7	**53.0**
SSIM (↑)	0.27	0.52	**0.85**
MSSIM (↑)	0.23	0.59	**0.95**
LPIPS (↓)	0.14	0.12	**0.09**

Table 2: Quantitative evaluation of modulo in-pixel encoding combined with various reconstruction algorithms for experimentally captured data. Our algorithm processing the same modulo images as the others achieves significantly better results in all relevant metrics.

tion algorithm on the modulo camera. The global shutter speed in our simulations and with the prototype are always set to capture the desired level of detail in the dark regions, relying on the encoder and reconstruction algorithm to recover the brightest parts of the scene.

Limitations and Future Work. Although promising, the proposed system has several limitations. First, our reconstruction pipeline improves results over existing work by a large margin, yet it fails in some cases as shown in Figure 9. Thus, there is room for further improving the robustness of the algorithm. Second, our mantissa-based encoding scheme is intuitive and robust, but the question of what an optimal encoding scheme for HDR imaging or other applications remains. Some prior work has studied end-to-end-optimized in-pixel irradiance encoding [35], which could be a fruitful direction for (un)wrapping-based HDR cameras, such as ours. Yet, optimizing periodic objective functions, such as modulo and mantissa-like functions, is not trivial and requires additional research. Third, the class of computational HDR cameras we discuss here seeks to improve the dynamic range of sensors for bright scene parts, but it does not necessarily improve the black level or performance in low-light conditions. It would be valuable to study how in-pixel intelligence offered by programmable sensors could help imaging in low-light scenarios, although this is beyond the scope of our work. Fourth, in our experiments we ignore the effect of the color filter array (CFA), primarily because

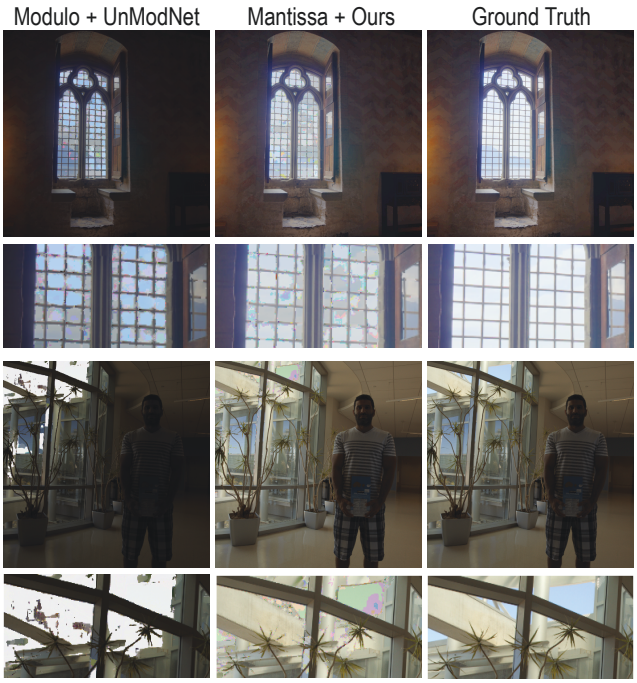

Figure 9: Limitations. Challenging areas for unwrapping often include regions with high spatial detail and wrapping or dense edges where it may be difficult for the networks to differentiate between wrap and texture edges. While our method is able to better reconstruct some of these areas than a modulo camera with the UnModNet algorithm, some artifacts remain.

our prototype is grayscale.

Furthermore, our SCAMP-5 prototype has many hardware limitations, including a high read noise level, low pixel fill factor, low resolution, lack of color filters, and a challenging software interface. Improving these aspects with better circuit design, 3D fabrication techniques, and improved firmware engineering could make this or related platforms better and more accessible to the computational photography community. The programmable sensor is a valuable tool in early experimentation. Ultimately, it could be replaced by a specialized CMOS image sensor device, implementing, in hardware, the optimized version of the mantissa-like encoding.

Conclusion. The emerging class of programmable sensors enables in-pixel intelligence, offering new imaging capabilities for computational photography systems. While our system demonstrates a new co-design of in-pixel irradiance encoding and decoding for snapshot HDR imaging, many other applications in computer vision, photography, and autonomous driving could be enabled by this platform. Our work takes first steps towards the vision of adaptive and domain-optimized computational cameras.

Acknowledgements

This project was in part supported by NSF Award 1839974, the NSF Graduate Research Fellowship, and a PECASE by the ARL.

References

[1] P.M. Acosta-Serafini, I. Masakiand and C.G. Sodini. A 1/3" VGA linear wide dynamic range CMOS image sensor implementing a predictive multiple sampling algorithm with overlapping integration intervals. *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference*, 2003. 3

[2] P.M. Acosta-Serafini, I. Masakiand and C.G. Sodini. Predictive multiple sampling algorithm with overlapping integration intervals for linear wide dynamic range integrating image sensors. *IEEE Transactions on Intelligent Transportation Systems*, 2004. 3

[3] Manoj Aggarwal and Narendra Ahuja. Split aperture imaging for high dynamic range. *International Journal of Computer Vision*, 58(1):7–17, 2004. 1, 3

[4] Masheal M Alghamdi, Qiang Fu, Ali Kassem Thabet, and Wolfgang Heidrich. Reconfigurable snapshot hdr imaging using coded masks and inception network. In *Vision Modeling and Visualization*, 2019. 1, 3

[5] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers. *Advanced High Dynamic Range Imaging: Theory and Practice*. AK Peters (CRC Press), 2011. 1

[6] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356, 2006. 3

[7] Ayush Bhandari, Felix Krahmer, and Ramesh Raskar. On unlimited sampling. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 31–35. IEEE, 2017. 4

[8] Ayush Bhandari, Felix Krahmer, and Ramesh Raskar. On unlimited sampling and reconstruction. *IEEE Transactions on Signal Processing*, 2020. 3, 4

[9] Matthew G. Brown, Justin Baker, Curtis Colonero, Joe Costa, Tom Gardner, Mike Kelly, Ken Schultz, Brian Tyrrell, and Jim Wey. Digital-pixel focal plane array development. In Manijeh Razeghi, Rengarajan Sudharsanan, and Gail J. Brown, editors, *Quantum Sensing and Nanophotonic Devices VII*, volume 7608, pages 765 – 774. International Society for Optics and Photonics, SPIE, 2010. 3

[10] Stephen J Carey, Alexey Lopich, David RW Barr, Bin Wang, and Piotr Dudek. A 100,000 fps vision sensor with embedded 535gops/w 256× 256 simd processor

array. In *2013 Symposium on VLSI Circuits*, pages C182–C183. IEEE, 2013. 2, 3, 7

[11] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. SIGGRAPH*, pages 369–378, 1997. 1, 2

[12] Piotr Dudek. Adaptive sensing and image processing with a general-purpose pixel-parallel sensor/processor array integrated circuit. In *Int. Workshop on Computer Architecture for Machine Perception and Sensing*, pages 1–6. IEEE, 2006. 3

[13] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017. 1, 3, 5, 7

[14] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177:1–177:10, 2017. 1, 3

[15] Gustav Theodor Fechner. Elements of psychophysics, 1860. 1948. 5

[16] Jorge Fernández-Berni, Ricardo Carmona-Galán, and Ángel Rodríguez-Vázquez. Flip-q: A qcif resolution focal-plane array for low-power image processing. In *Low-Power Smart Imagers for Vision-Enabled Sensor Networks*, pages 67–109. Springer, 2012. 3

[17] Eric R. Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The quanta image sensor: Every photon counts. *Sensors*, 16(8), 2016. 3

[18] D. C. Ghiglia and M. D. Pritt. *Two-dimensional phase unwrapping: theory, algorithms, and software*. Wiley, 1998. 3

[19] Mohit Gupta, Daisuke Iso, and Shree K. Nayar. Fibonacci exposure bracketing for high dynamic range imaging. In *2013 IEEE International Conference on Computer Vision*, pages 1473–1480, 2013. 1, 2

[20] Saghi Hajisharif, Joel Kronander, and Jonas Unger. Adaptive dualiso hdr reconstruction. *EURASIP Journal on Image and Video Processing*, 2015(1):41, 2015. 1, 3

[21] S. W. Hasinoff, F. Durand, and W. T. Freeman. Noise-optimal capture for high dynamic range photography. In *Proc. CVPR*, pages 553–560, 2010. 1, 2

[22] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 35(6):192, 2016. 1, 2

[23] IEEE Computer Society. Ieee standard for floating-point arithmetic. IEEE STD 754-2019, 2019. 2

[24] Atul Ingle, Andreas Velten, and Mohit Gupta. High flux passive imaging with single-photon sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6760–6769, 2019. 3

[25] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based High Dynamic Range Video. *ACM Trans. Graph. (SIGGRAPH Asia)*, 32(6), 2013. 1, 3

[26] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Trans. Graph. (SIGGRAPH)*, 22(3):319–325, 2003. 1, 3

[27] Peter M Knoll. Hdr vision for driver assistance. In *High-Dynamic-Range (HDR) Vision*, pages 123–136. Springer, 2007. 1

[28] Florian Lang, Tobias Plötz, and Stefan Roth. Robust multi-image hdr reconstruction for the modulo camera. In *German Conference on Pattern Recognition*, pages 78–89. Springer, 2017. 3

[29] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proc. ECCV*, pages 596–611, 2018. 1, 3

[30] Markus Loose, Karlheinz Meier, and Johannes Schemmel. A self-calibrating single-chip cmos camera with logarithmic response. *IEEE Journal of Solid-state circuits*, 36(4):586–596, 2001. 1, 2, 3

[31] Alexey Lopich and Piotr Dudek. An 80× 80 general-purpose digital vision chip in 0.18 μm cmos technology. In *IEEE Int. Symposium on Circuits and Systems*, pages 4257–4260, 2010. 3

[32] Mann, Picard, S. Mann, and R. W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of IS&T*, pages 442–448, 1995. 1, 2

[33] Rafat Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph. (SIGGRAPH)*, 30(4):40:1–40:14, 2011. 7

[34] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018. 1, 3

[35] J.N.P. Martel, L.K. Müller, S.J. Carey, P. Dudek, and G. Wetzstein. Neural Sensors: Learning Pixel Exposures for HDR Imaging and Video Compressive Sens-

ing with Programmable Sensors. *Proc. IEEE ICCP*, 2020. 1, 3, 9

[36] Julien NP Martel, Lorenz K Müller, Stephen J Carey, and Piotr Dudek. Parallel hdr tone mapping and autofocus on a cellular processor array vision chip. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1430–1433. IEEE, 2016. 3

[37] Mitsuhito Mase, Shoji Kawahito, Masaaki Sasaki, Yasuo Wakamori, and Masanori Furuta. A wide dynamic range cmos image sensor with multiple exposure-time signal outputs and 12-bit column-parallel cyclic a/d converters. *IEEE Journal of Solid-State Circuits*, 40(12):2787–2795, 2005. 3

[38] M. McGuire, W. Matusik, H. Pfister, B. Chen, J. F. Hughes, and S. K. Nayar. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, 27(2):32–42, 2007. 1, 3

[39] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum (Eurographics)*, volume 28, pages 161–171. Wiley Online Library, 2009. 1, 2

[40] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020. 1, 3

[41] Laurence Meylan, Scott Daly, and Sabine Süsstrunk. The reproduction of specular highlights on high dynamic range displays. *IS&T/SID 14th Color Imaging Conference (CIC)*, 2006. 3

[42] Wei Miao, Qingyu Lin, Wancheng Zhang, and Nan-Jian Wu. A programmable simd vision chip for real-time vision applications. *IEEE Journal of Solid-State Circuits*, 43(6):1470–1479, 2008. 3

[43] Shree K Nayar and Vlad Branzoi. Adaptive dynamic range imaging: Optical control of pixel exposures over space and time. In *IEEE Int. Conference on Computer Vision (ICCV)*, page 1168, 2003. 1, 3

[44] Shree K Nayar, Vlad Branzoi, and Terry E Boult. Programmable imaging: Towards a flexible camera. *Int. Journal of Computer Vision*, 70(1):7–22, 2006. 1, 3

[45] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 472–479, 2000. 1, 3

[46] Jun Ohta. *Smart CMOS image sensors and applications*. CRC press, 2020. 1

[47] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. 4, 14

[48] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 1

[49] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: On-the-fly reverse tone mapping of legacy video and photographs. *ACM Trans. Graph. (SIGGRAPH)*, 26(3), 2007. 3

[50] Jehyuk Rhee, Dongwon Park and Youngjoong Joo. Analysis and Design of a Robust Floating Point CMOS Image Sensor. *IEEE Sensors Journal*, 2009. 3

[51] Angel Rodríguez-Vázquez, Rafael Domínguez-Castro, Francisco Jiménez-Garrido, Sergio Morillas, Alberto García, Cayetana Utrera, Ma Dolores Pardo, Juan Listan, and Rafael Romay. A cmos vision system on-chip with multi-core, cellular sensory-processing front-end. In *Cellular nanoscale sensory wave computing*, pages 129–146. Springer, 2010. 3

[52] Mushfiqur Rouf, Rafał Mantiuk, Wolfgang Heidrich, Matthew Trentacoste, and Cheryl Lau. Glare encoding of high dynamic range images. In *Proc. CVPR 2011*, pages 289–296, 2011. 1, 3

[53] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Trans. Graph. (SIGGRAPH)*, 39(4), 2020. 1, 3

[54] Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. High dynamic range display systems. *ACM Trans. Graph. (SIGGRAPH)*, 23(3):760–768, 2004. 1

[55] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph. (SIGGRAPH Asia)*, 31(6):203:1–203:11, 2012. 1, 3

[56] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum*, volume 35, pages 153–163, 2016. 1, 3

[57] Viraj Shah and Chinmay Hegde. Signal reconstruction from modulo observations. In *2019 IEEE Global Con-*

ference on Signal and Information Processing (GlobalSIP), pages 1–5. IEEE, 2019. 3

[58] Viraj Shah, Mohammadreza Soltani, and Chinmay Hegde. Reconstruction from periodic nonlinearities, with applications to hdr imaging. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 863–867. IEEE, 2017. 3

[59] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1386–1396, 2020. 1, 3

[60] Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. *ACM Trans. Graph. (SIGGRAPH)*, 30(4):41:1–41:10, 2011. 1, 3

[61] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in hdr deghosting: a survey and evaluation. In *Computer Graphics Forum*, volume 34, pages 683–707. Wiley Online Library, 2015. 3

[62] Brian Tyrrell, Kirk Anderson, Justin Baker, Robert Berger, Matthew Brown, Curtis Colonero, Joseph Costa, Brian Holford, Michael Kelly, Eric Ringdahl, Kenneth Schultz, and James Wey. Time delay integration and in-pixel spatiotemporal filtering using a nanoscale digital cmos focal plane readout. *IEEE Transactions on Electron Devices*, 56(11):2516–2523, 2009. 3

[63] R Wagner, Ákos Zarándy, and Tamás Roska. High dynamic range perception with spatially variant exposure. In *IEEE Int. Workshop*, 2004. 3

[64] Gordon Wetzstein, Ivo Ihrke, and Wolfgang Heidrich. Sensor saturation in fourier multiplexed imaging. In *Proc. CVPR*, pages 545–552. IEEE, 2010. 1, 3

[65] David XD Yang, Abbas El Gamal, Boyd Fowler, and Hui Tian. A 640/spl times/512 cmos image sensor with ultra wide dynamic range floating-point pixel-level adc. In *IEEE Solid-State Circuits Conference*, pages 308–309, 1999. 3

[66] Ákos Zarándy. *Focal-plane sensor-processor chips*. Springer Science & Business Media, 2011. 3, 7

[67] Wancheng Zhang, Qiuyu Fu, and Nan-Jian Wu. A programmable vision chip based on multiple levels of parallel processors. *IEEE Journal of Solid-State Circuits*, 46(9):2132–2147, 2011. 3

[68] H. Zhao, B. Shi, C. Fernandez-Cull, S. Yeung, and R. Raskar. Unbounded high dynamic range photography using a modulo camera. In *Proc. ICCP*, pages 1–10, 2015. 1, 2, 3, 5, 6, 9, 14

[69] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Unmodnet: Learning to unwrap a modulo image for high dynamic range imaging. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 4, 5, 6, 9, 14

Supplemental Material
MantissaCam: Learning Snapshot High-dynamic-range Imaging with Perceptually-based In-pixel Irradiance Encoding

S1. Pipeline Details

S1.1. Mantissa Dataset creation

There are several ways to encode the mantissa. When working with synthetic data, the simplest way is to just take the log of the signal and then take the modulo of the resulting value. Recall log of anything below 1 is a negative value, which would not be conceivable with the hardware. Instead, only after the pixel saturates do we take the log to simulate the mantissa. When we saturate the pixel, the subsequent wrap will require twice the intensity to wrap again. To create the training dataset, we create the mantissa image and the corresponding winding number image. For each pixel ij,

$$\text{mantissa}_{ij} = \begin{cases} I_{ij}, & \text{if } I_{ij} < I_{max} \\ \log_\alpha(I_{ij}) \% I_{max}, & \text{otherwise.} \end{cases} \quad (1)$$

$$\text{winding number}_{ij} = \begin{cases} 0, & \text{if } I_{ij} < I_{max} \\ \lfloor \log_\alpha(I_{ij}) \rfloor + 1, & \text{otherwise.} \end{cases} \quad (2)$$

where $\%$ denotes the modulo operation and $\lfloor \cdot \rfloor$ denotes the floor function. For our dataset and experiments, we set $I_{max} = 1$ and $\alpha = 2$.

S1.2. Network Architecture

In this subsection, we describe the architecture for the single pass winding number prediction network (also see Figure S1). The extracted edges from the edge prediction network, along with the mantissa image, are fed into the network via feature extraction by a 7×7 convolutional layer, an instance norm, ReLU, and a non-local block for the extracted edge features. These images are then concatenated and sent through a squeeze-and-excitation block to perform dynamic channel-wise feature recalibration. The base network is an attention unet, pioneered by Oktay et al. [2]. The backbone is the U-Net where the expanding path has attention gates added, along with the skip connections. Skip connections allow features extracted from the contracting path to be used in the expanding path. The attention block places more emphasis on the features of the skip connections.

S2. Training Procedures

S2.1. RGB training on HDR images

For training our network for RGB, we trained the edge network for 400 epochs on the synthetic data at a learning rate of .0001 using an ADAM optimizer in Pytorch. From a dataset of 593 images, we randomly split it into 400 training images and 193 test images. We augment the training images by scaling the HDR image and calculating the corresponding mantissa and winding numbers.

S2.2. Training Procedure for SCAMP-5 Prototype

We retrain the edge prediction network for the captured grayscale dataset as described in the paper. Both UnModNet and our method use the same edge prediction network. The other parts of the respective pipelines are retrained on the captured dataset using a similar procedure as used for the synthetic data described above.

S2.3. Baseline Comparison

Graph Cuts Graph Cuts was implemented following the original ModuloCam paper [4] using the same custom potential function. Reaching out to the authors confirmed the method, which can be successful for some clearly wrapped modulo images, however requires delicate parameter tuning for each of the many layer unwrappings of each image. We chose a set of parameters to best unwrap the whole set. PSNR, SSIM, and MSSIM scores were comparable to those found in UnModNet [5] when they implemented the MRF algorithm.

Modulo Encoding with UnModNet We retrained UnModNet, the state of the art for unwrapping modulo images, with the same training process and same dataset as in [5] and results were comparable to those reported in the paper. In areas of dense wrappings, the pipeline struggles to stop unwrapping, leading to patches of white.

Mantissa Encoding with UnModNet One of our baseline experiment is to use the pipeline of UnModNet with the forward imaging model of MantissaCam. We trained the pipeline using the same training procedure as described above. We noted the layer-by-layer unwinding did not work well with the reconstruction from the mantissa encoding as errors in winding number manifest in exponentially bad errors. Indeed, missing a wrap results in much worse errors in MantissaCam (because of the exponential function used when reconstructing) than in ModuloCam, resulting in huge artifacts. Besides, the nature of the layer-by-layer unwrapping is prone to propagating errors.

Modulo Encoding with Our Network To combat propagation of errors in unwinding, we directly predict the wind-

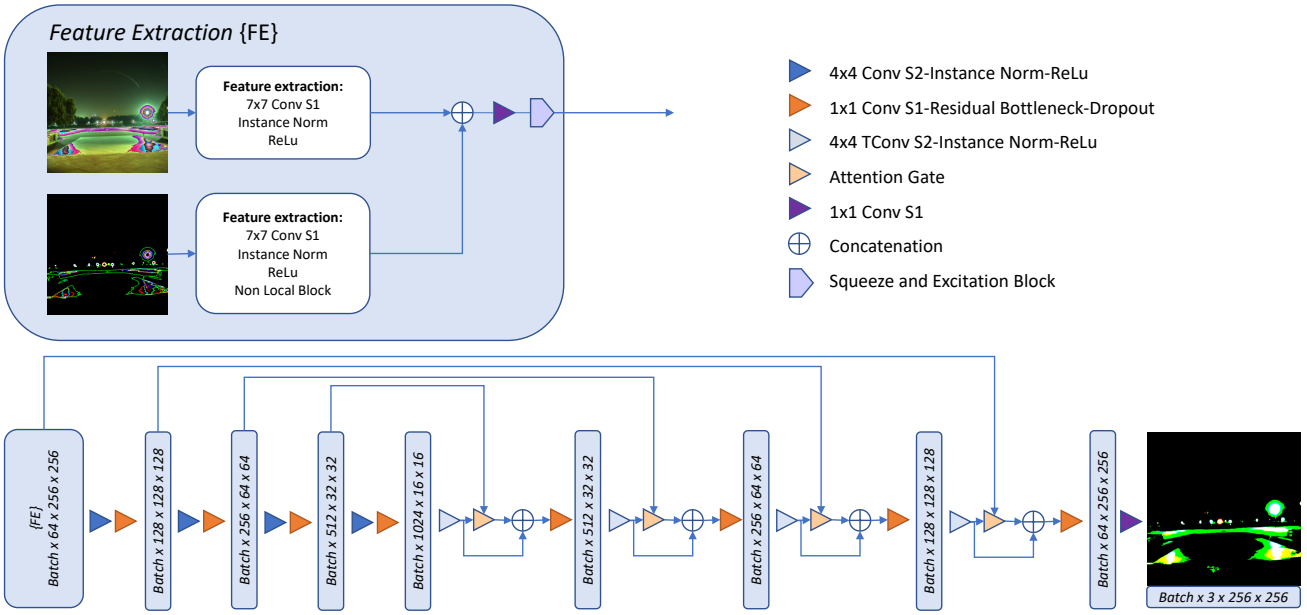

Figure S1: Attention UNet architecture of the winding number prediction network.

ing number in a single pass through an attention-unet instead of predicting a mask. Again, we train using the same training procedure as UnModNet. Results are promising, however, the network still struggles when the modulo image has very tight wrappings (of the order of 1–2 pixels width).

Mantissa Encoding with Our Network Introducing the mantissa allows us to spatially spread out the wraps as we get to higher and higher irradiance levels. This leads to preservation of more detail. Results comparing these methods, excluding the UnModNet for the mantissa, are shown in the paper and in the additional figures in Section 4.

S2.4. Additional Implementation Details

We compare the full reconstructed HDR image with the ground truth HDR image to calculate PSNR and Q-Score (2). We then tonemap both the ground truth and the predicted HDR images, all using the Reinhard Algorithm with gamma = 1, intensity = 1. The tonemapped images are then compared to calculate SSIM and MSSIM values. Inference time for our method is much faster than for the UnModNet or graph cuts due to the single pass architecture, as opposed to the iterative unwrapping that can unwrap as high as the default of 15 max iterations.

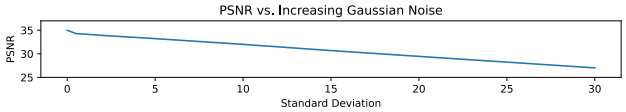

Figure S2: my caption of the figure

S3. Additional Details of Experimental Results

Currently, mantissa images cannot be directly captured in SCAMP-5. However we implemented a procedure on SCAMP-5 to capture modulo images as described in the main paper.

We also implemented a bracketed exposure procedure directly on the camera in order to get reference HDR images. Exposure bracketing is performed by capturing 5 images doubling the exposure time between each exposure, starting from a configurable short exposure time.

S4. Additional Results

See Figures S3 and S4 for additional results. From left to right, each row shows the modulo image, the graph cuts method, UnModNet + modulo, Ours + modulo, the mantissa image, Ours + mantissa, and the ground truth image. All tonemapped images follow the tone-mapping described in Section 3. Additionally, we performed a study on the

effects of noise. Our networks and comparisons were not trained on noisy images, so as we increase additive Gaussian noise, the PSNRs decrease, as shown in figure S2. However, if the networks are trained with real data, they are able to capture the effects of noise, as demonstrated by the results from our reconstruction algorithm on the captured images with our prototype. Figures S5–S9 show additional results for captured data with the SCAMP-5.

References

[1] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. SIGGRAPH*, pages 369–378, 1997. 1, 2

[2] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. 4, 14

[3] Sylvain Paris, Samuel W Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Trans. Graph.*, 30(4):68, 2011.

[4] H. Zhao, B. Shi, C. Fernandez-Cull, S. Yeung, and R. Raskar. Unbounded high dynamic range photography using a modulo camera. In *Proc. ICCP*, pages 1–10, 2015. 1, 2, 3, 5, 6, 9, 14

[5] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Unmodnet: Learning to unwrap a modulo image for high dynamic range imaging. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 4, 5, 6, 9, 14

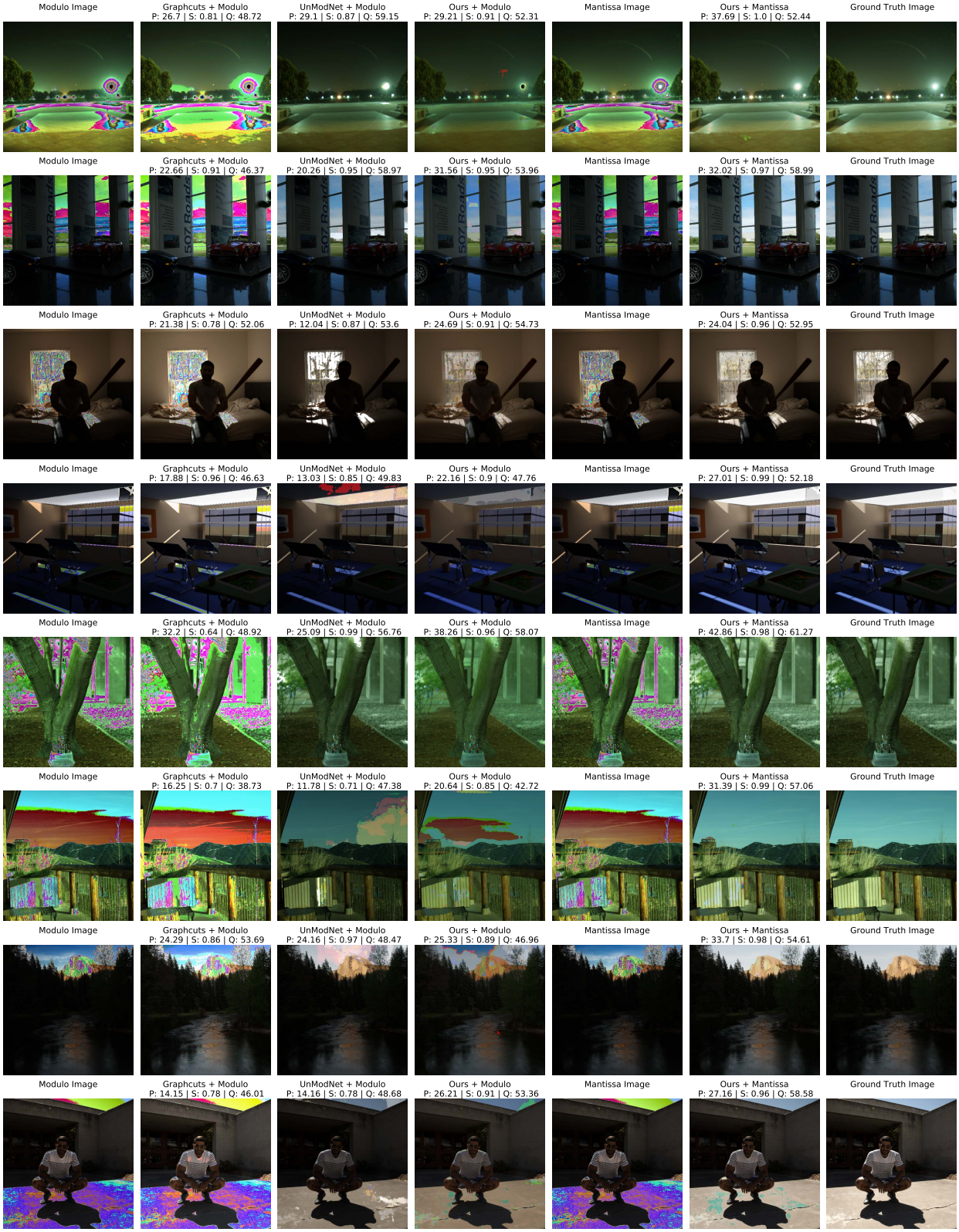

Figure S3: More results showing the comparison between different baselines and encodings. Ours + mantissa is better able to keep details in the high intensity areas. PSNR (P), SSIM (S), and Q-Scores (Q) are shown about the reconstructed images.

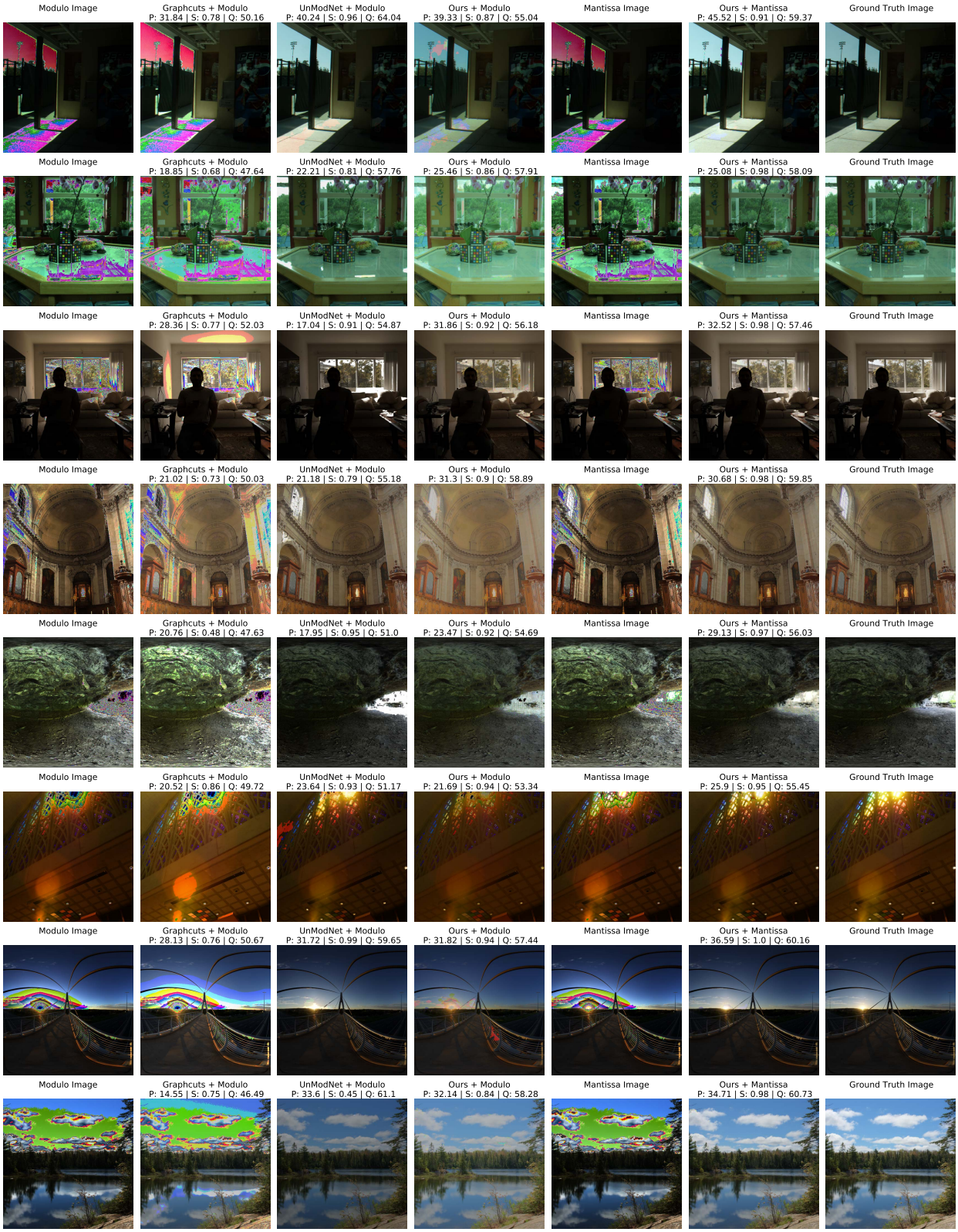

Figure S4: More results comparing the different reconstruction and encoding methods.

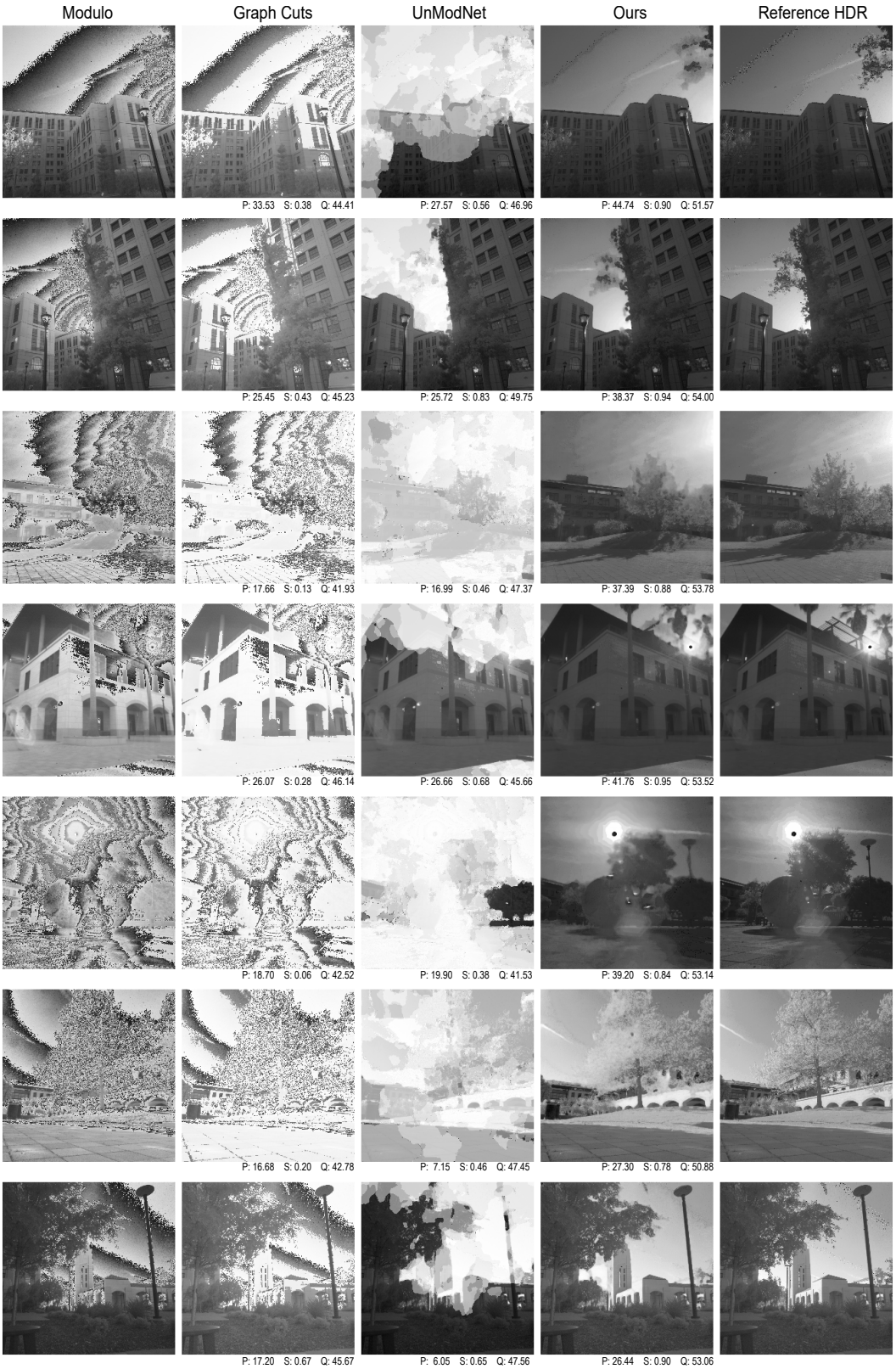

Figure S5: Comparisons on captured data.

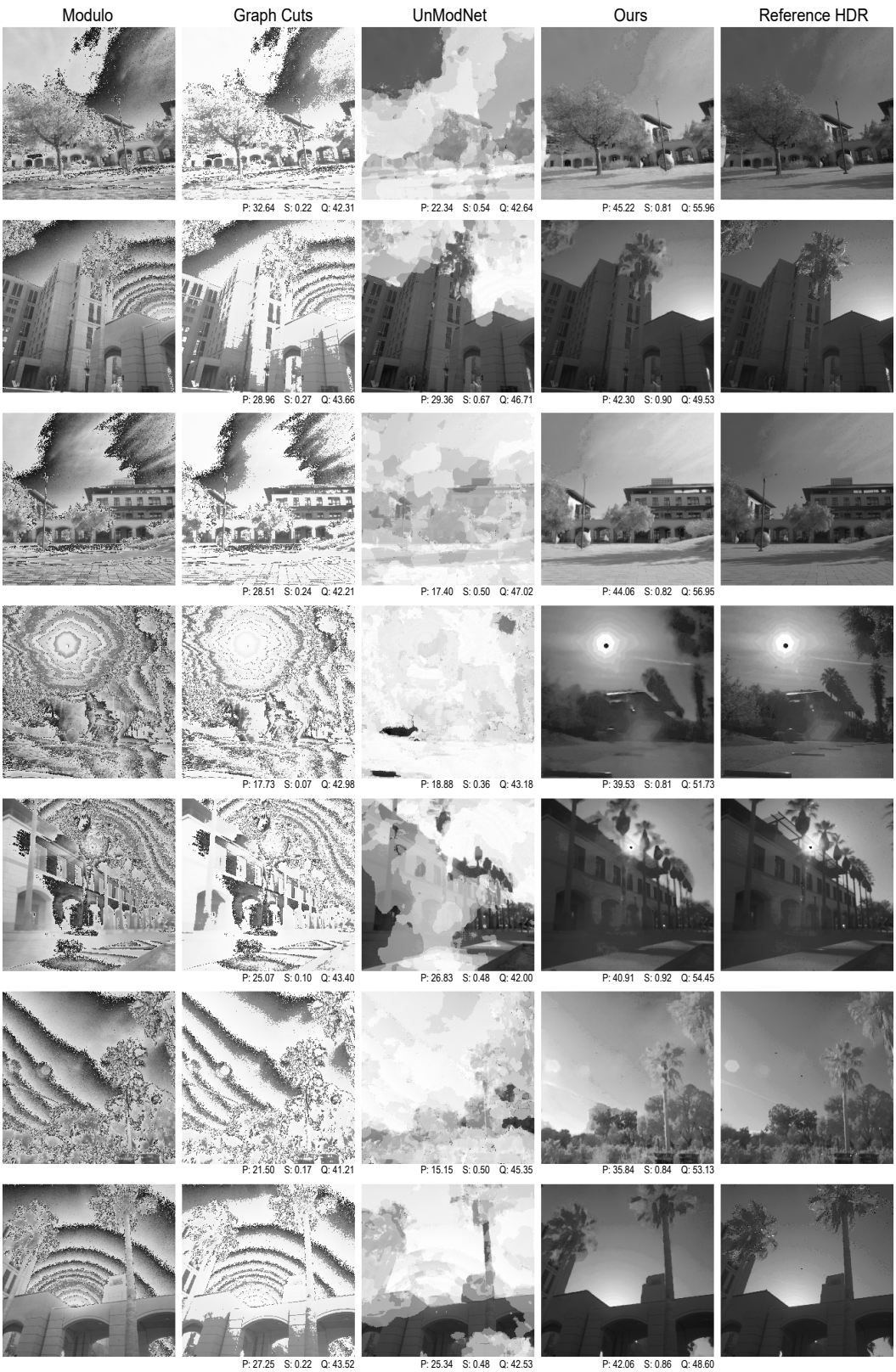

Figure S6: Comparisons on captured data.

Figure S7: Comparisons on captured data.

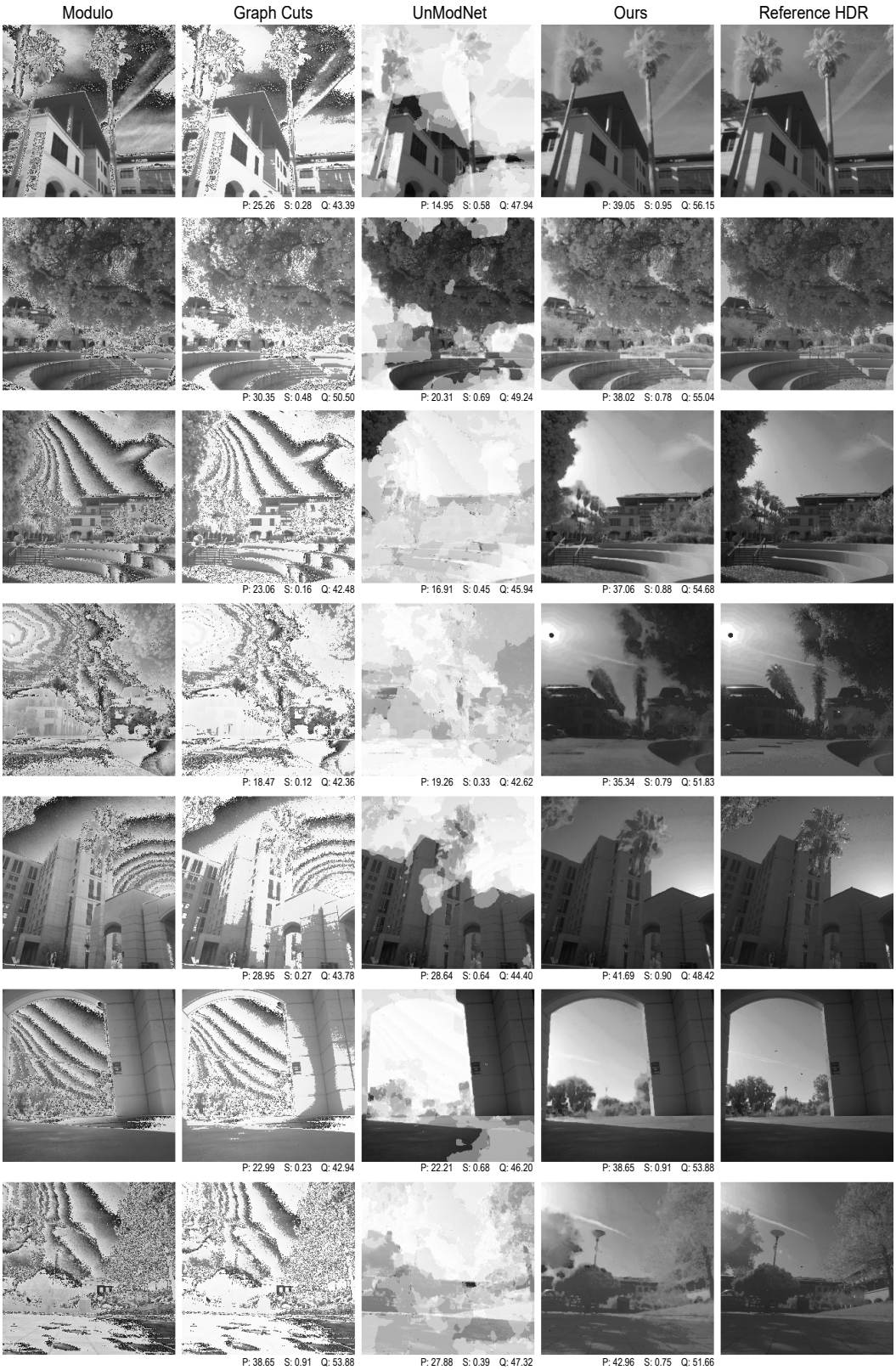

Figure S8: Comparisons on captured data.

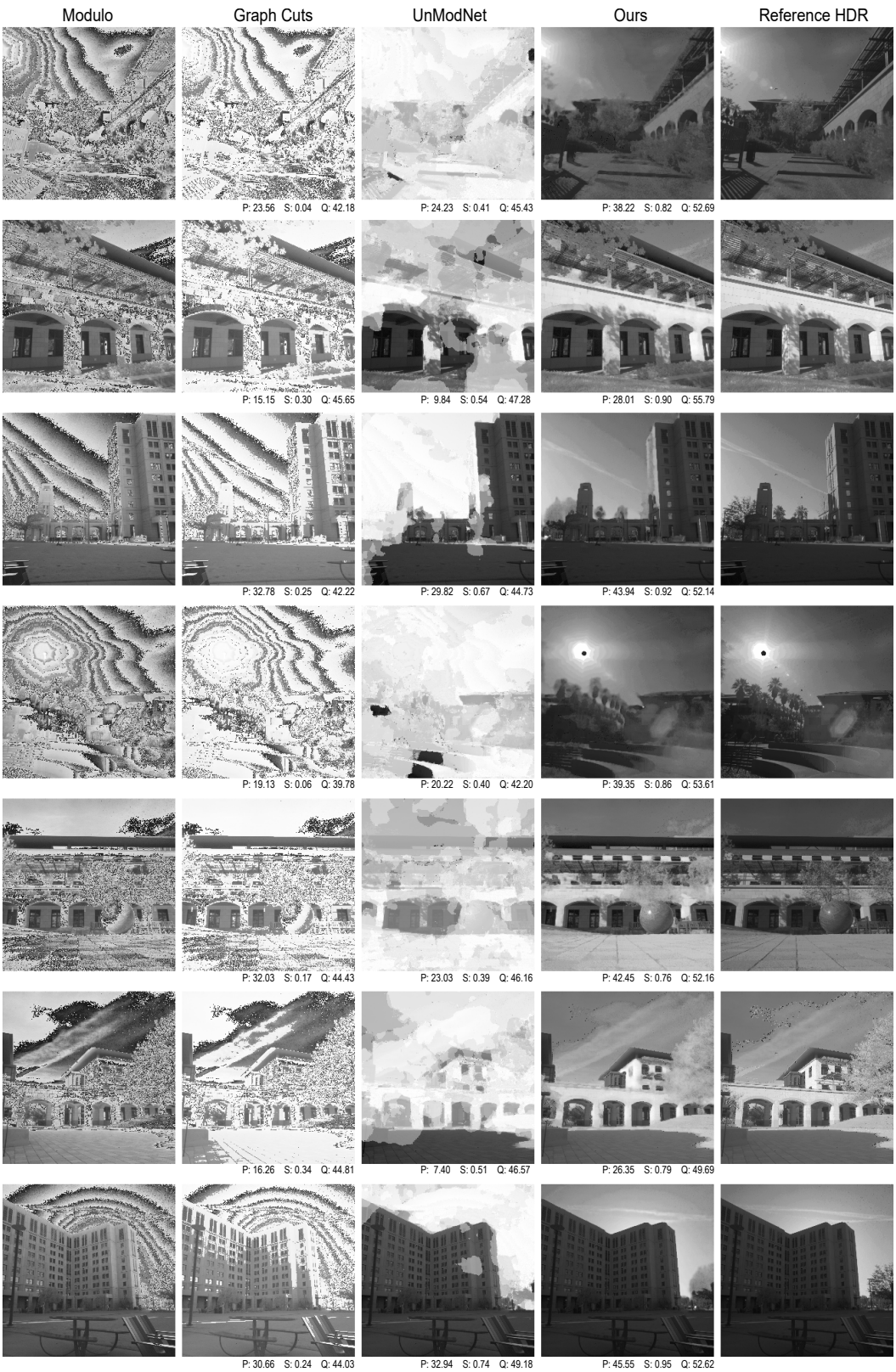

Figure S9: Comparisons on captured data.