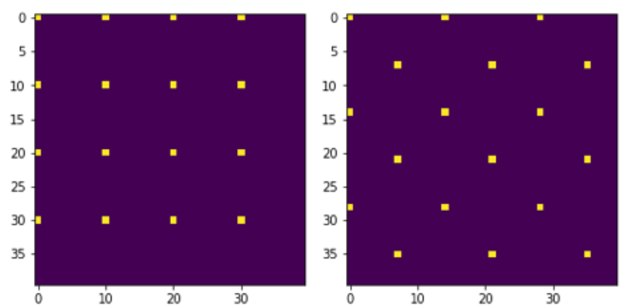
https://doi.org/10.2352/ISSN.2470-1173.2020.7.ISS-229

Expanding dynamic range in a single-shot image through a sparse grid of low exposure pixels

Leon Eisemann; Stuttgart Media University (HdM); Stuttgart, Germany
Jan Froehlich; Stuttgart Media University (HdM); Stuttgart, Germany
Axel Hartz; Stuttgart Media University (HdM); Stuttgart, Germany
Johannes Maucher; Stuttgart Media University (HdM); Stuttgart, Germany

Abstract

Camera sensors are physically restricted in the amount of luminance which can be captured at once. To achieve a higher dynamic range, multiple exposures are typically combined. This method comes with several disadvantages, like temporal or alignment aliasing. Hence, we propose a method to preserve high luminance information in a single-shot image. By introducing a grid of highlight preserving pixels, which equals 1% of the total amount of pixels, we are able to sustain information directly in-camera for later processing. To provide evidence, that this number of pixels is enough for gaining additional dynamic range, we use a U-Net for reconstruction. For training, we make use of the HDR+ dataset, which we augment to simulate our proposed grid. We demonstrate that our approach can preserve high luminance information, which can be used for a visually convincing reconstruction, close to the ground truth.



(a) LE-pixel with $k = 10$ (b) LE-pixel with $k = 10$ rotated 45°

Figure 1. *The used grids in this paper in figure (a) the standard grid as proposed, and the same grid rotated* 45° *(b). The rotated distance is calculated through Pythagorean theorem, then rounded into* $\mathbb{N}$.

Introduction

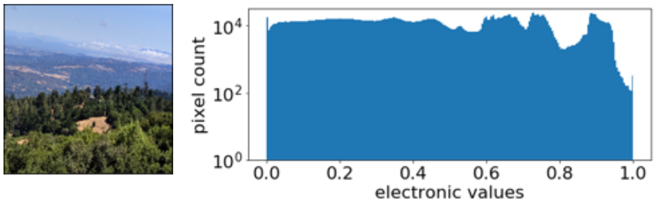
Current camera sensors are physically restricted in the amount of photons, which can be captured by the full well capacity of the individual sensor elements. This results in an overall limited luminance range for the whole sensor, which is usually controlled by aperture and electronic gain. Especially in the field of HDR creation this limitation is problematic, as the dynamic range of most sensors is smaller than the dynamic range of the scene to capture. To overcome this limitation several solutions were introduced. Firstly HDR-Burst, where a certain amount of images with different increasing exposures are shot in a short timespan to reduce the amount of photons per exposure and hence gain a high virtual full well capacity for the final exposure, which is one of the most commonly used methods [1]. These images are then aligned and processed to combine the contained information [1]. Similarly methods include RED Digital Cinemas HDRx, which combines one normal and one short exposure frame to gain additional information in the highlights [2] and DualISO, where the dynamic range is extended through different AD converter or denoising algorithms to get a low-noise signal to preserve high- and lowlight information [3]. Another approach is to combine two camera sensors, as proposed by *Tocci et al.[4]* and applied by *Froehlich et al.[5]* who used a stacked camera system, one for low- and one for high luminance, to preserve a high dynamic range, which is then later combined in post-production. Although these techniques show impressive results, they come with several disadvantages, like temporal aliasing, the need to align the individual images / videos and longer processing times or enormous costs and camera system size increase, in case of *Froehlich et al.[5]*. On the other side of the spectrum are technologies, for example the work of *Hasinoff et al.[1]* which propose a Hybrid Dynamic Range Autoencoder for predicting HDR Images based on LDR input, or the work of *Banterle et al.[6]*, which used inverse ToneMapping to achieve the same goal.

Methods

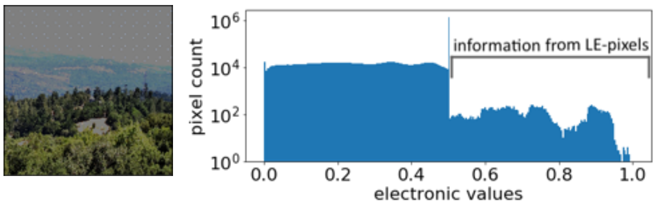
To solve some of the mentioned disadvantages, we propose a different method with the aim to preserve more information in high luminance image areas for later processing and enable sensors to gain a higher dynamic range. Similar to other approaches we also adapt different exposures / sensor sensitivity to preserve high luminance information. In contrast to related techniques we propose a system, that does not concatenate the sensor temporarily or physically for capturing more information. Instead our method is based on an $k \times k$ grid of pixels, as shown in figure 1, which are altered on a sensor to be less light sensitive and as a result highlight preserving. As we also tried a 45° rotation of the grid, it was therefor calculated as

$$\begin{aligned} &\text{grid}^{\text{rot}} = k^{\text{rot}} \times k^{\text{rot}} + (1.5 * k^{\text{rot}}) \times (1.5 * k^{\text{rot}}) \\ &\text{with} \quad k^{\text{rot}} = k * \sqrt{2} \quad \text{and} \quad k^{\text{rot}}, k \in \mathbb{N} \end{aligned} \tag{1}$$

Due to technical limitations k^{rot} was rounded to $\mathbb{N}$.
With $k = 10$, which is mainly used in this paper, only every $100th$ pixel is modified, which equals 1% of the total amount of pixels. This value for k was chosen as it was found in internal studies at HdM, that images with an amount of 1% replaced dead pix-



(a) Augmented ground truth image



(b) Augmented image with $k = 10$, clipped with $\beta = 0.5$

Figure 2. *Comparison of the ground truth in figure (a) against the augmented network input, with our proposed low exposure pixel grid, rotated 45°(b). In (b) all values above 0.5 therefore derive from our low exposure grid. For our proof of concept, the images were saved as compressed jpegs and then reloaded at a later point to calculate the histograms.*

els can not be distinguished from the original images, for typical 4K+ spatial resolution and viewing conditions of up to 1.5 picture heights. Hence, it is an amount which a manufacturer can sacrifice without problem, because even if the technique can not provide the desired result, the pixels could be marked as dead without a significant impact on image quality.

As the grid is applied by means of a filter on the sensor, its information can be captured at the same time of exposure, without temporal artefacts, which makes it suitable for motion picture applications. On top of that, the image characteristic of the remaining pixels is not impacted and the low exposure (LE) pixels' solely purpose is to gain additional dynamic range. Moreover this method can be easily applied to sensors, as there is no significant change to the pixel design necessary. The required grid structure could be applied by adding a new filter layer, by changing the shape of the micro-lenses or by covering parts of the pixel. This allows manufacturers to deploy this method in a fast and easy way and deliver the reconstruction logic later on, as algorithmic approach for their cameras over software updates or in the proposed way as SDK addition.

For testing this hypothesis and to provide empirical evidence, that a small amount of low exposure pixels can boost the performance of highlight reconstruction and thus extend the dynamic range of sensors, we augment images to simulate a small full well capacity, add our theoretic LE-Pixels and attempt to gain the missing information back. As U-Nets have shown impressive results in reconstruction challenges, we utilize a modified version of it. The encoder part of an U-Net, see figure 4, is trained to transform an high dimensional input to a low-dimensional representation, which the decoder part is trained to reconstruct again [7]. Through the low dimensional latent representation the U-Net automatically has to perform a feature engineering while training, which allows us a faster and more flexible experimentation [7, 8]. Additionally, U-Nets allow us to offer an automated pro-

(a) Input (b) without skip (c) with skip (d) ground truth

Figure 3. *Magnification of reconstructions without (b) and with (c) skip-connections. The autoencoder is able to reconstruct highlevel information, but without the skip-connections detail information in higher frequency regions is missing and prediction is blurred.*

cess without user-defined hyperparameters. Taken into account that convolutional networks are not fully predictable and results sometimes vary, we build our baseline, the reconstruction without LE-pixels, with the exact same network and training parameters as the reconstruction with altered pixels. Thereby we ensure that both methods have the same initial base for recreating the original image. To make the results as comparable as possible, randomizations are done with the same seed. Networks are trained for the same number of epochs, where the weights of the epoch with the lowest validation failure is taken. In this way we ensure to compare the best result of each method against each other. Since we were bound to a limited timeframe as well as technical equipment we could not modify a real camera, so we simulated a camera through data augmentation, as described in the following section.

Image Dataset

One of the key challenges for a learning based highlight reconstruction with a decent resolution is to gather a large enough dataset. Especially in our case, the resolution was a keypoint to simulate a realistic camera and grid structure. In contrast to similiar work, e.g. *Eilertsen et al.[9]*, who collected a large HDR dataset as their basis, we used SDR data due to the fact that aggregation and curation of HDR data is a time consuming process. Based on the assumption that our theory works independent of color / quantization domains, we do not apply any domain transformations and only simulate the information loss. For a fast acquisition of a large dataset, which contains a sufficient amount of information in high luminance regions, the HDR+ Dataset by *Hasinoff et al.[1]* was chosen. The dataset consists of 3640 images total, which were created through exposure stacking and were then tonemapped into sRGB domain [1]. As the number of images is not sufficient enough for training a good generalizing network, the data was additionally augmented. Therefor the images were transformed and flipped, with reflection of the content at the image border to prevent blank spaces. As it was important to keep a realistic imaging, no color-shift or noise were applied in the augmentation. The simulation of LE-Pixels was then applied onto the augmented data so the grid keeps a static position while the scenery changes, as it would be in a real world camera.

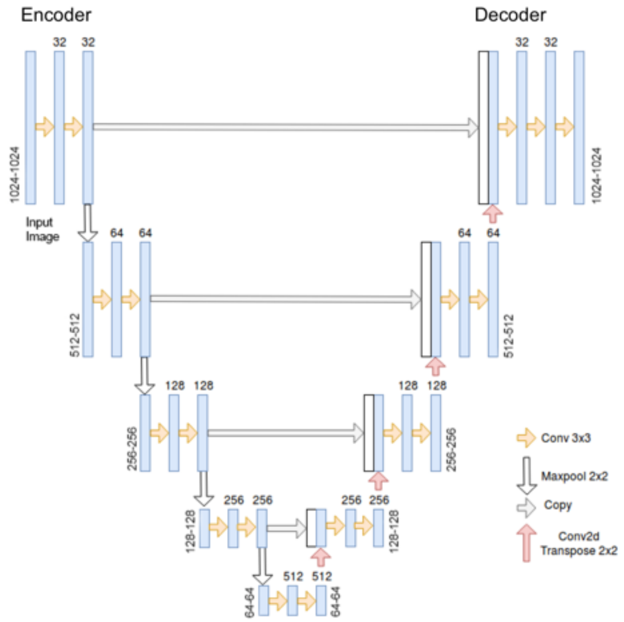


Figure 4. *U-Net used in this paper. Skip-connections are used to transfer data from the encoder to the decoder to boost the performance of the network. This architecture is not restricted to a fixed resolution, therefrom the layer resolutions are given based on the 1024×1024 images, which where used in training.*

For creation of the artificial dataset, all images were processed in sRGB Domain with the following function:

$$\begin{aligned} &\mathrm{img}^{\mathrm{dualiso}} = \mathrm{mask} \times \mathrm{img}^{\mathrm{org}} + (1-\mathrm{mask}) \times \mathrm{img}^{\mathrm{clip}} \\ &\text{with} \quad \mathrm{img}^{\mathrm{clip}} = \min(\beta, img^{\mathrm{org}}) \quad \text{and} \quad \beta = 0.5 \end{aligned} \tag{2}$$

Consequently the non LE-pixels $\mathrm{img}^{\mathrm{clip}}$ where clipped at half their electronic values (EV), as shown in figure 2b. We are aware of the fact that pixels under different exposures show different noise levels, as it is dependent on the signal level [1]. Since this is not our focus in this proof of concept work and autoencoders have shown impressive result in denoising challenges, we therefore do not apply any additional noise [7, 9].

Network Structure

As our focus was to provide a universal approach of recovering the preserved information, we propose a generic U-Net for this process. Referring to *Eilertsen et al.[9]*, our design does not make use of a fully connected layer for the latent representation and instead uses a multichannel low resolution representation of the input data [9]. Therefor this fully convolutional network (FCN) approach is resolution independent, as long as the input dimension is a multiple of the encoder downscaling factor [9]. Since the latent representation of the network is defined as $\frac{width^{\mathrm{input}}}{16} \times \frac{height^{\mathrm{input}}}{16} \times 512$ the resolution must be a multiple of 16. The down-conversion to the latent representation inside an U-Net means that high resolution information is lost and not usable in the decoder, wherefore predictions are lacking them [9]. To overcome this limitations skip-connections are used in

(a) MSE (b) Lin + MSE (c) $\mathrm{L}^{\mathrm{mix}}$ (d) $\mathrm{L}^{\mathrm{mix}}$ 45°

Figure 5. *7× Magnification of the reconstruction using the MSE (a), with the grid structure still visible as it yields no error. The linear interpolation layer (b) improves the results, but generates dark spots. The mixed loss function (c) shows nearly no grid structure left in the prediction, for the normal (c) as well as the 45° rotated grid (d). All images were sharpened with the same settings for better visibility of the LE-Pixel visibility problem.*

U-Nets, which transfer information from the encoder into the decoder directly [10]. Our particular network structure adds skip-connections between layers with the same spatial resolution in both the encoder and decoder [11]. To achieve this, the output of the encoder layer is concatenated along the feature axis of the decoder layer [10]. For a given layer in the encoder with $W \times H \times K$ the resulting decoder layer has the shape $W \times H \times 2K$ [9]. These additional feature maps are then reduced by the decoder in the next convolution step [10, 11, 9]. An example of the impact of the introduced skip-connections is displayed in figure 3. Adding the information transfer enables the network for a better reconstruction of high frequency information. Our final structure, as shown in figure 4, is mostly inspired by the work of *Mansar [11]* and *Ronneberger et al. [10]* as this approach showed good performance in image denoising and high flexibility between use cases [11, 10]. Our ambition is to provide evidence and give a universal approach, thus we do not introduce special domain transfer or use case specific changes to the final network in contrast to *Eilertsen et al.[9]* or *Park et al.[12]*. Furthermore, to present a convenient and replicable system, we used the keras functional API with tensorflow as back-end[1].

The firsts results of the network contained excessive artefacts from the LE-Grid, especially in the highlights of the image, as seen in figure 5. This originates from a combination of mean squared error (MSE) and max pooling, as the LE-Pixels inherit the highest values, causing them to overweigh in the max pooling layers. In addition, the LE-Pixels inherit the unchanged original values and as a consequence have no contribution to the MSE. To overcome this problem we temporarily introduced a specialized linear interpolation layer to mask the LE-Pixels in the network output. As this did not produce the desired results, as can be seen in the following section and figure 5, this layer has been discarded and is no longer utilized in the final notebook.

Loss Function

Despite the main goal of restoring information in high luminance areas we decided against a cost function that is formulated in linear quantization domain. One of the reasons was that the available training data contained huge variation above our defined clipping point, which in linear domain would have led to an unsteady cost estimation [9]. Additionally it would result

[1]The network is back-end independent, except the loss function which depends on tensorflow for SSIM

(a) Input (b) Baseline rec. w/o LE-Pixels (c) Without skip with LE-Pixels (d) Linear Interpolation and MSE (e) Ours, Mixed Loss Grid 45° rotated (f) Ground Truth

Figure 6. *Comparison of some of the methods mentioned in this paper, except for baseline (b) all methods could use the additional information to restore huge amount of the structure in the clipped regions.*

in an underestimation in lower range luminance [9]. To provide a general cost function for reconstruction we initially used MSE between the network output and the unaltered original image. In the process of experimentation it became apparent that the network predicted inaccurate color temperatures in the highlights of the images, while using MSE as loss function. Especially in skies the network resulted in a warmer color than the original image. In addition even after intensive experimentation with training parameters, the network still had visible problems with the grid structure, which was still shown in the output. Despite introducing a linear interpolation layer, the grid was noticeable in homogeneous areas and the reconstruction of image structures was lagging behind. To put more weight on visual perception as well as stabilizing color reconstruction, the loss function was altered to a combination of multi-scale structural similarity (MS-SSIM) and mean absolute error (MAE/L1) as recommended by *Zhao et al.[13]*.

The new loss L^{mix} was then calculated with

$$L^{mix} = \alpha \times (1 - L^{MS\text{-}SSIM}) + (1-\alpha) \times L^{L1} \quad (3)$$

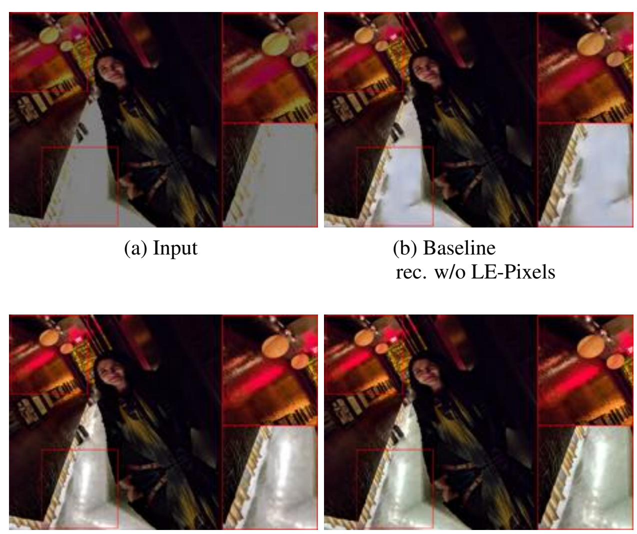
with $\alpha = 0.84$ as proposed by the original paper [13]. As shown in figure 5, the changed loss function is capable of removing the grid artefacts in the output and stabilizing the reconstruction of color temperatures.

Metric		*Input*	*Baseline*	*LE*	*LE* 45°
PSNR	mean	19.3359	26.8580	31.0810	31.2140
	std.	4.3816	3.4907	3.6427	3.6648
SSIM	mean	0.8968	0.9488	0.9541	0.9549
	std.	0.0639	0.0315	0.0304	0.0301
MSE	mean	0.0159	0.0026	0.0010	0.0010
	std.	0.0109	0.0018	0.0007	0.0007

Table 1. *Comparison of different metrics evaluated over the validation set. Metrics where calculated on per image basis against ground truth, which where then averaged.*

Training

Training of the networks was performed with the ADAM optimizer, with a learning rate of 1e-4. The networks where trained for 10 epoches using a mini-batchsize of 2 due to limitations in processing power. Training with this settings takes around 6 hours on a NVIDIA GTX 1080. For a more flexible training process the keras built-in, ReduceLROnPlateau was used for adaptive learning-rate reduction when necessary. The reconstruction interference time is about 180ms, which makes the U-Net approach currently to slow for real-time applications. As mentioned ear-



(a) Input (b) Baseline rec. w/o LE-Pixels

(c) Ours Mixed Loss, Grid 45° rotated (d) Ground truth

Figure 7. *Zoom-ins of reconstructions of complex light situation. As the baseline (b) is not able to estimate the structure, as well as the luminance. Our method (c) in contrast can not only estimate the correct brightness of the highlights, but also reconstruct the structures close to the ground truth (d).*

(a) Input (b) Prediction (c) Ground truth

Figure 8. *7× Magnification of reconstruction. The prediction of our method (b) can not reconstruct all diagonal structures in the image, due to undersampling of the our proposed grid structure.*

lier an algorithmic approach would be more suitable in real world camera systems, considering the resources available on embedded hardware.

Results

In this section we present a number of examples, to verify the quality of our proposed method. Additional examples can be found in the supplementary material and on the project website [14]. Furthermore the jupyter notebooks which contain all code needed for training and prediction, as well as the corresponding CNN weights can be downloaded from: `https://github.com/leisemann/low-exposure-pixel-grid`

Test errors

To verify our assumptions and theories along the paper, we evaluate the success of these in table 1. The error of each configuration was calculated on 182 saved predictions from the validation set, at the scale of 1024×1024. The results shown in this paper were likewise generated using the same image size. Error was averaged over all images to prevent outliers in the data. Both our proposed grids reach a significantly better result over all metrics. It was possible to lower the MSE below 50%, compared to the baseline. Furthermore, PSNR shows a remarkable increase of more than 15% and SSIM demonstrates a notably better result.

Predictions

Figure 6 demonstrates a set of predictions from the validation set, that have been transformed as the original training data and simulate a camera with the matching grid structure. The examples demonstrates successful highlight reconstruction, compared to the baseline. The prediction without LE-Pixel grid is not able to restore any high luminance structures in the upper example, whereas our method can restore a significant amount thereof, even in situations were only a small number of unclipped pixels is left. In dayscenes, colors and intensities of high luminance, reflecting surfaces can be recovered in a convenient way, even sensitive gradients on the horizon. The same applies for the lower example, where the baseline reconstruction is unable to restore any of the highlight information, while our method successfully is able to use the additional information for a plausible result. Even in highly saturated areas where light reflections are close to monochromatic, brightness and color could be matched as well as complex light situations, as displayed in figure 7. In contrast, the baseline can neither estimate the right luminance nor the right shape of the highlights.

The difference between, the horizontal grid and its 45° rotated counterpart is marginally in terms of metrics, as displayed in table 1. On a perceptive level, the horizontal grid shows a better reconstruction on horizontal and vertical structures, but lags behind in predicting diagonal structures. In comparison the 45° rotated grid shows opposing performance, where diagonal structures are better reconstructed. As horizontal and vertical structures dominate typical images, the horizontal grid is recommended.

Limitations

The limitations of this approach are content-dependent, as the network produces convincing results in large structures like clouds, it shows artefacts when trying to reconstruct small structures, as displayed in figure 8. As this is hardly surprising with $k = 10$, which leads to 1% of information left, the method suffers from undersampling / aliasing in this areas. Thus, clipped structures smaller then the LE-pixel grid will mostly be restored with matching luminance and color, but with vague guesses of the original structure. We also experimented with smaller LE-Grids, where already 2% LE-Pixels showed a significant improvement. An important point to mention here is that the undersampling problem derives partially from our ground truth, since the HDR+ dataset contains more information / structures in the highlights, as this would be the case in a real world camera [1], where the algorithm would be applied before tonemapping. In real world application this limitation is most probably still existing, but therefore significantly less visible. Additionally we tested this method only in the context of still images, so for applications on video, further work is needed, especially as videos introduce a temporal challenge, which makes it necessary for predictions to be continuous over time.

Conclusion

Preserving and reconstructing highlights is an important and challenging task in the development of better imagery. To give another method to preserve high luminance information in camera, we present a $k \times k$ with $k = 10$ grid of highlight preserving pixels to sustain information directly in camera for later processing. To provide evidence that a small number of pixels is enough to gain additional dynamic range, we use a fully convolutional autoencoder for reconstruction, as one possibility of a fully automated process. The functionality, quality and drawbacks of the method are demonstrated through a number of examples.

References

[1] 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 (Proc. SIGGRAPH Asia)*, vol. 35, no. 6, 2016.

[2] RED Digital Cinema, "High Dynamic Range Video with HDRx," `https://www.red.com/red-101/hdrx-high-dynamic-range-video`, Accessed: 2019-06-27.

[3] Saghi Hajisharif, Joel Kronander, and Jonas Unger, "Adaptive dualiso hdr reconstruction," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, pp. 41, 2015.

[4] Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep

Sen, “A versatile hdr video production system,” in *ACM SIGGRAPH 2011 Papers*, New York, NY, USA, 2011, SIGGRAPH ’11, pp. 41:1–41:10, ACM.

[5] Jan Froehlich, “Encoding high dynamic range and wide color gamut imagery,” Mar. 2018.

[6] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers, “Expanding low dynamic range videos for high dynamic range applications,” in *Proceedings of the 24th Spring Conference on Computer Graphics*. ACM, 2008, pp. 33–41.

[7] G. E. Hinton, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[8] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[9] 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)*, vol. 36, no. 6, pp. 178, 2017.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015.

[11] Youness Mansar, “Fingerprint Denoising and Inpainting using Fully Convolutional Networks,” Aug. 2018.

[12] Jae Sung Park, Jae Woong Soh, and Nam Ik Cho, “High Dynamic Range and Super-Resolution Imaging From a Single Image,” *IEEE Access*, vol. 6, pp. 10966–10978, 2018.

[13] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, “Loss Functions for Neural Networks for Image Processing,” *arXiv:1511.08861 [cs]*, Nov. 2015, arXiv: 1511.08861.

[14] Leon Eisemann, Jan Froehlich, Axel Hartz, and Johannes Maucher, “Expanding dynamic range in a single-shot image through a sparse grid of low exposure pixels,” `https://drive.google.com/drive/folders/1_ELo4ilz5lelT7HIP33hXD8uSENYdCvH?usp=sharing`, Accessed: 2019-06-30.

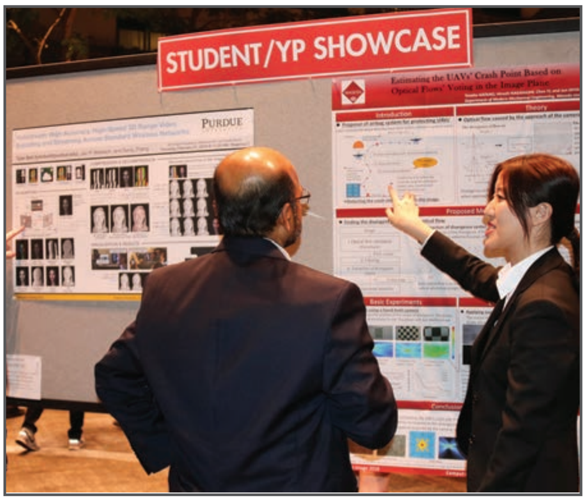
JOIN US AT THE NEXT EI!

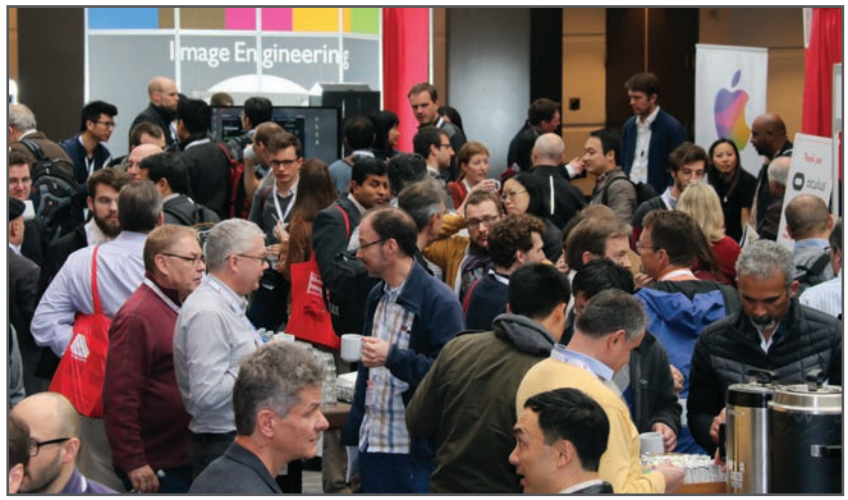
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •

• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T

imaging.org