Neuromorphic Camera Guided High Dynamic Range Imaging

Jin Han[1] Chu Zhou[1] Peiqi Duan[2] Yehui Tang[1] Chang Xu[3] Chao Xu[1] Tiejun Huang[2] Boxin Shi[2*]
[1]Key Laboratory of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University
[2]National Engineering Laboratory for Video Technology, Dept. of Computer Science, Peking University
[3]School of Computer Science, Faculty of Engineering, University of Sydney

Abstract

Reconstruction of high dynamic range image from a single low dynamic range image captured by a frame-based conventional camera, which suffers from over- or under-exposure, is an ill-posed problem. In contrast, recent neuromorphic cameras are able to record high dynamic range scenes in the form of an intensity map, with much lower spatial resolution, and without color. In this paper, we propose a neuromorphic camera guided high dynamic range imaging pipeline, and a network consisting of specially designed modules according to each step in the pipeline, which bridges the domain gaps on resolution, dynamic range, and color representation between two types of sensors and images. A hybrid camera system has been built to validate that the proposed method is able to reconstruct quantitatively and qualitatively high-quality high dynamic range images by successfully fusing the images and intensity maps for various real-world scenarios.

1. Introduction

High Dynamic Range (HDR) imaging is a widely used technique that extends the luminance range covered by an image. A batch of HDR imaging techniques have developed in recent decades by computer vision and graphics community, as summarized by Sen and Aguerrebere [48]. Traditional methods include taking multiple Low Dynamic Range (LDR) images under different exposures, then merging them with different weights to reproduce an HDR image [6]. Another approach is hallucinating texture details from a single LDR image, which is called inverse tone mapping [3]. Inverse tone mapping is obviously an ill-posed problem, that relies on predicting badly exposed regions from neighboring areas [54] or priors learned through deep neural networks [7].

In recent years, some specially designed neuromorphic cameras, such as DAVIS [4], have drawn increasing attention of researchers. Neuromorphic cameras have unique features different from conventional frame-based cameras, they are particularly good at sensing very fast motion and high dynamic range scenes ($1\mu s$ and 130 dB for DAVIS240). The latter characteristic can be utilized to form an *intensity map*, which encodes useful information lost in conventional imaging by a dynamic range capped camera due to over- and/or under-exposure. Despite the distinctive advantages in dynamic range, neuromorphic cameras generally bear low spatial resolution (180×240 for DAVIS240) and do not record color information, resulting in intensity maps less aesthetically pleasing as LDR photos captured by a modern camera. It is therefore interesting to study the fusion of LDR images and intensity maps with mutual benefits being combined for high-quality HDR imaging.

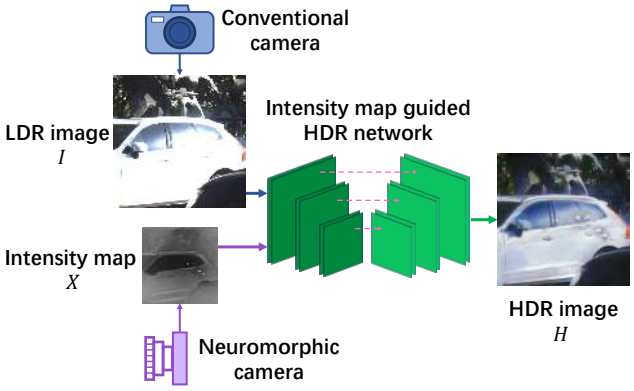

Figure 1. An intensity map guided HDR network is proposed to fuse the LDR image from a conventional camera and the intensity map captured by a neuromorphic camera, to reconstruct an HDR image.

In this paper, we propose a neuromorphic camera guided HDR imaging method. Directly stitching the intensity map and LDR image from two types of sensors is expected to produce poor HDR reconstruction, due to the great domain gap on spatial resolution, dynamic range, color representation and so on. To address these issues, we propose the intensity map guided HDR network, with specific modules designed for each type of gap. As illustrated in Fig. 1, the

*Corresponding author: shiboxin@pku.edu.cn.

proposed network successfully takes as input two types of images and reconstructs a high-quality HDR image.

The main contributions of this paper can be summarized as follows:

1) We propose an information fusion pipeline to reconstruct an HDR image by jointly taking a single LDR image and an intensity map. This pipeline demonstrates the principles and design methodologies to bridge the gap on resolution, dynamic range, and color representation between two types of sensors and images (Sec. 3.1).
2) According to the proposed pipeline, an intensity map guided HDR network is constructed to achieve the faithful reconstruction. We design specific modules in the network to address each type of gap in a more effective and robust manner (Sec. 3.2).
3) We build a hybrid camera system to demonstrate that the proposed method is applicable to real cameras and scenes (Sec. 4.2). Extensive experiments on synthetic data, and real-world data captured by our hybrid camera validate that the proposed network is able to reconstruct visually impressive HDR images compared with state-of-the-art inverse tone mapping approaches.

2. Related Work

Image-based HDR reconstruction. The classic HDR imaging method was proposed by Debevec and Malik [6] by merging several photographs under different exposures. However, aligning different LDR images may lead to ghosting in the reconstructed HDR result due to misalignment caused by camera movement or changes in the scene. This problem incurs a lot of research on deghosting in HDR images [26, 38, 49, 56]. Instead of merging multiple images, inverse tone mapping was proposed by Banterle *et al*. [3], whose intent is to reconstruct visually convincing HDR images from a single LDR image. This ill-posed problem was attempted to be solved by several optimized approaches [30, 34, 43].

In recent years, Convolutional Neural Networks (CNNs) have been applied to a lot of HDR image reconstruction tasks. Kalantari and Ramamoorthi [25] firstly aligned images under different exposures using optical flow, and then fed them to a neural network which merged LDR images to reconstruct an HDR image. Eilertsen *et al*. [7] used a U-Net [44] like network to predict the saturated areas in LDR images, and applied a mask to reserve non-saturated pixels in LDR images, then fused the masked image with predicted image to get the HDR results. Endo *et al*. [8] clipped the badly exposed pixels at first. It predicted the LDR images under multiple exposures, then merged these LDR images using classic method [6]. ExpandNet [33] adopted three branches of encoders, concentrating on different level features, then it concatenated and fused the features to get HDR images. Trinidad *et al*. [52] proposed PixelFusionNet to fuse misaligned images from multi-view under different exposures for dynamic range expansion.

Computational HDR imaging. HDR imaging problem would become less ill-posed by using computational approaches or even unconventional cameras that implicitly or explicitly encode expanded dynamic range of the scene. Nayar *et al*. [36] added an optical mask on a conventional camera sensor to get spacially varying pixel exposures. Tocci *et al*. [51] implemented an HDR-video system that used a beam splitter to simultaneously capture three images with different exposure levels, then merged them into an HDR image. Hirakawa and Simon [18] placed a combination of photographic filter over the lens and color filter array on a conventional camera sensor to realize single-shot HDR imaging. Zhao *et al*. [58] used a modulo camera that is able to wrap the high radiance of dynamic range scene periodically and save modulo information, then used Markov Random Field to unwrap and predict the HDR scene radiance pixel-wisely.

Inspired by the mechanism of human retina, neuromorphic sensors such as DAVIS [4] (Dynamic and Active Pixel Vision Sensor), ATIS [40] (Asynchronous Time-based Image Sensor), FSM [59] (retina-inspired Fovea-like Sampling Model), and CeleX [20] detect the changes of scene radiance asynchronously. This series of non-conventional sensors surpass conventional frame-based cameras in various aspects [10] including high dynamic range. Reconstructing from raw events/spikes data has shown great potential in recovering very high dynamic range of the scene [21, 41, 42, 46, 59]. But there are few attempts trying to combine them with conventional imaging to produce more visually pleasing HDR photos with higher resolution and realistic color appearance.

3. Proposed Method

3.1. LDR Image and Intensity Map Fusion Pipeline

As illustrated in Fig. 1, our goal is to reconstruct an HDR image given the input of an LDR image I and an intensity map X. Such a fusion pipeline can be conceptually illustrated using Fig. 2, which contains four key steps:

Color space conversion. Most conventional cameras record color images in *RGB* format and each channel contains pixel values represented by 8-bit integers. There exists a nonlinear mapping between scene radiance and the pixel values in the camera pipeline, so we have to firstly map LDR images to linear domain via the inverse camera response function (CRF) f^{-1}. To fuse with the one-channel intensity map, we then convert color space of the LDR image from *RGB* to *YUV*. The *Y* channel I_Y indicates the luminance of I which is in the same domain of X, and *U, V*

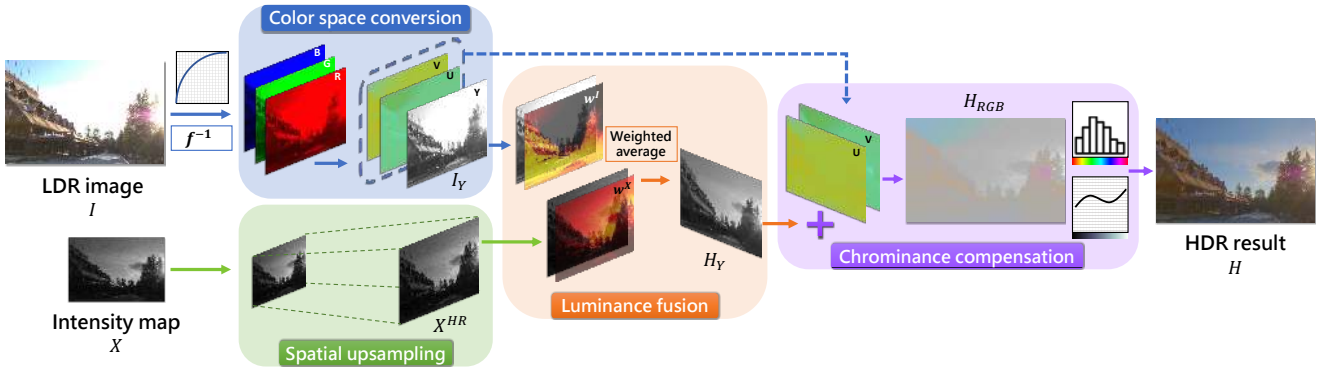

Figure 2. The conceptual pipeline of intensity map guided HDR imaging, consisting of four steps: color space conversion of the LDR image, spatial upsampling of the intensity map, luminance fusion to produce HDR image in luminance domain, and chrominance compensation that refills the color information to get a colorful HDR result.

channels contain the color information. We use I_Y to fuse with intensity map and reserve *U, V* channels as chrominance information to be added back later.

Spatial upsampling. To bridge the resolution gap between X and I_Y, we need to do an upsampling operation for the intensity map to make it have the same size as I_Y. The upsampling operation $S(\cdot)$ is defined as follows:

$$X^{HR} = S(X), \tag{1}$$

where X^{HR} is the upsampled intensity map. $S(\cdot)$ can be any upsampling operator such as nearest neighbor or bicubic interpolation, or a pre-trained neural network for super-resolution.

Luminance fusion. To expand the dynamic range of I_Y using X_{HR}, an intuitive solution is to define a weighting function which indicates the pixels that should be retained for fusion and those should be discarded. This can be implemented by adopting a similar merging strategy proposed by Debevec and Malik [6]. The fused value of H_Y is calculated as follows:

$$H_Y = W(I_Y, X^{HR}) = \frac{w^I I_Y + w^X X^{HR}}{w^I + w^X}, \tag{2}$$

where w^I and $w^X \in [0, 1]$ indicate corresponding weights for different types of input signals. A straightforward way to determine the weight values is to set a threshold τ (*e.g.*, $\tau > 0.5$) manually. Pixel values (normalized to $[0, 1]$) lying in the effective range $[1-\tau, \tau]$ are given larger weights to retain the information, while values out of the range are either too dark (under-exposed) or too bright (over-exposed), hence smaller weights are given to discard such information. A binary mask could be calculated based on the threshold, which is the simplest way to get a weight map. Another option is to set weights as a linear ramp,

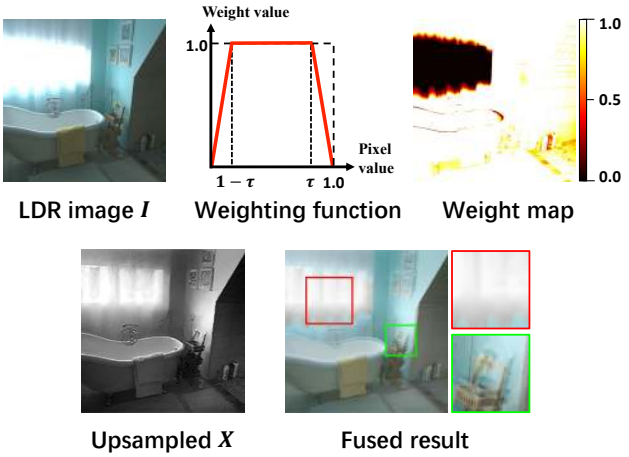

Figure 3. An example of fusing an intensity map and an LDR image using a linear ramp as weighting function. Such a straightforward fusion strategy results in various unpleasant artifacts (color distortion in the red box and blurry artifacts in the green box).

which is similar to the pixel-wise blending in [7]. Such a weighting function can be expressed as

$$w_i = \frac{0.5 - \max(|I_i - 0.5|, \tau - 0.5)}{1 - \tau}. \tag{3}$$

The weighting function and calculated weight map are shown in the first row of Fig. 3.

Chrominance compensation. H_Y now contains HDR information in high resolution, but only in luminance domain. The color information can be compensated from U, V channels of I, *i.e.*, I_U and I_V. Denote $C(\cdot)$ as the color compensation operator, this procedure can be represented as

$$H = C(H_Y, I_U, I_V), \tag{4}$$

which combines H_Y with I_U, I_V, and converts it back to *RGB* color space.

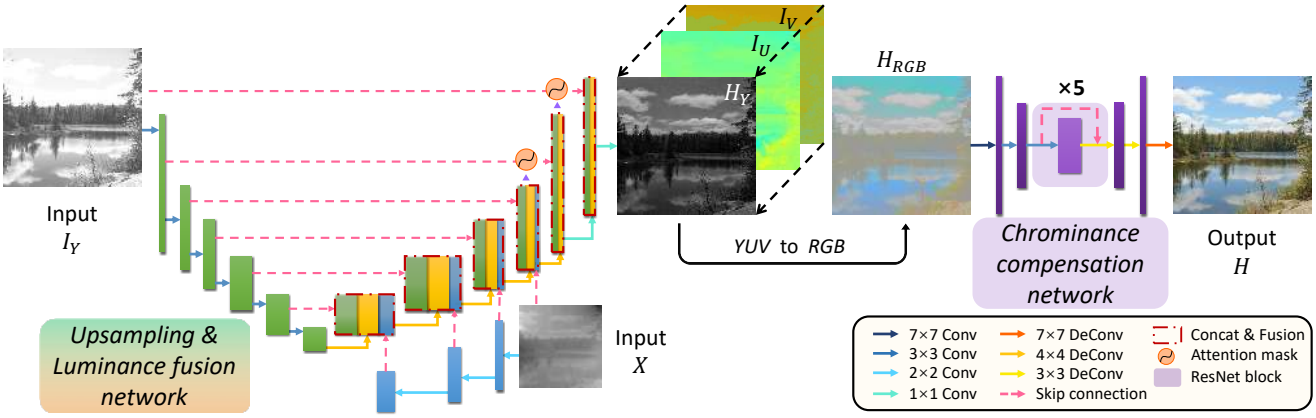

Figure 4. Overview of the intensity map guided HDR network architecture. It contains upsampling & luminance fusion network for HDR reconstruction in luminance domain, which is a double-encoder network with the input of I_Y and X, followed by the chrominance compensation network which takes the H_{RGB} as input to refine the color information.

Due to the dynamic range gap between H_Y and $I_U(I_V)$, directly combining them leads to unnatural color appearance, such as H_{RGB} shown in the color compensation block of Fig. 2. Realistic color appearance could be recovered using some color correction methods, such as color histograms adaptive equalization [5] or tone curve adjustment [31].

The synthetic example in Fig. 3 demonstrates that simply applying the conceptual pipeline in Fig. 2 may not achieve a satisfying HDR image. From the green box, we can easily observe blurry artifacts caused by misalignment between I and X (we add displacement between I and X in synthetic data to simulate the real setup), and unrealistic color recovery in the red box due to dynamic range gap.

To address these issues, we translate the pipeline in Fig. 2 as an end-to-end network $\mathcal{F}(\cdot)$:

$$H = C\left(W\left(f^{-1}(I_Y), S(X)\right), I_U, I_V\right) = \mathcal{F}(I, X; \theta), \quad (5)$$

where θ denotes parameters of the network. We will next introduce the specific concerns in realizing each of the four steps using deep neural networks.

3.2. Intensity Map Guided HDR Network

In this section, we describe the details of the proposed network, whose architecture is shown in Fig. 4. First of all, inverse CRF and color space conversion are conducted offline as a preprocessing to input. Then for each pixel in the input I_Y, the proposed network learns to extend the bit-width with the information encoded in X. We design specific modules in the network in accordance to the remaining three steps described in Sec. 3.1. Upsampling and luminance fusion can be realized by deconvolutional layers and skip connections in the network, respectively. Therefore, we design the network using the U-Net [44] architecture with double encoder (encoder of I_Y and X) and one decoder, the encoder of X guides the decoder to reconstruct HDR images at multiple representation scales.

Upsampling using deconvolution. We perform the upsampling operation by deconvolutional filters in the decoder. The purpose of decoder is to reconstruct H_Y from the fused latent representation, which has been encoded by the two encoders. Deconvolution has the ability to diffuse information from a small scale feature map to a larger feature map with learnable parameters. Feature maps from X at multiple scales guide the decoder to upsample with extended dynamic range information. Compared to a naive upsampling operation $S(\cdot)$, the deconvolutional layers learn a comprehensive representation from the image context to realize upsampling operation by end-to-end back-propagation, rather than simply rely on interpolation from nearby pixels.

Luminance fusion with attention masks. Information fusion in the luminance domain is the key step for dynamic range expansion. The proposed architecture applies skip-connections, which transfer feature maps between encoders and decoder to incorporate both rich textures in I_Y and high dynamic range information in X. Deeper networks have been shown to benefit from skip-connections in a variety of tasks [16, 19]. However, simply concatenating feature maps from two encoders is expected to be influenced by the dynamic range gap between the two input images. So we fuse the concatenated tensor by 1×1 convolution before each deconvolutional layer. As stated in the luminance fusion part of Sec. 3.1, a weighting function $W(\cdot)$ is added to determine the regions to retain, and the regions to be complemented by the other input. The weighting function can be implemented by introducing attention mechanism [23] in the network to assign different importance to different parts of an image. We choose to use self-attention gate [47]

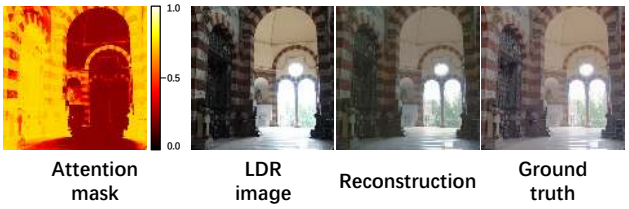

Figure 5. An example of attention mask calculated from self-attention module, which is added on LDR image to filter the badly exposed regions and reserve useful information for reconstruction.

as a mask added on I_Y to preserve the relevant information. There is no mask being added on X, and we use it just to complement the information that I_Y lacks. The attention mask is computed by 1×1 convolution, then activated by a non-linear function as implemented in Attention U-Net [39]. Then the element-wise multiplication of attention mask and input feature map from I_Y is able to filter the badly exposed pixels and reserve areas with valid information for reconstruction. Compared to assigning weights intuitively like Eq. (2), our attention mask is computed from feature maps of two input images, and the learnable parameters can be trained end-to-end. Fig. 5 demonstrates the effectiveness of attention mask.

Chrominance compensation network. Given the HDR image in luminance domain H_Y, we combine it with chrominance information I_U and I_V from the LDR image, then convert it to *RGB* color space to recover color appearance. The directly converted H_{RGB} may suffer from color distortion due to the dynamic range gap between H_Y and $I_U(I_V)$. Because the luminance values of *Y* channel are stored in high precision format (*e.g.*, float), while the *UV* values directly inherited from I are still in 8-bit integer format. As a result, the converted H_{RGB} tends to be dim and less colorful after tone mapping, as shown in Fig. 4. Therefore, we propose a chrominance compensation network to realize color correction for H_{RGB}. The network is an autoencoder [17] architecture with residual blocks [16], as the skip-connections in residual blocks learn to compensate for the difference between input and output images. It recovers the chrominance information on each pixel and learns to reconstruct realistic color appearance in HDR images, as demonstrated in the Output H of Fig. 4.

3.3. Loss Function

The proposed network reconstructs images in the linear luminance domain, which covers a wide range of values. Calculating loss between the output images H_Y and ground truth $\hat{H_Y}$ directly may cause the loss function being dominated by large values of H_Y, while the effect of small values tends to be ignored. Therefore, it is reasonable to compute loss function between H_Y and $\hat{H_Y}$ after tone mapping. The range of pixel values are compressed by the following function proposed by [25] after normalized to $[0, 1]$:

$$\mathcal{T}(H_Y) = \frac{log(1 + \mu H_Y)}{log(1 + \mu)}, \tag{6}$$

where $\mathcal{T}(\cdot)$ is the tone mapping operator and μ (set to be 5000) denotes the amount of compression. The tone mapping operator is computationally effective and differentiable, thus easy for back-propagation.

We train our network by minimizing the loss function which has two parts: pixel loss and perceptual loss [24]. Pixel loss computes the ℓ_1 norm distance between $\mathcal{T}(H_Y)$ and $\mathcal{T}(\hat{H_Y})$:

$$\mathcal{L}_{pixel} = \left\| \mathcal{T}(H_Y) - \mathcal{T}(\hat{H_Y}) \right\|_1. \tag{7}$$

Since the LDR images and intensity maps are taken from different cameras, the misalignment in the input pair is unavoidable. We try to solve this problem by adding the perceptual loss, which is defined for tone-mapped images based on feature maps extracted from the VGG-16 network [50] pre-trained on ImageNet [45]:

$$\mathcal{L}_{perc} = \sum_h \Big(\left\| \phi_h(\mathcal{T}(H_Y)) - \phi_h(\mathcal{T}(\hat{H_Y})) \right\|_2^2 + \left\| G_h^\phi(\mathcal{T}(H_Y)) - G_h^\phi(\mathcal{T}(\hat{H_Y})) \right\|_2^2 \Big), \tag{8}$$

where ϕ_h denotes the feature map convoluted from h-th layer of VGG-16, G_h^ϕ is the Gram matrix of feature maps ϕ_h of two input images. Both of the two parts are computed by ℓ_2 norm. We use the layers 'relu4_3' and 'relu5_3' of VGG-16 network in our experiments, because high-level features are less sensitive to unaligned pair of images.

To summarize, our total loss is written as:

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{pixel} + \alpha_2 \mathcal{L}_{perc}, \tag{9}$$

where α_1 and α_2 are the weights for different parts of loss function. We set as $\alpha_1 = 100.0$ and $\alpha_2 = 3.0$.

As for the chrominance compensation network, we also apply pixel loss and perceptual loss, and a slight difference is that the weight for perceptual loss α_2 is 5.0.

3.4. Methods of Acquiring Intensity Maps

The intensity maps can be acquired by various types of neuromorphic cameras.

Indirect approach. For neuromorphic cameras such as DAVIS [4], the output is a sequence of events data, containing the time, location, and polarity information of log intensity changes in a scene. The high dynamic range scene radiance is recorded in a *differential* manner by event cameras. Among many methods of reconstructing intensity maps from events, we choose the E2VID [41] network to

accumulate streams of events as a spatial-temporal voxel grid for reconstruction because it reconstructs the most convincing HDR intensity maps according to our experiments.

Direct approach. Intensity maps can also be acquired from spike-based neuromorphic cameras such as FSM [59]. Each pixel of the FSM responds to luminance changes independently as temporally asynchronous spikes. The accumulator at each pixel accumulates the luminance intensity digitalized by the A/D converter. Once the accumulated intensity reaches a pre-defined threshold, a spike is fired at this time stamp, and then the corresponding accumulator is reset in which all the charges are drained. To get the intensity map, we apply a moving time window to *integrate* the spikes in a specific period, and the intensity map can be computed by counting these spikes pixel-wisely [59]. FSM generates 40000 time stamps per second. We find that setting the window size as 1000 time stamps ($1/40$ second) reconstructs valid HDR intensity maps in our experiments.

3.5. Implementation Details

Dataset preparation. Learning-based methods rely heavily on training data. However, there are no sufficiently large-scale real HDR image datasets. Therefore, we collect HDR images from various image sources [1, 2, 11, 12, 37, 57] and video sources [9, 29]. Since the proposed network has two different types of images as input, we synthesize LDR images from HDR images like taking photos with a virtual camera [7]. The irradiance values of HDR images are scaled by random exposure time and then distorted by different non-linear response curves from DoRF [15]. As for the intensity maps, we simulate them in accordance with those acquired from neuromorphic cameras. Event camera estimates the gradients of the scene, and reconstructs the intensity maps using event data. So we firstly compute the gradients of tone-mapped HDR images and reconstruct intensity maps using Poisson solver [27] to simulate such process. During the training, we apply data augmentation to get an augmented set of pairwise data. We resize the original HDR images and then crop them to 512×512 at random locations followed by random flipping. We select viewpoints that an image covers the scene region with HDR information. Intensity maps are augmented using the same operations except cropping to the size of 256×256 as a low-resolution input.

Training strategy. The proposed network is implemented by PyTorch, and we use ADAM optimizer [28] during the training process with a batch size of 2. 600 epochs of training enables the network to converge. The initial learning rate is 10^{-5}, during the first 400 epochs it is fixed, in the next 200 epochs, it decays to 0 with a linear strategy. The encoder of I_Y is a VGG-16 network [50]. We convert I from *RGB* to *YUV* color space and duplicate I_Y twice to form a 3-channel tensor, since the input channel size of VGG-16 network is supposed to be three. The network is initialized with Xavier initialization [13]. We use instance normalization (IN) [53] instead of batch normalization (BN) [22] after each deconvolutional layer in the decoder. The output of network is activated by Sigmoid function that maps pixel values to the range of $[0, 1]$.

4. Experiments

4.1. Quantitative Evaluation using Synthetic Data

Fig. 6 shows reconstruction results of the proposed and other comparing methods. To the best of our knowledge, the proposed framework is the first one to combine LDR images with intensity maps to realize HDR images reconstruction. Therefore, we compare to three state-of-the-art deep learning based inverse tone mapping methods: DrTMO [8], ExpandNet [33], HDRCNN [7]; and traditional method [6] merging an over- and an under-exposed images with the exposure ratio of $50 : 1$. For the sake of fairness, we omit the comparison to merging three or more LDR images with different exposures. Thanks to the complementary dynamic range information provided by intensity maps, the proposed approach is able to recover rich texture details in the images such as clouds or the outline of the intense light source. For example, in the top row of Fig. 6, the leaves around the street lamp (green box) and the trunk (red box) are clearly visible in our results, which are similar to the ground truth, while this is not the case for other inverse tone mapping methods. Although merging two LDR images extends the dynamic range (more reliable than single-image solutions), it cannot recover rich details of a scene due to the limited dynamic range covered by two LDR images.

Besides visual comparison, we conduct quantitative evaluations using the widely adopted HDR-VDP-2.2 metrics [35], which computes the visual difference and predicts the visibility and quality between two HDR images. It produces the quality map and Q-Score for each HDR image to indicate the quality of reconstruction. Fig. 7 shows the quality maps and Q-Scores of different methods. The quality maps show the difference probability between a predicted HDR image and the ground truth. Results show that the proposed approach achieves more similarities to the ground truth and higher Q-Score in HDR image reconstruction compared to other methods.

4.2. Qualitative Evaluation using Real-world Data

In order to demonstrate the effectiveness of the proposed method on real-world scenarios, we build a hybrid camera system [55], which is composed of a conventional camera (Point Grey Chameleon 3) and a neuromorphic camera (DAVIS240 [4] or FSM [59]) with the same F/1.4 lens, as illustrated in Fig. 8. There is a beam splitter in front of the two sensors, which splits the incoming light and sends them

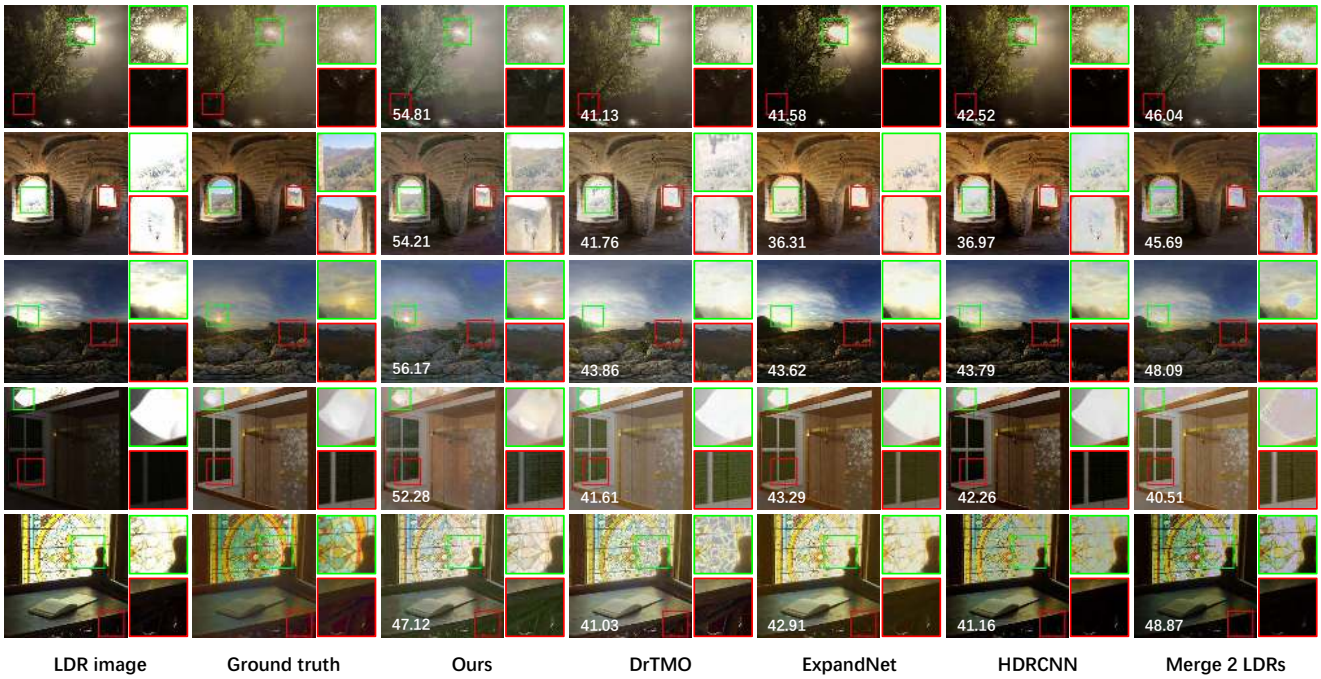

Figure 6. Visual comparison between the proposed method and state-of-the-art deep learning based inverse tone mapping methods: DrTMO [8], ExpandNet [33], HDRCNN [7]; and a conventional approach merging two LDR images [6]. The Q-Scores are labeled in each image. Please zoom-in electronic versions for better details.

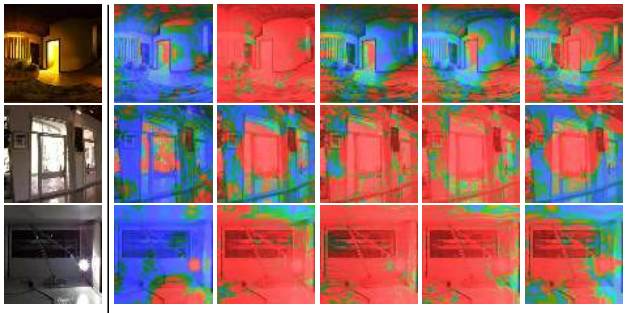

Figure 7. Comparisons on quality maps and Q-Scores calculated from HDR-VDP-2.2 evaluation metrics [35]. Visual differences increase from blue to red. Q-Scores average across all images in the whole test dataset.

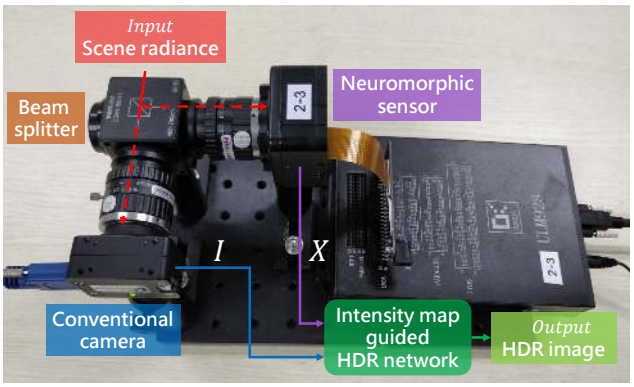

Figure 8. The prototype of our hybrid HDR imaging system composed of a conventional and a neuromorphic camera.

to different sensors simultaneously. We conduct geometric calibration and crop the center part from two camera views to extract the well-aligned regions as I and X.

We take photos for both indoor and outdoor high dynamic range scenes and reconstruct the HDR images using our method. In Fig. 9, the input images are first fused in the luminance domain (the third column) and then compensated by the chrominance information (the last column). Results show that the proposed method could fuse the input I and X to reconstruct high-quality HDR images. For example, the outline of car roof (the first row) is over-exposed due to the strongly reflected light, but the detailed texture could be captured by the neuromorphic cameras, and recovered in the fusion results using our method. Some artifacts could be observed in our results, such as in the glass window of the second row. This is brought by the intensity map during image reconstruction using stacked events [41], since we need to shake the DAVIS to trigger events, where blur and noise may occur.

4.3. Ablation Study

To validate the effectiveness of the proposed architecture and each specific module, we conduct ablation study on three variants as follows. Visual comparison for different variants is shown in Fig. 10.

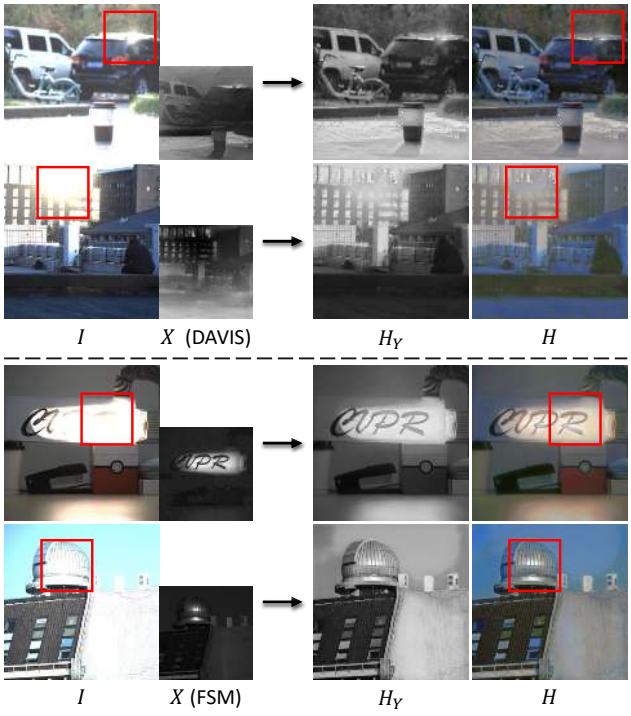

Figure 9. Real data results reconstructed by the proposed network. The LDR images are captured by conventional cameras and the intensity maps are acquired by DAVIS (the upper two rows) and FSM (the lower two rows).

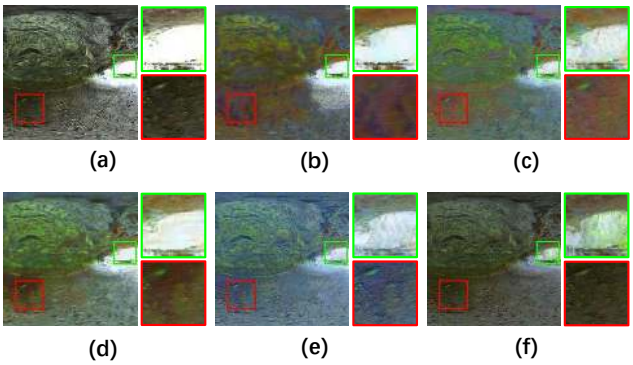

Figure 10. Visual comparison of different variants of the proposed method. (a) Input LDR image, (b) without attention mask, (c) single encoder architecture, (d) adding adversarial loss, (e) intensity map guided HDR network, and (f) ground truth.

Without attention masks. We investigate the effectiveness of the attention mask module. As shown in Fig. 10 (b), the green box of over-exposed region is similar to the input LDR image. Without the guidance of attention masks, it is difficult for the network to correctly distinguish information to reserve or discard. Hence leads to some artifacts.

Single encoder architecture. We compare our network with a single encoder architecture without the encoder of X. This can be achieved by scaling X to the same size of I_Y and concatenating them at first, then sending the 4-channel tensor to a single encoder. In this case, two images from different domains are directly combined. As Fig. 10 (c) shows, although the over-exposed region is recovered, obvious artifacts can be observed in the under-exposed regions.

Adding adversarial loss. We investigate adding adversarial loss [14] $\mathcal{L}_{adv}$ to the total loss function. By treating the proposed network as the generator, we train a discriminator simultaneously to distinguish the real or fake HDR image. The example in Fig. 10 (d) shows that the adversarial loss may lead to undesired mosaics in both over- and under-exposed areas.

We also conduct HDR-VDP-2.2 metrics for different models, and the average Q-Scores are calculated for evaluation, which are shown in the following results. Without attention masks: 45.55, single encoder: 49.74, adding adversarial loss: 45.22, complete model: **51.68**. These results demonstrate our complete model achieves the optimal performance with these specifically designed strategies.

5. Conclusion

We proposed a neuromorphic camera guided HDR imaging method, which fuses the LDR images and the intensity maps to reconstruct visually pleasing HDR images. A hybrid camera system has been built to capture images in real-world scenarios. Extensive experiments on synthetic data and real-world data demonstrate that the proposed method outperforms state-of-the-art comparing methods.

Discussion. Considering the limitation of GPU memory, we use 512×512 LDR images and 256×256 intensity maps to train the networks. However, our model can handle higher resolution LDR images once we upsample the intensity maps to the corresponding scale level with LDR images.[1] Apart from that, increasing the resolution of output images might also be achieved by a pre-trained super-resolution network [32].

Limitations and future work. Due to the distortion of camera lens and different field of views of the two sensors in our hybrid camera system, the pixels that are better aligned are mostly centered on the whole image plane. Therefore, although this paper demonstrates convincing evidence of fusing two types of images for HDR imaging, the final quality still has a gap between merging multiple images of different exposures, captured by a modern DLSR with tens of millions of pixels. To realize this, using a better designed hybrid camera is our future work.

6. Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant No. 61872012, No. 61876007, and Beijing Academy of Artificial Intelligence (BAAI).

[1] Higher resolution results can be found in the supplementary material.

References

[1] Funt et al. HDR dataset. https://www2.cs.sfu.ca/~colour/data/funt_hdr/#DESCRIPTION.

[2] sIBL archive. http://www.hdrlabs.com/sibl/archive.html.

[3] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proc. of International Conference on Computer Graphics and Interactive Techniques*, 2006.

[4] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[5] Vasile V Buzuloiu, Mihai Ciuc, Rangaraj M Rangayyan, and Constantin Vertan. Adaptive-neighborhood histogram equalization of color images. *Journal of Electronic Imaging*, 10(2):445–460, 2001.

[6] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of ACM SIGGRAPH*, 1997.

[7] 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 (Proc. of ACM SIGGRAPH Asia)*, 36(6):178:1–178:15, Nov. 2017.

[8] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 36(6):177:1–177:10, Nov. 2017.

[9] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Proc. of Digital Photography X*, 2014.

[10] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Joerg Conradt, Kostas Daniilidis, and Scaramuzza. Davide. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405*, 2019.

[11] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

[12] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2019.

[13] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2010.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of International Conference on Neural Information Proceeding Systems*, 2014.

[15] Michael D Grossberg and Shree K Nayar. What is the space of camera response functions? In *Proc. of Computer Vision and Pattern Recognition*, 2003.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition*, 2016.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[18] Keigo Hirakawa and Paul M Simon. Single-shot high dynamic range imaging with conventional camera hardware. In *Proc. of Internatoinal Conference on Computer Vision*, 2011.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. of Computer Vision and Pattern Recognition*, 2017.

[20] Jing Huang, Menghan Guo, and Shoushun Chen. A dynamic vision sensor with direct logarithmic output and full-frame picture-on-demand. In *Proc. of International Symposium on Circuits and Systems*, 2017.

[21] S. Mohammad Mostafavi I., Lin Wang, Yo-Sung Ho, and Kuk-Jin Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. of Computer Vision and Pattern Recognition*, 2019.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[23] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision*, 2016.

[25] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 36(4):144–1, 2017.

[26] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *Proc. of Internatoinal Conference on Computational Photography*, 2006.

[27] H Kim, A Handa, R Benosman, SH Ieng, and AJ Davison. Simultaneous mosaicing and tracking with an event camera. In *Proc. of the British Machine Vision Conference*, 2014.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, and Jonas Unger. Unified HDR reconstruction from raw CFA data. In *Proc. of Internatoinal Conference on Computational Photography*, 2013.

[30] Pin-Hung Kuo, Chi-Sun Tang, and Shao-Yi Chien. Content-adaptive inverse tone mapping. In *Proc. of Visual Communications and Image Processing*, 2012.

[31] Markku Lamberg, Joni Oja, Tero Vuori, and Kristina Bjorknas. Image adjustment with tone rendering curve, June 9 2005. US Patent App. 10/729,872.

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a

generative adversarial network. In *Proc. of Computer Vision and Pattern Recognition*, 2017.

[33] 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, 2018.

[34] Belen Masia, Sandra Agustin, Roland W. Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 28(5):160:1–160:8, Dec. 2009.

[35] Manish Narwaria, Rafal Mantiuk, Mattheiu P. Da Silva, and Patrick Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24:010501, 2015.

[36] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proc. of Computer Vision and Pattern Recognition*, 2000.

[37] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in LDR and HDR images. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2015.

[38] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, 2014.

[39] 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. *arXiv preprint arXiv:1804.03999*, 2018.

[40] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. An asynchronous time-based image sensor. In *Proc. of International Symposium on Circuits and Systems*, 2008.

[41] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2019.

[42] Christian Reinbacher, Gottfried Graber, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *arXiv preprint arXiv:1607.06283*, 2016.

[43] 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 Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 26(3), July 2007.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[46] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Proc. of Asian Conference on Computer Vision*, 2018.

[47] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*, 2018.

[48] Pradeep Sen and Cecilia Aguerrebere. Practical high dynamic range imaging of everyday scenes: Photographing the world as we see it with our own eyes. *Signal Processing Magazine*, 33(5):36–44, 2016.

[49] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 31(6):203:1–203:11, Nov. 2012.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[51] Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile HDR video production system. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 30(4):41:1–41:10, July 2011.

[52] Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, and Janne Kontkanen. Multi-view image fusion. In *Proc. of Internatoinal Conference on Computer Vision*, 2019.

[53] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[54] Lvdi Wang, Li-Yi Wei, Kun Zhou, Baining Guo, and Heung-Yeung Shum. High dynamic range image hallucination. In *Proc. of the Eurographics conference on Rendering Techniques*, 2007.

[55] Zihao Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020.

[56] Greg Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of Graphics Tools*, 8(2):17–30, 2003.

[57] Feng Xiao, Jeffrey M DiCarlo, Peter B Catrysse, and Brian A Wandell. High dynamic range imaging of natural scenes. In *Proc. of Color and Imaging Conference*, 2002.

[58] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *Proc. of Internatoinal Conference on Computational Photography*, 2015.

[59] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *Proc. of International Conference on Multimedia and Expo*, 2019.