Inverting the Imaging Process by Learning an Implicit Camera Model (Supplementary Material)

Xin Huang[1], Qi Zhang[2], Ying Feng[2], Hongdong Li[3], Qing Wang[1]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[2] Tencent AI Lab [3] Australian National University

1. Additional Implementation Details

1.1. Network Details

All networks in our framework are based on the MLP architecture. The deformation network $\mathcal{D}$ is a 4-layer MLP with 256 channels of each layer. The atlas network $\mathcal{A}$ consists of 4 layers with 512 channels. The offset network $\mathcal{O}$ and weight network $\mathcal{W}$ also have 4 layers, each with 64 channels. The tone-mapping network $\mathcal{T}$ is composed of three MLPs, each of 2 layers with 128 channels, to fit the response functions of R, G, and B channels respectively. Rectified Linear Unit (ReLU) activations are adopted between inner layers of networks, and the outputs of the last layers are passed through a tanh activation, except for the weight network $\mathcal{W}$ which takes softmax activation instead.

1.2. Training Details

Without the supervision of all-in-focus HDR images, our framework is sensitive to the initial values of parameters of the network. During the initial bootstrapping phase (10k iterations), we firstly train the deformation network $\mathcal{D}$ by mapping pixel position $\mathbf{p} = (x, y, i)$ (normalized to range $[-1, 1]$) to coordinate (x, y). It enforces the deformations to be initialized as zero, considering that the static background occupies a large proportion of the image sequence. The optical flow estimated by the off-the-shelf method [7] is not always accurate due to the different focuses and exposures of input images. Therefore, during the training phase, the flow loss weight λ_f gradually decays to 0 over the course of optimization.

2. Additional Experiments and Results

2.1. Comparisons with Traditional Methods

Compared with optimization-based traditional methods on the all-in-focus HDR imaging task, our neural camera model enables recovering irradiance maps from the image stack where exposure and defocused blur vary simultaneously. As shown in Fig. 2 (a), HF fails to recover an all-in-focus image due to the different exposures of input images, and PM is similar. Although HF+PM and PM+HF can recover all-in-focus HDR images from 9 images, our method takes only 3 images as input and outperforms them.

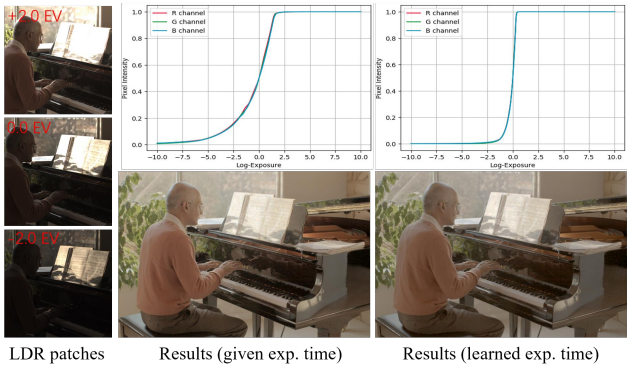


Figure 1. Results comparison of our method with given exposure time and learned exposure time on an ME scene. The left column shows the zoom-in patches of input images. The top row shows CRFs learned by *tone mapper*. The bottom row shows our tone-mapped HDR results.

2.2. Larger Sampling Patch

The evaluation of our generated LDR and defocused images are presented in Fig. 2 (b). Using depth maps is helpful to generate accurate defocused blur. However, our method can also produce photorealistic defocused blur without the monocular depth. Using 3×3 samples to represent the PSF is not enough when the degree of blur is too large, which causes aliasing artifacts. The blur pattern of defocus is more reasonable when we use 5×5 pixels to infer a defocused pixel, as shown in Fig. 2 (b).

2.3. Evaluation of Implicit Camera Model

We evaluate the pre-trained camera model on a new scene. The results are shown in Fig. 2 (c). One can see that our model achieves a competitive result using a pretrained

Figure 2. (a) Comparisons with two off-the-shelf optimization-based algorithms. "HF" is Helicon Focus software for all-in-focus images and "PM" is Photomatix software for HDR images. (b) Evaluations of our generated LDR and defocused images. (c) Evaluations on a new scene by freezing the pre-trained camera model. The first and third rows are all-in-focus HDR images. The PSNR and SSIM are presented in the lower left corner. **Please open with PDF reader for zoom-in**.

CRF. Unlike our *tone mapper*, pixel positions are fed into our *blur generator* to produce blending weights and offsets, which causes the learned blur generator to depend on the trained scene. Therefore, the performance decreases when we freeze the pre-trained blur generator and then train our model on a new scene.

2.4. Learned Exposure

The EXIF tags may be unavailable for compressed images from the internet. So we can also learn the exposure time for each image during the optimization. Figure 1 shows the recovered HDR images with given exposure time or learned exposure time on the ME dataset. As can be seen, there is a scale difference between the two CRFs but HDR images with abundant details are well reconstructed by the two models, which illustrates learning exposure time is feasible.

2.5. The Number of Input (Training) images.

To evaluate the influence of input images, different combinations of exposures and focuses are evaluated in our method. Figure 3 shows three sets (3 images, 5 images, and 9 images) of input images. Ideally, 3 images are enough for our method to recover all-in-focus HDR images. However, in some special cases when images focus on over-exposed or under-exposed areas, the method produces results with artifacts (*e.g.* artifacts on the head of the dog when the model is trained on 3 images). In these cases, raising the number of input images to 5 or 9 yields better results.

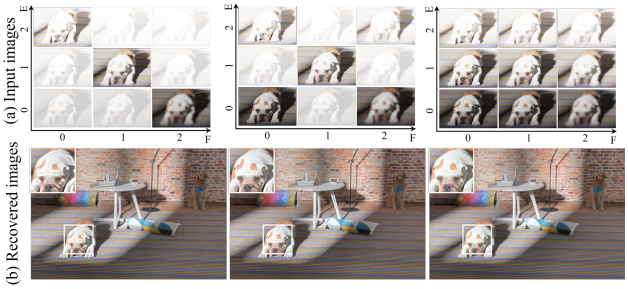


Figure 3. Results comparison of our method with different sets of input images on a synthetic MFME scene. (a) Top row shows the zoom-in patches of the input images. F denotes the focus and E denotes the exposure. (b) Bottom row shows the corresponding tone-mapped results. **Better viewed on screen with zoom in.**

2.6. More Results

In Fig. 8 and 9, we show the comparisons of our method with the two-stages methods for recovering all-in-focus and HDR images from the images with different focuses and exposures. As one can see, The results by FusionDN [10] + PM [5] have distorted colors. U2Fuison [9] + PM [5] and MFF-GAN [13] + PM [5] produce better results with consistent colors, but both methods fail to deal with the ghosting near the object boundary, such as the cups in Fig. 9 (green insets). Compared with the two-stage methods, our method produces all-in-focus and HDR images with sharp boundaries and details.

In Fig. 10, we show the comparisons of our method with multi-focus image fusion (MFIF) methods for recovering all-in-focus images from a near-focused image and a far-

Figure 4. Visualization of our results for video deblurring.

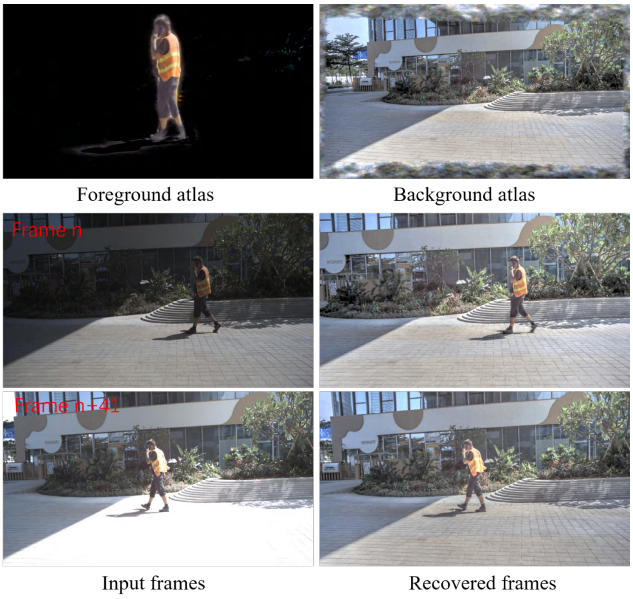


Figure 5. Visualization of our results for HDR video reconstruction. The visualizations of the neural atlases are shown in the first row.

focused image. Similarly, the results by FusionDN [10], U2Fuison [9] and MFF-GAN [13] all have ghosting near boundaries, while our results are clearer and have a consistent color with input images. Figure 11 presents the recovered HDR results of our method and the state-of-the-art HDR imaging methods (HDR-GAN [4], AHDRNet [11], and DeepHDR [8]) for dynamic scenes. Three SOTA methods fail to recover the textures outside the window in the top scene, due to there the large over-exposed region in the reference image. Compared with them, our methods produce superior results. Besides, our method does not produce artifacts on the moving objects, such as the hand of the baby in the bottom scene, which demonstrates that our method can fit the scene with mall motions.

3. Applications

3.1. Controllable Rendering

The other consequence of our implicit camera model is that it enables rendering images with modified camera settings. When we keep the *blur generator* and *tone mapper* during the inference, our method can control the focus and exposure of rendered images. The degree of defocus blur is greatly related to depth. For example, the points at the same depth should have a consistent blur on images. To control the focus correctly, we concatenate the position $\mathbf{p}$ with the corresponding depth and feed them into the *blur generator* to learn the PSF, where the depth is estimated using a single image depth estimation method [12]. We have tried to modify the focus of the images from our MFME dataset, but we find the depth estimation method failed to predict accurate depths. Consequently, we evaluate the focus control on a Lytro dataset [3]. The scene contents in the Lytro dataset are relatively simple, so we can estimate the depth accurately. To render images with varying focus, we interpolate the image indices i that are fed into the *blur generator*. Additionally, we control the exposure of rendered images by modifying the exposure time Δt. The exposure control is evaluated on the ME dataset. Figure 7 shows the controllable rendering of our method. We see the focus of the images (top row) smoothly varies from the foreground to the background. The bottom row presents the modification of exposures, where the exposure of the renderings increases gradually.

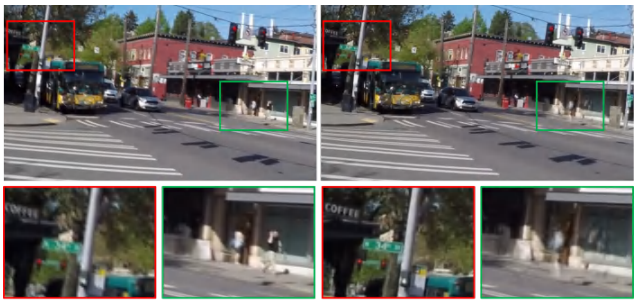
3.2. Video Enhancement

Our implicit camera model is also applicable to video enhancement combined with video scene representations. We adopt the layered neural atlases representation [2], which decomposes the video into a set of layered 2D atlases to deal with object motions and camera motions. We evaluate our model for video deblurring on Deep Video Deblurring (DVD) dataset [6] and HDR video reconstruction on the Deep HDR Video (DHV) dataset [1]. A video deblurring case is shown in Fig. 4. The input video of 100 frames with camera motion blur and our method recovers sharper textures. For the HDR video reconstruction task, the input is a video of 80 frames with alternating exposures. Note that, this input video contains a moving person, so the video is represented with two atlases: an atlas for the foreground and an atlas for the background. In Fig. 5, we show the

results for HDR video reconstruction. We can see that the scene contents are successfully split into two atlases and our method recovers the texture of over-exposed areas based on information from other frames with a lower exposure (see the ground in frame $n + 41$).

3.3. A Failure Case

Figure 6 shows a failure case where pedestrians on the street are missing in the recovered images since the people are too small to split into a single atlas and there are lots of self-occlusions. However, our camera model also successfully removes the camera motion blur of the video.



One input frame Recovered frame

Figure 6. One failure case of our method. The input video is challenging in that the pedestrians have complex self-occlusions. The pedestrians on the street are missing in our recovered frames (see the green insets), while our method removes the camera motion blur of the video (see the red insets).

References

[1] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *ICCV*, pages 2502–2511, 2021.

[2] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM TOG*, 40(6):1–12, 2021.

[3] Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25:72–84, 2015.

[4] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE TIP*, 30:3885–3896, 2021.

[5] Photomatrix. Photo editing software for hdr & real estate photography. https://www.hdrsoft.com/, 2021.

[6] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, pages 1279–1288, 2017.

[7] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020.

[8] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*, pages 117–132, 2018.

[9] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, 44(1):502–518, 2020.

[10] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12484–12491, 2020.

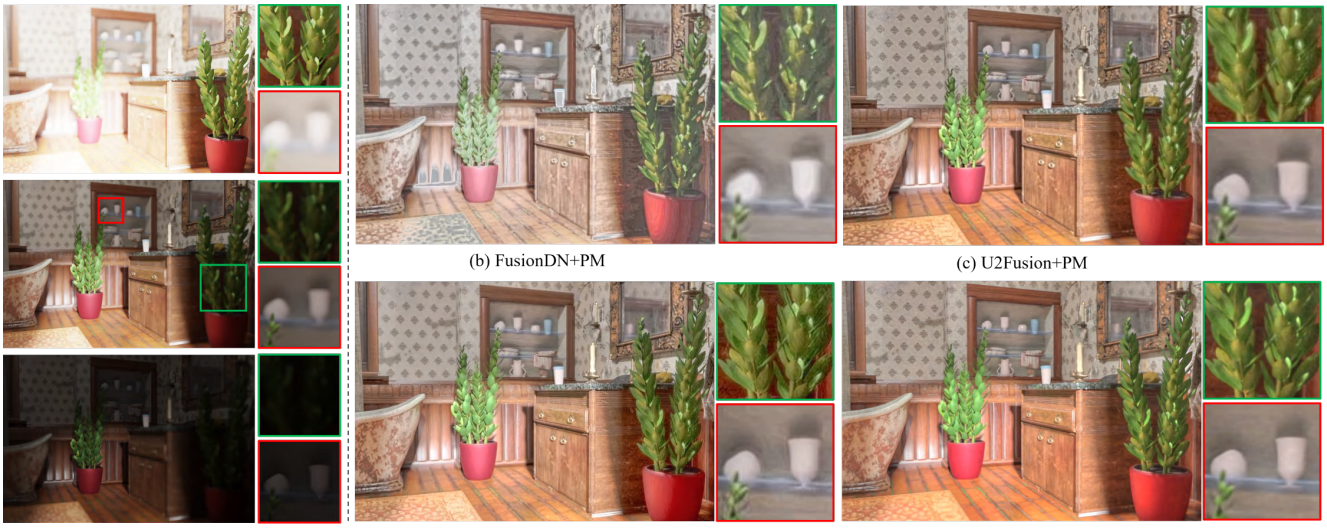
[11] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *CVPR*, pages 1751–1760, 2019.

[12] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE TPAMI*, 2021.

[13] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021.

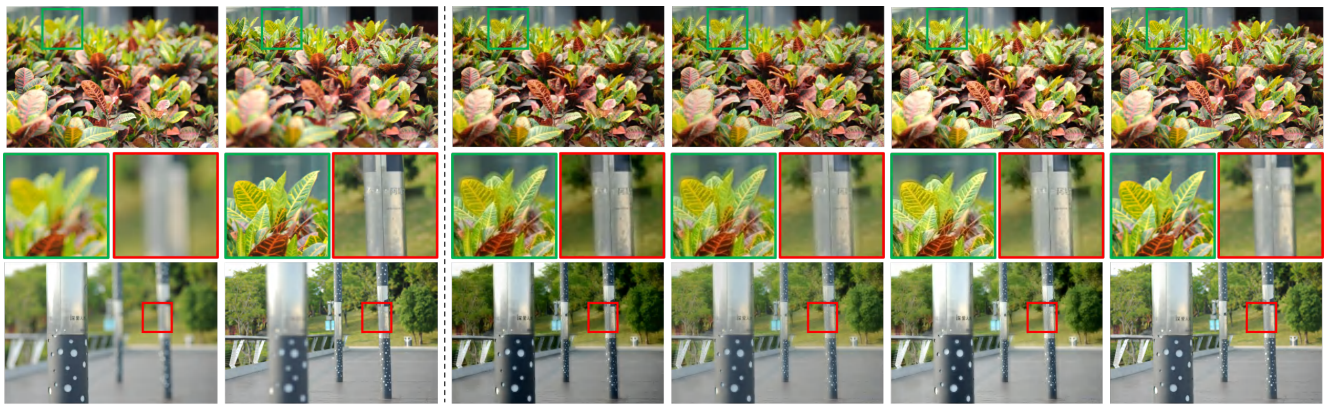
Figure 7. Controllable rendering results of our method. The leftmost and rightmost images are two training images, and the middle results are rendered with interpolated focus or exposure.

Figure 8. Example results of our method compared with two-stage methods on the MFME real dataset. "PM" denotes the HDR imaging method in Photomatix [5]. (a) Our input images with different focuses and exposures. (b-e) All-in-focus and HDR images produced by three two-stage methods and our method. The red and green insets show the zoom-in views of the images. All HDR images are tone-mapped for display.



(a) Input blurry LDR images (b) FusionDN+PM (c) U2Fusion+PM (d) MFF-GAN+PM (e) Ours

Figure 9. Example results of our method compared with two-stage methods on the MFME synthetic dataset. "PM" denotes the HDR imaging method in Photomatix [5]. (a) Our input images with different focuses and exposures. (b-e) All-in-focus and HDR images produced by three two-stage methods and our method. The red and green insets show the zoom-in views of the images. All HDR images are tone-mapped for display.



(a) Input images (b) FusionDN (c) U2Fusion (d) MFF-GAN (e) Ours

Figure 10. Example results of our method compared with MFIF methods on the MF dataset. (a) Two input images. One is near-focused and the other is far-focused. (b-e) All-in-focus results by MFIF methods and our method. The red and green insets show the zoom-in views of the images. **Better viewed on screen with zoom in.**

(a) Input LDR images (b) DeepHDR (c) AHDRNet (d) HDR-GAN (e) Ours

Figure 11. Example results of our method compared with HDR imaging methods on the ME dataset. (a) Three input images with different exposures. Exposure values (EVs) are shown in the upper left. The image highlight with a blue box denotes the reference image. (b-e) The recovered HDR images for the reference image. All HDR images are tone-mapped for display.