All-in-focus Imaging from Event Focal Stack
Supplementary Material

Hanyue Lou[†1,2] Minggui Teng[†1,2] Yixin Yang[1,2] Boxin Shi[*1,2]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{hylz, minggui_teng, yangyixin93, shiboxin}@pku.edu.cn

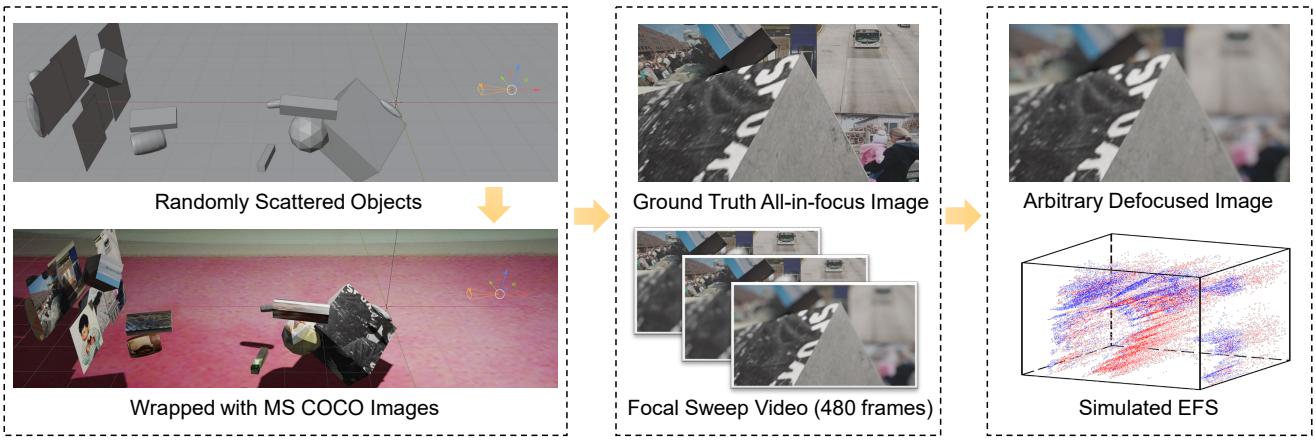


Figure 9. An illustration of our synthetic data generation pipeline.

6. Dataset

6.1. Training synthetic dataset

The pipeline for generating the training dataset is shown in the Figure 9. First, we pick geometric objects, scale them and scatter them at different depths in the scene randomly, forming 200 scenes. Second, to better match the data distribution to real-world images, we wrap the surfaces of the objects with images sampled from the MS-COCO dataset [6] as their textures. The sampled images are from *2017 Val images*, a subset of MS-COCO. The final step is to render 480 images as an image focal stack for each scene and then to render the ground truth image as an all-in-focus image with a small aperture.

After rendering the image focal stack, we input them into DVS-Voltmeter [5] to generate event streams. To improve the generalization of the model to unknown types of event cameras, we apply the 6 different camera parameters in DVS-Voltmeter randomly. Each camera parameter ($k_1 \sim k_6$) is randomly sampled from the range $[Min, Max]$ shown in Table 3. The reference columns "DVS346" and "DVS240" are the parameters calibrated on the event camera models DVS346 and DVS240, provided as DVS-Voltmeter preset configurations [5].

Table 3. Settings of DVS-Voltmeter [5] parameters.

Param.	range	DVS346	DVS240
k_1	$[4.0, 5.5]$	5.3	4.4
k_2	$[18, 25]$	20	23
k_3	$[5 \times 10^{-5}, 2.5 \times 10^{-4}]$	1×10^{-4}	2×10^{-4}
k_4	$[0.8 \times 10^{-7}, 1.2 \times 10^{-7}]$	1×10^{-7}	1×10^{-7}
k_5	$[3 \times 10^{-9}, 8 \times 10^{-8}]$	5×10^{-9}	5×10^{-8}
k_6	$[8 \times 10^{-6}, 1.2 \times 10^{-5}]$	1×10^{-5}	1×10^{-5}

6.2. Real-captured data

As shown in Figure 10, we build a hybrid camera system consisting of a machine vision camera (HIKVISION MV-CA050-12UC) and an event camera (PROPHESEE GEN4.0) with a beam splitter. For calibration, we use a

† Contributed equally to this work as first authors

* Corresponding author

Project page: https://hylz-2019.github.io/EFS

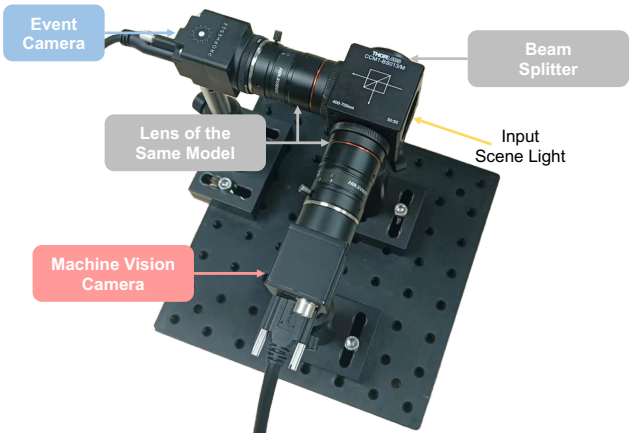


Figure 10. Our hybrid camera system.

Table 4. Ablation study on loss functions.

	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓
$\mathcal{L}_2$ only	33.04	0.9305	0.9605	0.1547
perc. only	31.51	0.8977	0.9567	0.1580
$\mathcal{L}_1$ + perc.	33.09	0.9300	0.9599	0.1515
$\mathcal{L}_2$ + perc.	**33.25**	**0.9323**	**0.9611**	**0.1510**

checkerboard to deal with homography and radial distortion between two views.

7. Ablation Experiment

7.1. Loss function

We ablate different loss functions ($\mathcal{L}_2$ loss only, perceptual loss only, $\mathcal{L}_1$ loss + perceptual loss) from the complete model ($\mathcal{L}_2$ loss + perceptual loss) and evaluate them quantitatively in Table 4. Results show that the combination of $\mathcal{L}_2$ loss and perceptual loss improves the performance of EvRefocusNet and EvMergeNet in reconstructing all-in-focus images.

7.2. Qualitative comparison

The qualitative comparison among the different ablation studies is shown in Figure 11. According to the results, our complete model can produce a sharper, all-in-focus image. Note that ET-Net [9] only reconstructs gray-scale images, thus, we only compare the results of "ET+MNet" with the gray-scale ground truth.

7.3. Analysis

To verify the effectiveness of each module, we conduct three ablation studies, shown Section 4.3, and the detailed analysis of each ablation study is listed as follows:

- "ET+MNet": Our EvRefocusNet takes a single defocused image with the corresponding event stream as input, while ET-Net [9] only utilizes the event stream, resulting in a lack of texture details.
- "RNet+GDF": Since the event stream provides high-temporal-resolution edge information, compared with gradient domain fusion [11], our EvMergeNet can predict more accurate weights for focal stack merging.
- "Uniform": By dynamically selecting refocus distances with our golden search method instead of sampling distances uniformly, our method can refocus to objects which would fall between the uniform samples otherwise, as illustrated in Figure. 2 (the blue cube is out of focus in all focal stack images). Our method also avoids refocusing on distances with no objects, which causes a waste of computation when distances are sampled uniformly.

Table 5. Quantitative results on the LiFF dataset [1].

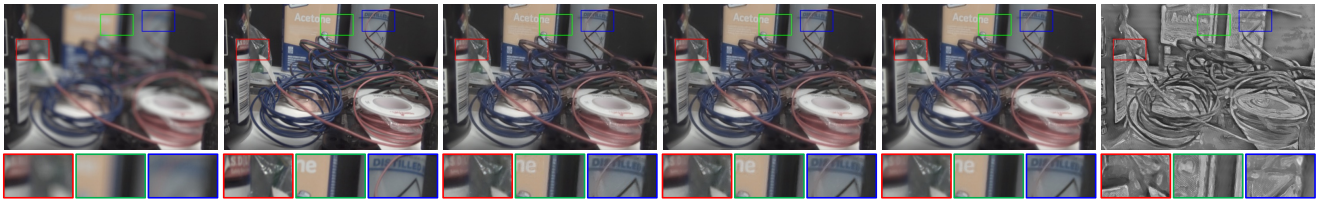
	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓
KPAC [8]	27.96	0.8396	0.9115	0.2473
IFAN [4]	29.59	0.8119	0.8679	0.3741
APL [10]	27.14	0.7758	0.8522	0.4060
DRBNet [7]	30.51	0.8639	0.9278	0.2164
Ours	**33.25**	**0.9323**	**0.9611**	**0.1510**

8. Experiments with Image-based Methods

To compare with single-image-based methods comprehensively, we feed them with 10 images in the same scene, which are focused at different focal distances, obtain the 10 defocused deblurring images, and then calculate the average metric values as the final results. The quantitative result is shown in Table 5. Based on the results, our method still outperforms the state-of-the-art image-based methods.

9. Speed Variation Issue of Focal Plane

For convention image focal stack methods [2, 11], the focal plane must move at a stable speed. However, our EFS is less affected by this restriction. We take EFS by rotating the focus ring by hand, which inevitably makes the focal plane move at a varying speed. With the high temporal resolution property, the event camera can detect scene radiance changes at the microsecond level. Since our manual rotation speed is much slower than its temporal resolution, the performance of our method is robust to such speed variation. As the example in Figure 12 shows, our method can restore an all-in-focus image with consistently high quality, given EFS captured at different speeds. As we capture EFS manually, we show histograms of the number of events at each timestamp to partially reflect the speed variation when rotating the focus ring for EFS capture.



(a) Defocused Image (b) Ground Truth (c) Ours (d) Uniform (e) RNet+GDF (f) ET+MNet

Figure 11. Visual quality comparison of ablation studies on synthetic data.

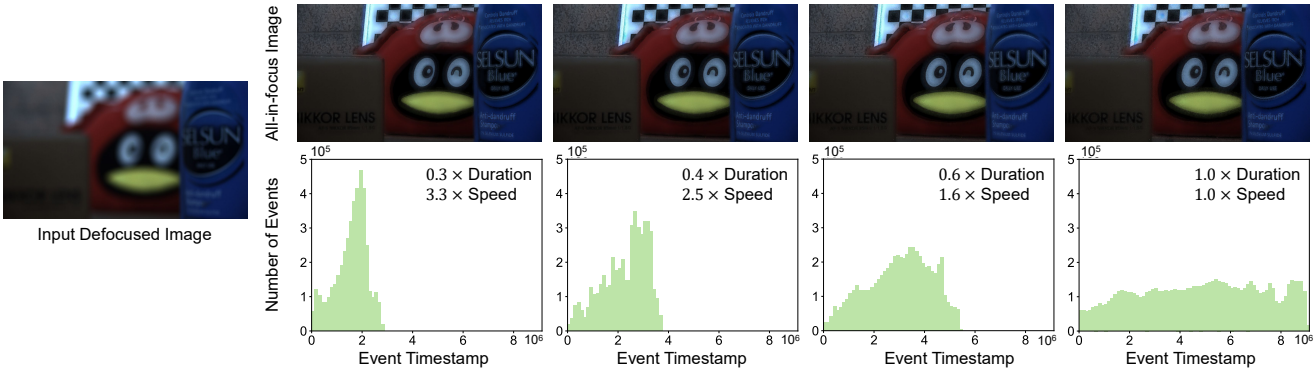


Figure 12. Results of our method given EFS captured at different speeds. Below each image, we show the histograms of the number of events at each timestamp, to partially reflect the speed variation when rotating the focus ring for EFS capture.

10. Network Details

In this section, we present architecture details of our EvRefocusNet (shown in Table 6) and EvMergeNet (shown in Table 7). Since we split the defocused image into $N \times N$ spatially non-overlapping image patches, in which N is set to 8 in our implementation, the reconstructed image focal stack is composed of 64 refocused images correspondingly.

11. Image Focal Stack

Our method can restore images refocused at arbitrary focus distances from a single defocused image and the corresponding EFS. The generated image focal stacks are shown in our project page.

12. More Results on Synthetic Dataset

In this section, we provide more qualitative comparisons among our method, DRBNet [7], IFAN [4], KPAC [8], and APL [10] on synthetic data, shown in Figure 13, Figure 14, and Figure 15.

13. More Results on Real Dataset

In this section, we provide more qualitative comparisons among our method, DRBNet [7], IFAN [4], KPAC [8], and APL [10] on real data, shown in Figure 16.

References

[1] Donald G. Dansereau, Bernd Girod, and Gordon Wetzstein. LiFF: Light field features in scale and depth. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[2] Berthold Klaus Paul Horn. Focusing. Technical report, MIT, 1968. 2

[3] 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. 4

[4] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2, 3, 5, 6, 7, 8

[5] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Proc. of European Conference on Computer Vision*, 2022. 1

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and

Table 6. Network details of EvRefocusNet. DenseConv modules are densely connected convolutional blocks [3]. ResBlock modules are residual blocks. Deconv modules are transposed convolutional blocks. All modules include batch normalization and activation functions.

EvRefocusNet	Input	Kernel Size	Stride	In Channels	Out Channels	Output
Conv1	Defocused Image	3	1	3	64	conv1
DenseConv1	conv1	3	1	64	128	denseconv1
Conv2	denseconv1	2	2	128	128	conv2
DenseConv2	conv2	3	1	128	256	denseconv2
Conv3	denseconv2	2	2	256	256	conv3
DenseConv3	conv3	3	1	256	512	denseconv3
ConvE1	Event Stack	3	1	64	64	conve1
DenseConvE1	conve1	3	1	64	128	denseconve1
ConvE2	denseconve1	2	2	128	128	conve2
DenseConvE2	conve2	3	1	128	256	denseconve2
ConvE3	denseconve2	2	2	256	256	conve3
DenseConvE3	conve3	3	1	256	512	denseconve3
Deconv2	[denseconv3, denseconve3]	2	2	1024	256	deconv2
Conv5	[denseconv2, denseconve2, deconv2]	1	1	768	128	conv5
DenseConv5	conv5	3	1	128	256	denseconv5
Deconv1	denseconv5	2	2	256	128	deconv1
Conv6	[denseconv1, denseconve1, deconv1]	1	1	384	64	conv6
ResBlock1	conv6	3	1	64	64	resblock1
ResBlock2	resblock1	3	1	64	64	resblock2
PredConv	resblock2	3	1	64	3	pred

Table 7. Network details of EvMergeNet. DenseConv modules are densely connected convolutional blocks [3]. ResBlock modules are residual blocks. Deconv modules are transposed convolution blocks. All modules include batch normalization and activation functions.

EvMergeNet	Input	Kernel Size	Stride	In Channels	Out Channels	Output
Conv1	Image Stack	3	1	$3 * 64$	64	conv1
DenseConv1	conv1	3	1	64	128	denseconv1
Conv2	denseconv1	2	2	128	128	conv2
DenseConv2	conv2	3	1	128	256	denseconv2
Conv3	denseconv2	2	2	256	256	conv3
DenseConv3	conv3	3	1	256	512	denseconv3
ConvE1	Event Stack	3	1	64	64	conve1
DenseConvE1	conve1	3	1	64	128	denseconve1
ConvE2	denseconve1	2	2	128	128	conve2
DenseConvE2	conve2	3	1	128	256	denseconve2
ConvE3	denseconve2	2	2	256	256	conve3
DenseConvE3	conve3	3	1	256	512	denseconve3
Deconv2	[denseconv3, denseconve3]	2	2	1024	256	deconv2
Conv5	[denseconv2, denseconve2, deconv2]	1	1	768	128	conv5
DenseConv5	conv5	3	1	128	256	denseconv5
Deconv1	denseconv5	2	2	256	128	deconv1
Conv6	[denseconv1, denseconve1, deconv1]	1	1	384	64	conv6
ResBlock1	conv6	3	1	64	64	resblock1
ResBlock2	resblock1	3	1	64	64	weights

C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. of European Conference on Computer Vision*, 2014. 1

[7] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 3, 5, 6, 7, 8

[8] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proc. of International Conference on Computer Vision*, pages 2642–2650, 2021. 2, 3, 5, 6, 7, 8

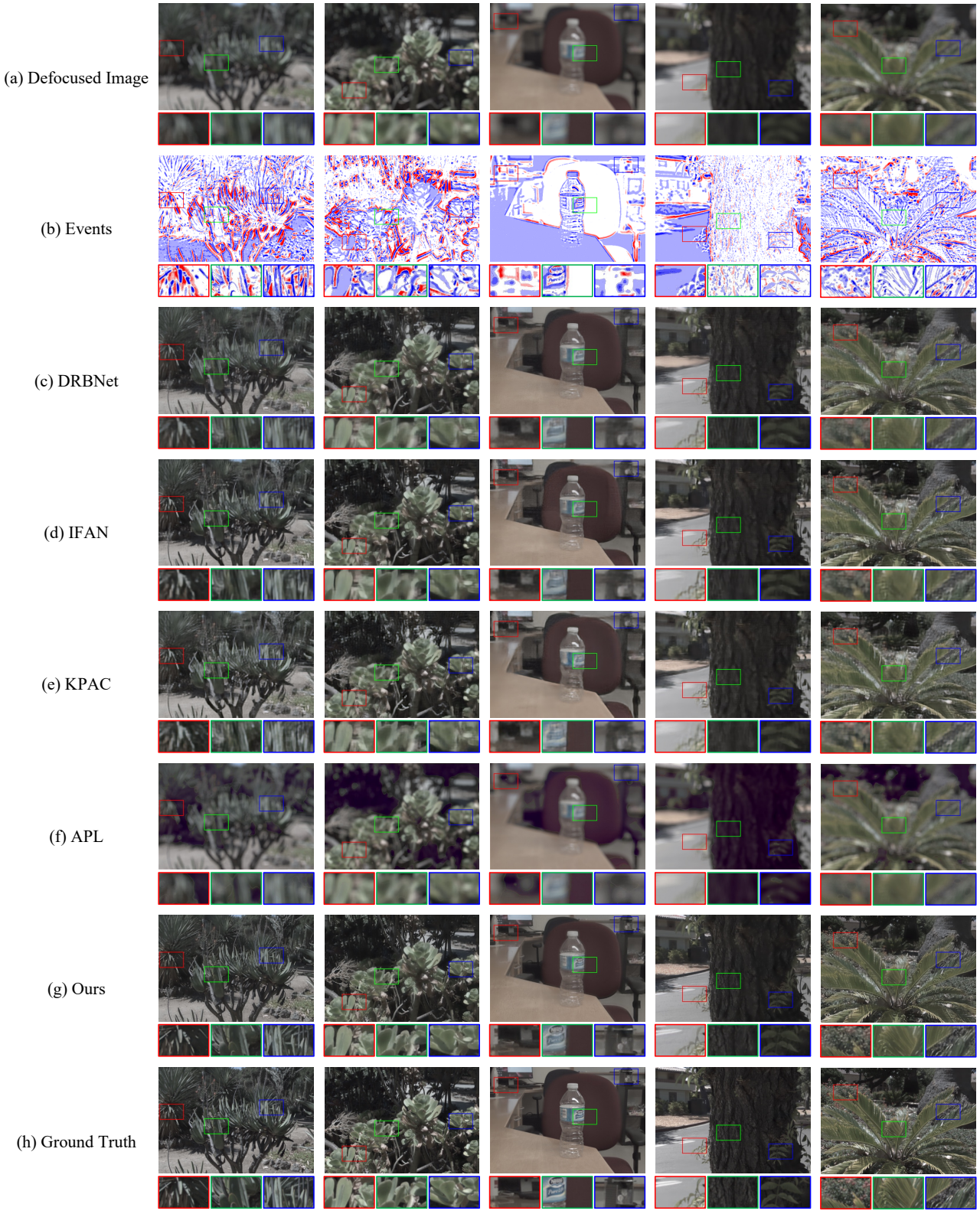


Figure 13. Visual quality comparison with image-based defocus deblurring methods on synthetic data (Part I). (a) Defocused image. (b) Events. (c)∼(g) All-in-focus results of DRBNet [7], IFAN [4], KPAC [8], APL [10], and ours. (h) Ground truth.

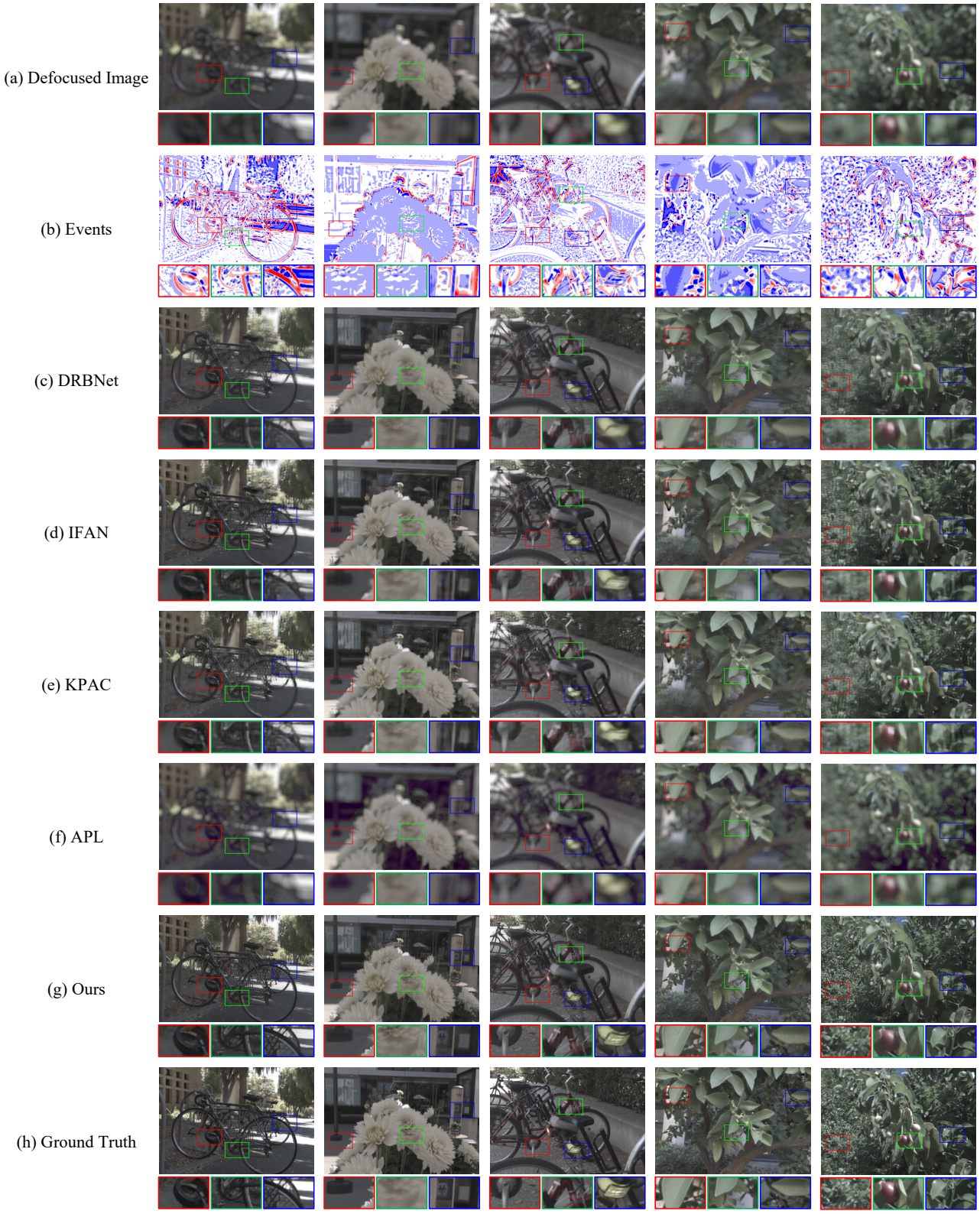


Figure 14. Visual quality comparison with image-based defocus deblurring methods on synthetic data (Part II). (a) Defocused image. (b) Events. (c)∼(g) All-in-focus results of DRBNet [7], IFAN [4], KPAC [8], APL [10], and ours. (h) Ground truth.

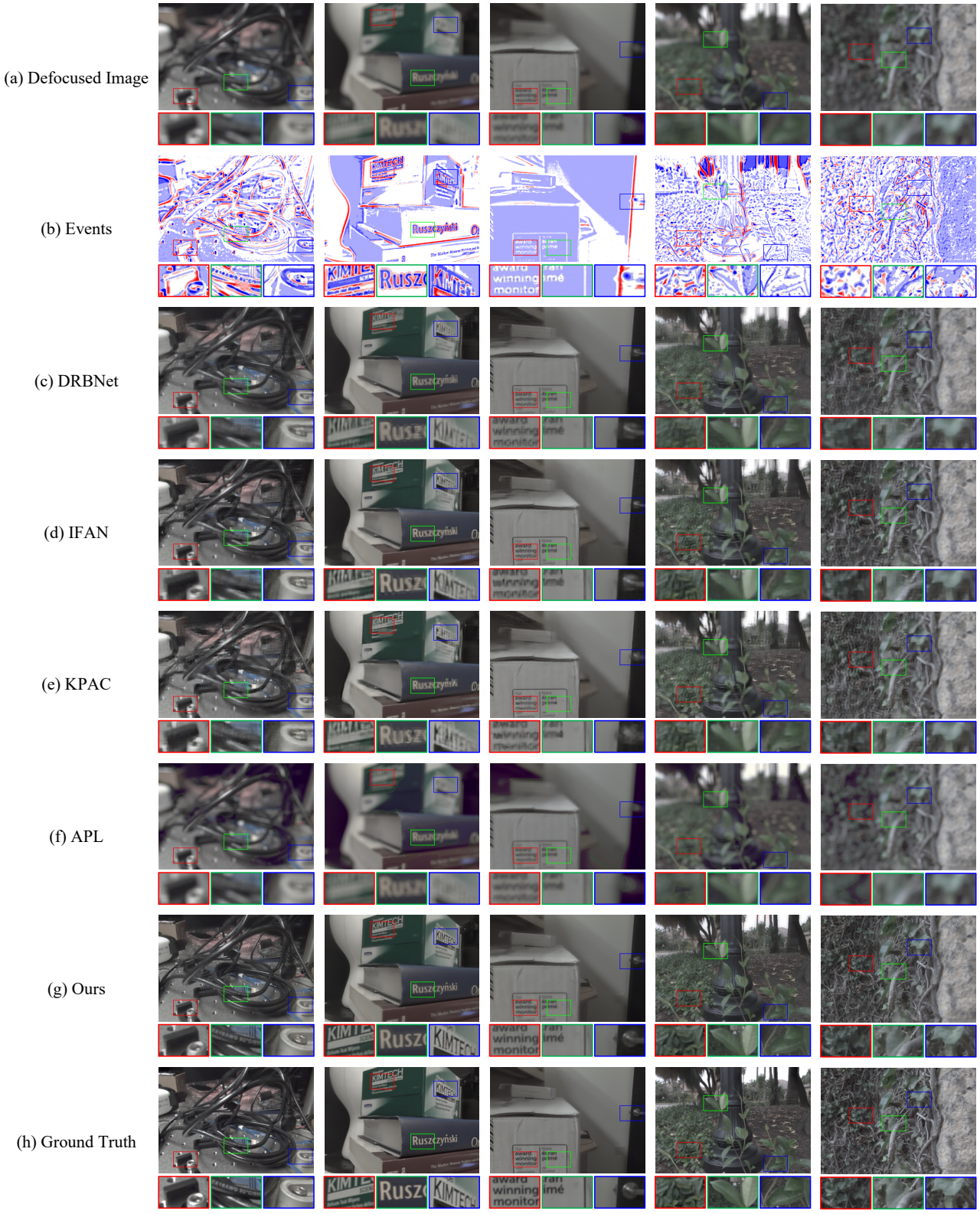


Figure 15. Visual quality comparison with image-based defocus deblurring methods on synthetic data (Part III). (a) Defocused image. (b) Events. (c)∼(g) All-in-focus results of DRBNet [7], IFAN [4], KPAC [8], APL [10], and ours. (h) Ground truth.

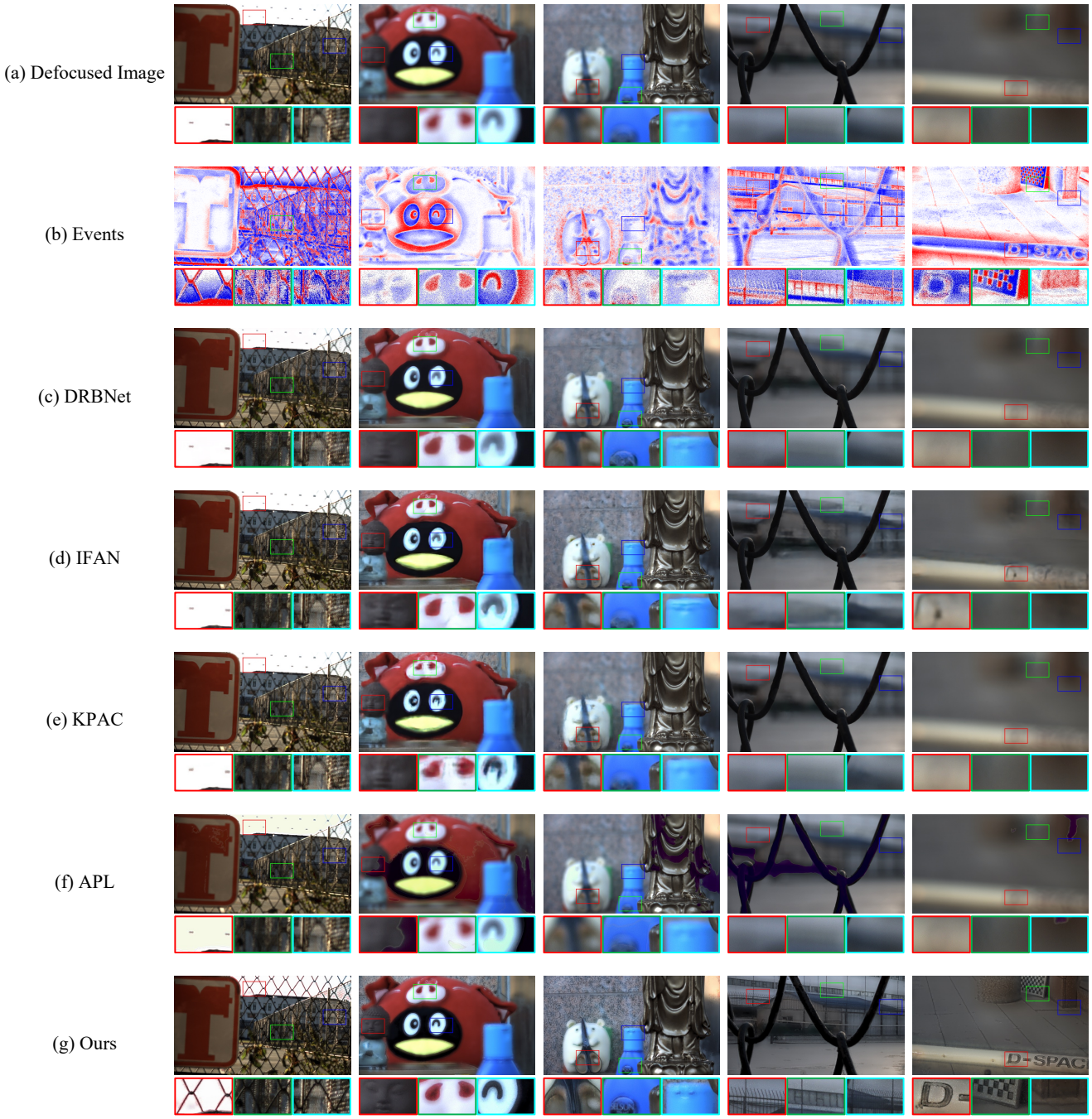


Figure 16. Visual quality comparison with image-based defocus deblurring methods on real data. (a) Defocused image. (b) Events. (c)∼(g) All-in-focus results of DRBNet [7], IFAN [4], KPAC [8], APL [10], and ours.

[9] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. of International Conference on Computer Vision*, 2021. 2

[10] Wenda Zhao, Fei Wei, You He, and Huchuan Lu. United defocus blur detection and deblurring via adversarial promoting learning. In *Proc. of European Conference on Computer Vision*, 2022. 2, 3, 5, 6, 7, 8

[11] Changyin Zhou, Daniel Miau, and Shree K. Nayar. Focal sweep camera for space-time refocusing. Technical report, Columbia University, 2012. 2