

Supplementary Material: Event-driven Video Frame Synthesis

Zihao W. Wang¹ Weixin Jiang¹ Kuan He¹ Boxin Shi² Aggelos Katsaggelos¹ Oliver Cossairt¹
¹ Northwestern University ² Peking University
 {winswang, weixinjiang2022}@u.northwestern.edu

A. Comparison of two binning strategies

The two binning strategies are first reiterated below:

- Binning 1: For an incoming event, if its spatial location already has an event in the current event frame, then cast it into a new event frame; Otherwise, this incoming event will stay in the current event frame. In this case, each event frame should only have three values, *i.e.*, $\{-1, 0, 1\}$.
- Binning 2: Similar to several previous work [1, 2, 4, 6], where events are stacked/integrated over a time window, we allow each event frame to have more than three values. However, since the "tanh" function in event sensing model only outputs values between -1 and 1, we modify our event sensing model to have a summation operation over several sub-event frames. That is, $\mathcal{E}_{b2} = \sum_t \mathcal{E}_t$.

We use a toy example to further analyze the performance of the two binning strategies, shown in Fig. S1. Assume there are two intensity pixels at different locations. During a certain amount of time, each intensity pixel outputs two intensity values, *i.e.*, a_1 and b_1 from Pixel 1 and a_2 and b_2 from Pixel 2. Assume in the same time window four events are fired from two event pixels. (Assume the locations of event pixels and intensity pixels match perfectly.) According to Binning 1, the four events result in 3 event frames. Therefore, two intermediate frames $[x_{11}, x_{21}]$ and $[x_{12}, x_{22}]$ can be interpolated accordingly. Binning 1 makes sufficient use of the temporal order of events, resulting in 6 constraints:

$$\begin{cases} \sigma(x_{11} - a_1) = 0 \\ \sigma(x_{12} - x_{11}) = 1 \\ \sigma(b_1 - x_{12}) = 1 \\ \sigma(x_{21} - a_2) = -1 \\ \sigma(x_{22} - x_{21}) = -1 \\ \sigma(b_2 - x_{22}) = 0, \end{cases} \quad (\text{S1})$$

where we use $\sigma(\cdot)$ to denote the event sensing model $\tanh\{\alpha(\cdot)\}$. Binning 2 integrates sub-event frames to-

gether. Therefore, it does not preserve the temporal order of events, resulting in ambiguity. In Eq. S2, the first equation has at least three solutions, *i.e.* $\{0, 1, 1\}$, $\{1, 0, 1\}$, $\{1, 1, 0\}$ corresponding to each "tanh" function respectively. This ambiguity is challenging to be solved by stochastic gradient descent.

$$\begin{cases} \sigma(x_{11} - a_1) + \sigma(x_{12} - x_{11}) + \sigma(b_1 - x_{12}) = 2 \\ \sigma(x_{21} - a_2) + \sigma(x_{22} - x_{21}) + \sigma(b_2 - x_{22}) = -2 \end{cases} \quad (\text{S2})$$

B. Statistics on real event streams (Binning 1)

We examine several event streams captured in real scenarios using our Binning 1 strategy. The results are shown in Fig. S2. We plot three metrics: 1) event density, defined as (# of events) / (full resolution) \times 100% per event frame; 2) event speed, defined as (event density) / (event frame duration); and 3) event frame duration, defined as the elapsed time from the first to the last event in the same frame. We observe that the event frame duration results in less variation than the event density and speed. An empirical mean of the event frame duration is $\sim 2\text{ms}$, corresponding to $\sim 500\text{FPS}$ and $\sim 16\times$ temporal upsampling from 30FPS (regular frame rate).

C. Visual results for comparing plug & play to one-time denoising

We compare two frameworks, *i.e.*, the plug & play [5] and the one-time denoising to investigate how to use our trained Residual Denoiser (RD). The plug & play framework decouples the forward physical model and the denoising prior using the ADMM technique [3]. For one time denoising, we apply the residual denoiser once after the DMR has converged. One time denoising is considered because it is considerably faster than plug & play. From a computation point of view, one epoch of our DMR has comparable computation time to one layer of a fully-connected CNN.

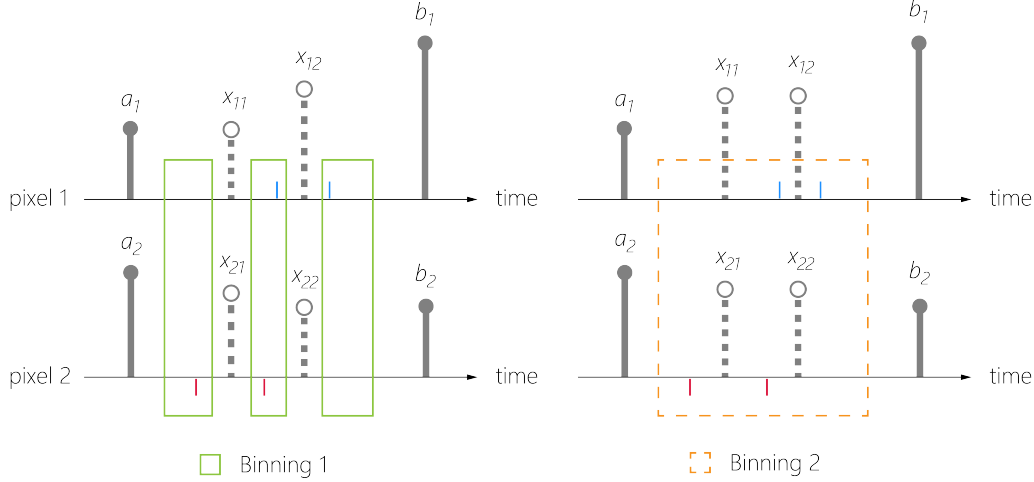


Figure S1: Comparison of two binning strategies.

However, our RD has 17 layers, which requires more computation time for the plug & play method. While we generally expect plug & play to perform better, our experimental results show that one-time denoising performs similar or even better than plug & play, shown in Table S3. (Visual results are included in supplementary material.) We reason that this is related to our training process and the initialization of the high-res tensor. Our differentiable model involves a temporal transition process from an existing frame to a future frame. We initialize the high-res tensor with the reference intensity frame. In each DMR iteration, the reconstruction process produces artifacts that are similar to the degradations in the initialized image. However, our denoiser is trained to “recognize” this “degradation” and remove these artifacts. Therefore, our denoiser is most useful and efficient when applied after the DMR has converged.

D. Additional results for RD compared to Gaussian denoisers

Additional results comparing our RD with state-of-the-art Gaussian denoisers are shown in Fig. S4.

References

- [1] P. Bardow, A. J. Davison, and S. Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 884–892, 2016. 1
- [2] S. Barua, Y. Miyatani, and A. Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *Proc. of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 1
- [3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternat-

ing direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 1

- [4] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3857–3866, 2019. 1
- [5] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg. Plug-and-play priors for model based reconstruction. In *Global Conference on Signal and Information Processing (GlobalSIP)*, pages 945–948. IEEE, 2013. 1
- [6] L. Wang, Y.-S. Ho, K.-J. Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1
- [7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 5
- [8] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 5

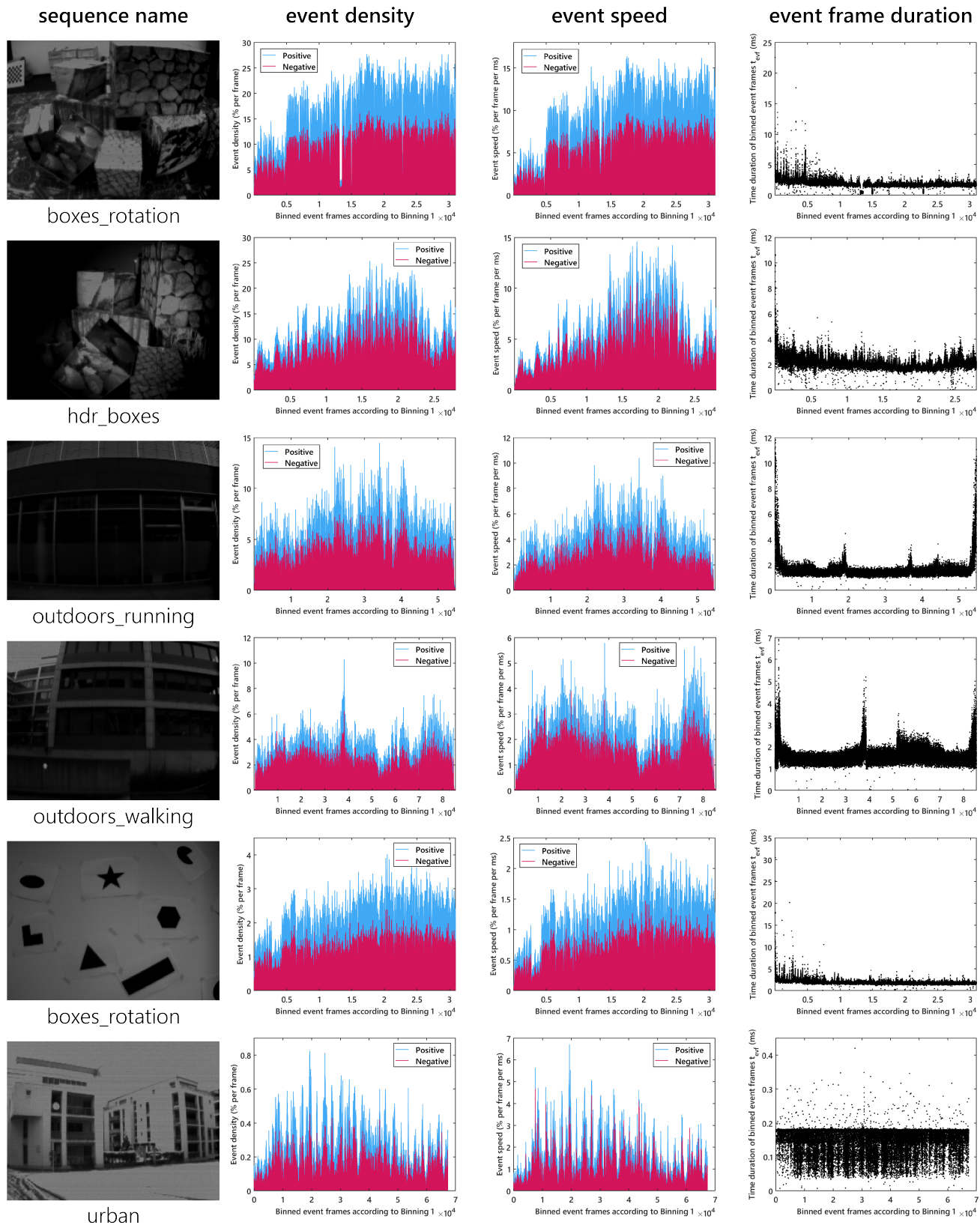


Figure S2: Statistics for using Binning 1 on real event streams.




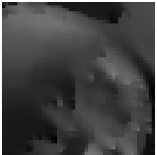














PSNR	SSIM	28.07	0.9506	24.53	0.8832	29.94	0.9345
plug & play							
		29.11	0.9646	24.89	0.8947	32.30	0.9776
one-time denoising							
ground truth							

Figure S3: Plug & play vs. one-time denoising.

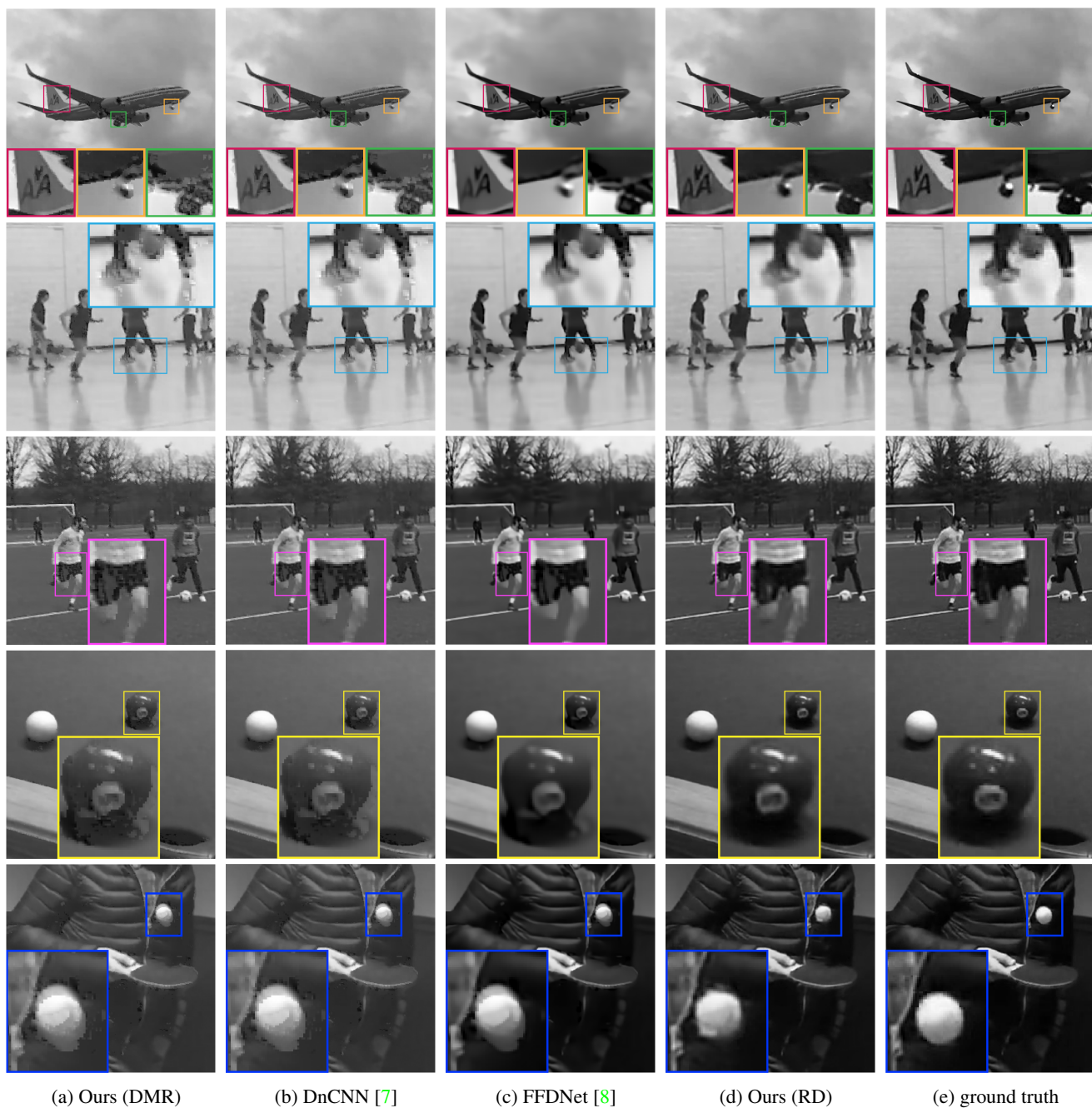


Figure S4: Comparison of denoising performance. Our learned Residual Denoiser (RD) reconstructs the intermediate frame (1-frame interpolation case) with fewer motion artifacts.