Supplementary Material for

Stereo Depth from Events Cameras: Concentrate and Focus on the Future

Yeongwoo Nam[1,2,*] Mohammad Mostafavi[3,*] Kuk-Jin Yoon[4] Jonghyun Choi[2,5,†]

[1]Saige Research [2]NAVER AI Lab. [3]Lunit [4]KAIST [5]Yonsei University

yw.nam@saigeresearch.ai, mostafavi@lunit.io, kjyoon@kaist.ac.kr, jc@yonsei.ac.kr

Note: We use blue color to refer to figures, tables, section numbers and citations **in the main paper** (*e.g.*, [17]). All red or green characters refer to figures, tables, section numbers and citations in this supplementary material.

1. Detail about Event Cameras (Section 2)

While conventional cameras present the scenes as frames, an event camera reports the scene as a stream of sparse and disconnected events in time, *i.e.*, per-pixel intensity changes larger than a predefined threshold. If a pixel value changes larger than a threshold, the event camera records a positive or negative sign value with the pixel location and the timestamp, asynchronously. The sign indicates if that pixel has received larger intensity changes than the threshold, *i.e.*, positive event, or its intensity for that pixel has dropped lower than another threshold, *i.e.*, minus event. Each event point in the stream is a quadruple; two for the pixel location, one for timestamp, and one for the sign. Each event is fired when it happens with very low latency, in the order of microseconds.

The asynchronous nature of events brings the unique capability of being immune to motion blur even under rapid scene changes and camera movements. The event camera also has a higher dynamic range that reveals scene details that ordinary cameras cannot sense and may miss. It also has additional benefits of low power consumption and low bandwidth requirements.

2. In-depth Comparison to [27] (Section 5)

Event and Intensity Fusion. In the state-of-the-art event intensity depth estimation method [27], their 'Recycling Network' is used for fusing events and intensity images. It is a light-weighted version of E2SRI [28,29] which is originally aimed for super-resolution. As their recycling network is based on a recurring neural network (RNN) structure [3], the data processing is not performed in parallel and is not time efficient to fuse the two modalities. Different to their method, the proposed concentration networks does not have any RNN structure, thus data is processed in parallel and is much faster. We reach almost $2\times$ inference speed in event-intensity design, *i.e.* 18.2 FPS in our design in comparison to 10 FPS in [27]. This is also presented in Table 1 of the main paper.

Intermediate Output of the Network. Our concentration network does not require any additional loss term as it only creates the appropriate event stack needed to solve the stereo matching problem. In contrast, the recycling network of the state-of-the-art model requires additional loss term such as LPIPS and L1 to produce image-like outputs by combining event stacks and intensity images to reconstruct a high dynamic range, blur-free image-like output.

Location to Fuse Events and Intensity Image. The recycling network combines event stacks and intensity images at the near-input level after a few convolution layers. As there is a domain gap between the event stack and the intensity image, combining two different visual signature at the input level may be problematic, especially if the stacked event information and its corresponding intensity image are not perfectly aligned due to warping errors. Instead of fusing the events and intensity image at the input level, we use a 1×1 convolution to perform feature-level fusion in the near-end features. As we combine the feature vectors, we relatively less suffer from the problems of combining input levels as our qualitative results suggest in Fig. 5 of the main paper and the supplemented Fig. 1 and Fig. 2.

3. Number of Events per Stack.

We investigate the effect of the number of events per event stack to the performance, and summarize results in Table. 1. While a large number of events in a stack would contain textural details, it also occurs overriding the previous events and even harms the visual details. If the number of events exceeds 1 million, the result is also bad as many

*: equal contribution. †: corresponding author. This work is done while YN and JC are an intern, AI tech advisor at NAVER AI Lab., respectively.

Table 1. **Effect of the number of events on depth estimation**. Starting from 250K events, the error metrics become lower when we stack more events, *i.e.*, 500K. But further adding events increases the error as the events start overriding previous events. The concentration network learns to find the best balance and creates an event stack with the most details that in return creates depth estimates with the least error values with a large gap to the manually set number of events.

Number of Events	MAE(↓)	1PE(↓)	2PE(↓)	RMSE(↓)
250,000	0.906	21.875	6.987	1.979
500,000	0.864	20.175	6.330	1.939
1,000,000	0.888	20.879	6.607	1.980
2,000,000	0.889	20.475	6.491	2.013
Concentrated Events	**0.831**	**18.875**	**5.757**	**1.880**

incoming new events overwrite previous events. But using a small number of events in a stack (*e.g.*, 250K) result in poor depth estimation, as few events would miss many details. Empirically, using 500K events the performance shows reasonable performance. In contrast, our concentrated stack outperforms them.

4. RGB and event fusion method ablation.

We concatenate RGB and event for the input to our model following our ablation settings as mentioned in the supplement.Our fusion significantly outperforms it (compare row 2 and 3) in all measures. Furthermore when comparing to the intensity only settings (row 4), our method (row 3) outperforms it in all metrics by noticeable margins.

Table 2. **RGB and event fusion ablation study**

Method	Modality	MAE(↓)	1PE(↓)	2PE(↓)	RMSE(↓)
Ours	E	0.797	18.053	5.369	1.799
Concatenate	E+I	0.505	8.345	1.800	1.279
Ours	E+I	**0.485**	**7.929**	**1.668**	**1.238**
EI-Stereo [2]	I	0.511	8.524	1.832	1.288

5. Additional Qualitative Results (Section 6.2)

We present additional qualitative results from the DSEC dataset. We compare our method to the baseline method [43] as the DSEC website challenge [1] kindly provided. By the courtesy of the authors of [27] to send their results submitted in the DSEC challenge to us, we compare qualitatively to their results for better understanding the differences. In the yellow highlighted boxes in the second row of Fig. 1, [43] does not recover the car. This is because the car is moving at a constant speed and few events occur. On the other hand, our method recovers the car relatively clearly. In addition, our results by the 'event-only' method (column-(d)) exhibit sharp and detailed edges when compared to the [43]; Bus stop in first, fourth row of Fig. 1, guard rails in second row of Fig. 2. Compared with [27] (column-(e)), our event-intensity method has fewer artifacts. These artifacts can generate quite large errors, such as highlighted boxes in fifth row of Fig. 1 and third row of Fig. 2. Furthermore, our event-intensity method (column-(f)) recovers further details such as the empty spaces of the guard rails in third row of Fig. 1.

References

[1] DSEC competition hosted by the cvpr 2021 workshop on event-based vision https://dsec.ifi.uzh.ch/cvpr-2021-competition-results. 2

[2] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021. 2

[3] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 1

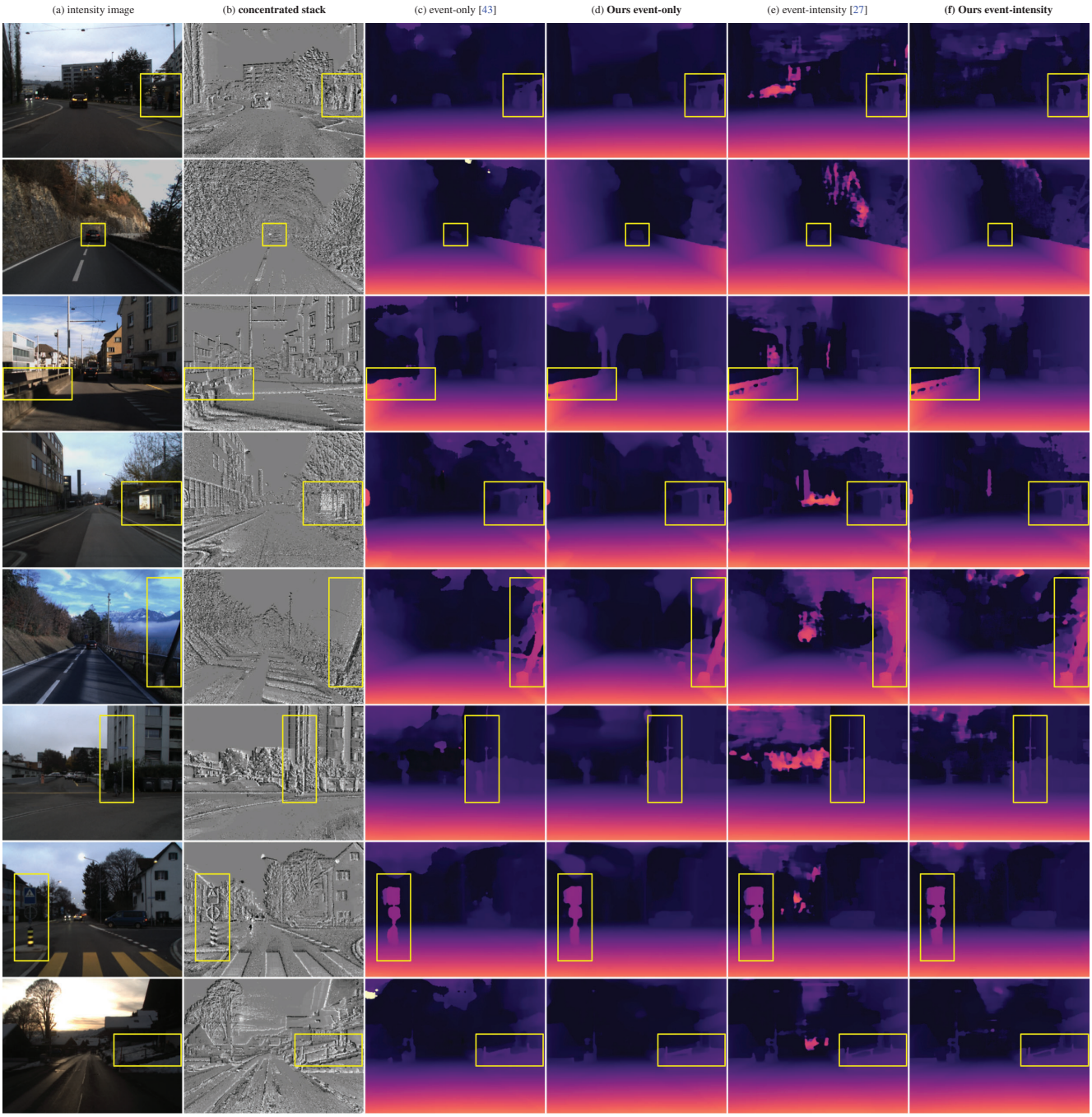

Figure 1. **Qualitative comparison on dense depth estimation**. We present our dense depth estimations using event-only (d) and events fused with intensity images (f) together with the (a) intensity image and (b) concentrated event stack for reference. We compare them to the (c) event-only [43] and (e) event-intensity [27] methods respectively.

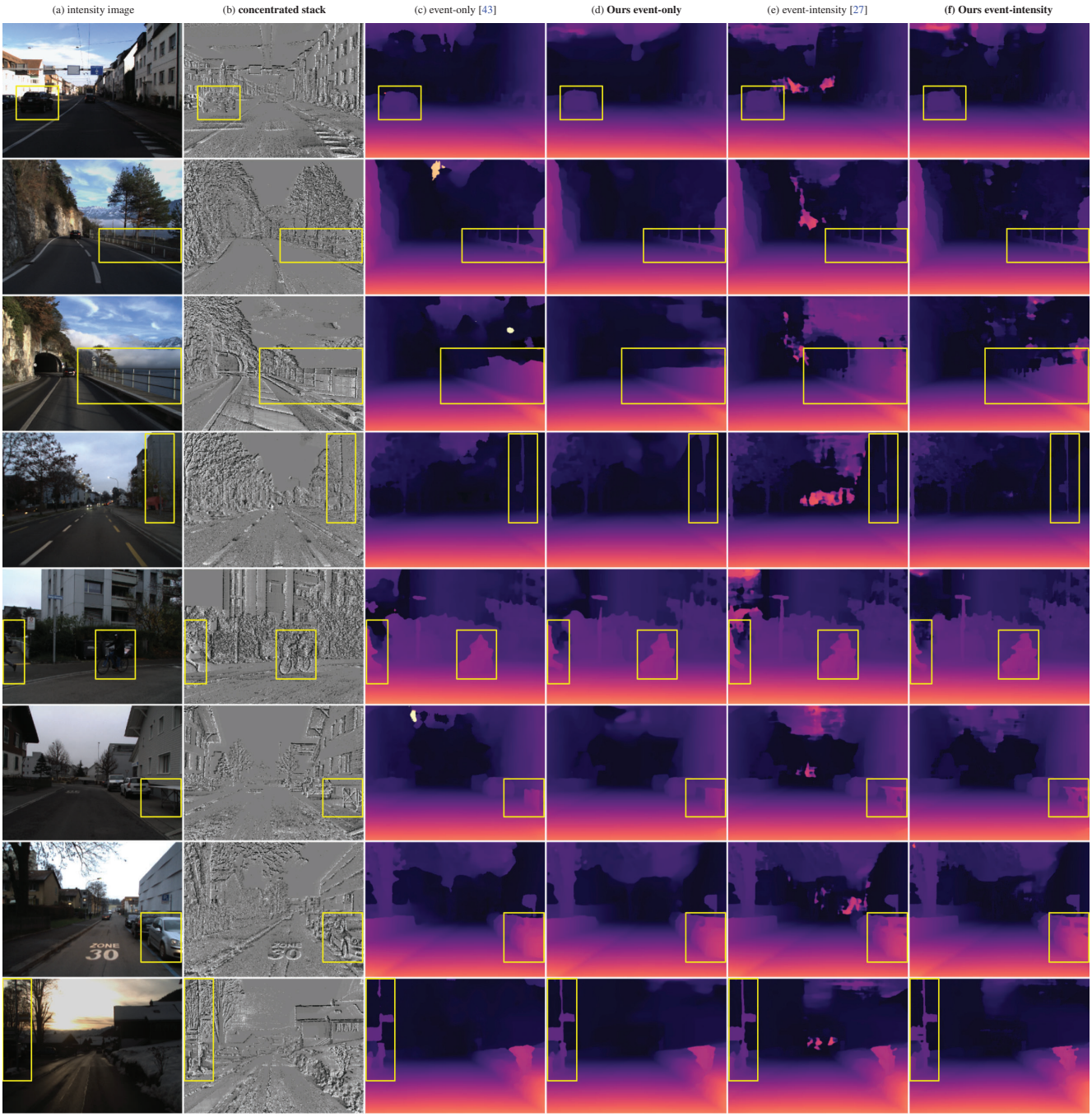

Figure 2. **More qualitative comparison on dense depth estimation**. We present our dense depth estimations using event-only (d) and events fused with intensity images (f) together with the (a) intensity image and (b) concentrated event stack for reference. We compare them to the (c) event-only [43] and (e) event-intensity [27] methods respectively.