# Unsupervised Joint Learning of Optical Flow and Intensity with Event Cameras

Supplementary Material

## 7. Additional Ablation Study

As an expansion of Sec. 4.4, we present the results of the ablation studies on loss weights and contrast threshold $C$.

### 7.1. Loss Weights

In addition to disabling some loss terms to show their effects (Sec. 4.4), an ablation study on the weight of CMax loss $\lambda_2$ is performed to show how the ratio between CMax and PhE influences model performance. The results are reported in the upper half of Tab. 7. It turns out that increasing or decreasing $\lambda_2$ does not lead to a further improvement in performance. Therefore, the choice of $\lambda_1$ and $\lambda_2$ to train our main model yields a sensible combination of the CMax and PhE loss terms.

### 7.2. Contrast Threshold

The contrast threshold $C$ can vary across event cameras and change even within the same dataset [41]. Hence, it is worth analyzing the influence of $C$ on model performance.

The works of [14, 15] have shown that the PhE and its linearized version are insensitive to the value of $C$, due to the PhE being calculated using thousands or millions of events instead of a few. The user just needs to set a mean value for $C$, and then the optimizer seeks a balanced motion and brightness to best explain the events. This finding has also been verified by the results in Fig. 4 in our main paper and Figs. 6 to 8 in this supplementary: our model was trained only on the DSEC dataset (Prophesee Gen3) with $C = 0.2$, but nevertheless it is able to predict precise optical flow and image intensity on several other datasets, such as ECD (DAVIS240C), HDR (Samsung DVS Gen3) and BS-ERGB (Prophesee Gen4).

Furthermore, we also train models with different $C$ values, and report the results in the lower half of Tab. 7. The results agree with the statements above; the accuracy remains approximately constant for different $C$ values.

| | | Flow | | | Intensity | | |
|---|---|---|---|---|---|---|---|
| Ablation | Value | EPE↓ | AE↓ | %Out↓ | MSE↓ | SSIM↑ | LPIPS↓ |
| $\lambda_2$ | 0.2 | 2.78 | 8.12 | 18.93 | 0.10 | 0.31 | 0.57 |
| | 5.0 | 2.34 | 9.73 | 16.42 | 0.10 | 0.31 | 0.56 |
| $C$ | 0.1 | 1.89 | 6.80 | 12.34 | 0.11 | 0.30 | **0.55** |
| | 0.4 | 1.88 | **6.42** | 12.32 | 0.10 | **0.32** | 0.56 |
| **Main** ($\lambda_2 = 1.0, C = 0.2$) | | **1.78** | 6.44 | **11.24** | **0.10** | 0.31 | 0.56 |

Table 7. Results of ablation studies on $\lambda_2$ and $C$.

## 8. DSEC: Training Sequence Selection

As mentioned in Sec. 5, all MB/USL methods rely on the brightness constancy assumption to estimate optical flow or image intensity. Events triggered by flickering lights or by hot pixels would undermine the estimation accuracy. To this end, the sequences in the training split of the DSEC dataset [13] are screened before being used for training, according to the data quality. Here we present the list of the selected training sequences in Tab. 8.

| | | | |
|---|---|---|---|
| zurich_city_02_a | zurich_city_02_b | zurich_city_02_c | zurich_city_02_d |
| zurich_city_02_e | zurich_city_03_a | zurich_city_04_a | zurich_city_04_b |
| zurich_city_04_c | zurich_city_04_d | zurich_city_04_e | zurich_city_04_f |
| zurich_city_05_a | zurich_city_06_a | zurich_city_07_a | zurich_city_08_a |
| zurich_city_11_a | zurich_city_11_b | interlaken_00_c | interlaken_00_d |
| interlaken_00_e | interlaken_00_f | interlaken_00_g | thun_00_a |

Table 8. Sequences from the DSEC training split that were used for training our model.

## 9. BS-ERGB: Evaluation Sequence Cropping

Here we describe the removal of the last seconds of the may29_rooftop sequences, for the intensity evaluation on the BS-ERGB [42] dataset (970 × 625 px resolution), as presented in Tab. 9. We perform the cropping because the camera is not moving in the last seconds (mentioned in Sec. 4.1) of those sequences, thus the recorded event data is pure noise.

| Sequence | Start Time [s] | End Time [s] |
|---|---|---|
| may29_rooftop_handheld_01 | 0.0 | 24.0 |
| may29_rooftop_handheld_02 | 0.0 | 17.0 |
| may29_rooftop_handheld_03 | 0.0 | 14.0 |
| may29_rooftop_handheld_05 | 0.0 | 9.5 |

Table 9. Details of the cropping of the may29_rooftop sequences in the BS-ERGB dataset.

## 10. Additional Qualitative Results

In this section, we present additional qualitative results of our model on the high-resolution BS-ERGB [42] datasets in Fig. 6, including (a) input events, (b) image of warped events (IWE) with the predicted flow, (c) optical flow, (d) image intensity and (e) reference images. A qualitative comparison between our method and other baselines on the same dataset is also presented in Fig. 7, where some regions

of interest (ROIs) are highlighted for further zoomed-in visualization in Fig. 8.

Figure 6 demonstrates that our model recovers precise optical flow and image intensity on unseen data (i.e., not used for training). In column c (optical flow), independent moving objects (IMOs) are clearly identified with respect to the background (e.g., the cars in the first row and the motorbike in the second row). Besides, flow discontinuities agree with the contours of different objects at different depths (e.g., the fence in the third and fourth rows and the bench in the last row). For image intensity, our model reconstructs fine details at the valid pixels (e.g., the contours of objects), while it leverages the total variation regularization to partially fill in the regions that lack texture and rarely trigger events.

Figure 7 confirms that our model produces competitive results compared to baseline methods. Our reconstructed images are overall sharper, and are more precise in HDR conditions. To highlight this, we select two HDR regions (i.e., the stairs of the building and the bench on the rooftop) and present their zoomed-in versions in Fig. 8. It can be clearly seen that: E2VID and SPADE-E2VID report poor HDR performance; FireNet produces intensity images with low contrast, where strong artifacts (like spider webs) caused by the rectification of event data using the camera's intrinsic parameters are clearly observed; BTEB's intensity images are blurred; ET-Net oversmooths the fine textures on the building wall, and shows strange wrong textures on the bench (especially the stone part in the bottom right). In contrast, our reconstructed intensity reveals sharp edges and fine details in HDR illumination, where frame-based cameras suffer from under/over exposure problems.

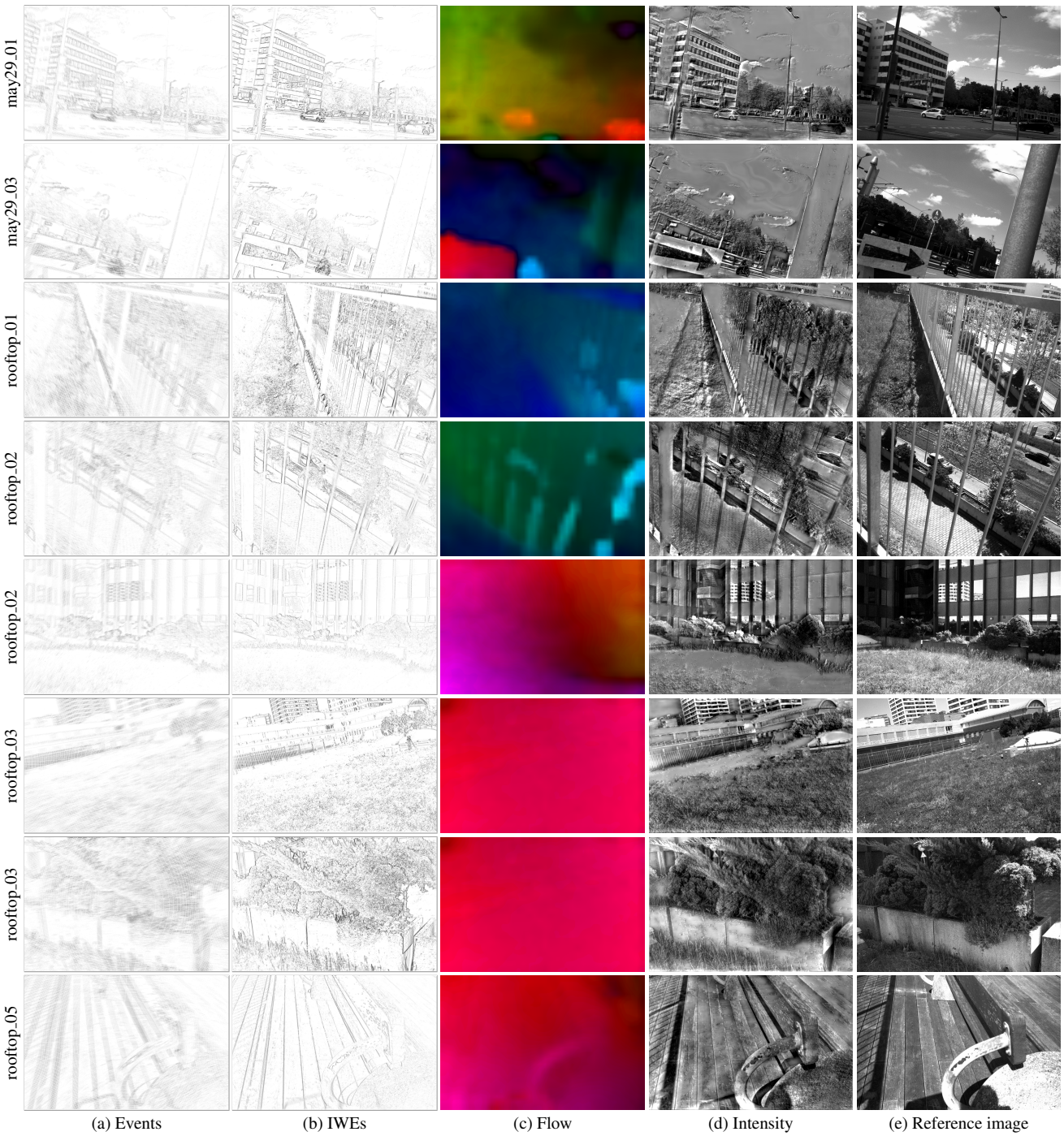


Figure 6. *Additional qualitative results on the BS-ERGB dataset*. From left to right: (a) input events; (b) image of warped events (IWE) with our predicted flow; (c) our predicted flow; (d) our predicted intensity; (e) reference image.

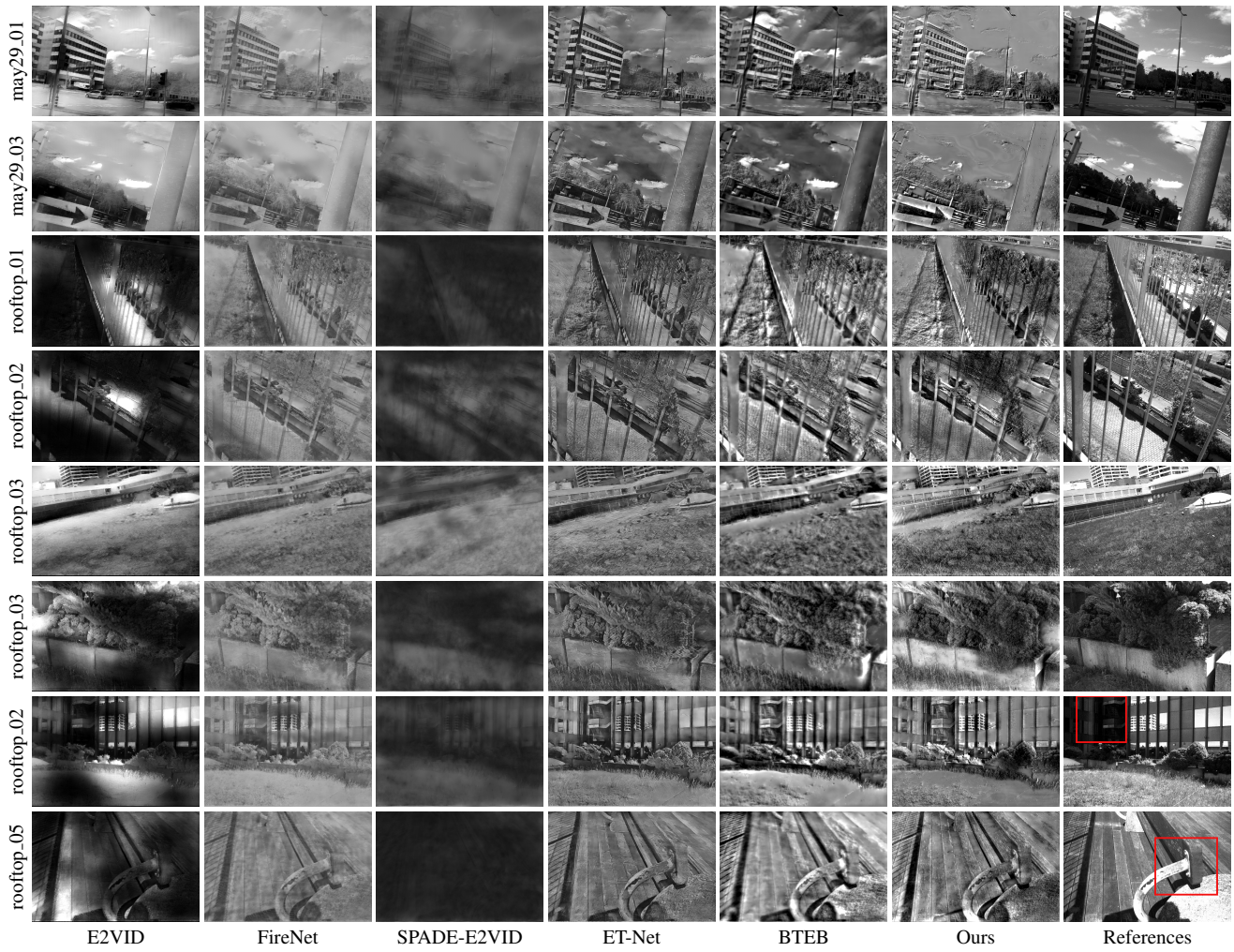


Figure 7. *Additional qualitative comparison of image intensity reconstruction on the BS-ERGB dataset*. For the last two rows, the regions marked by red boxes are enlarged in Fig. 8.

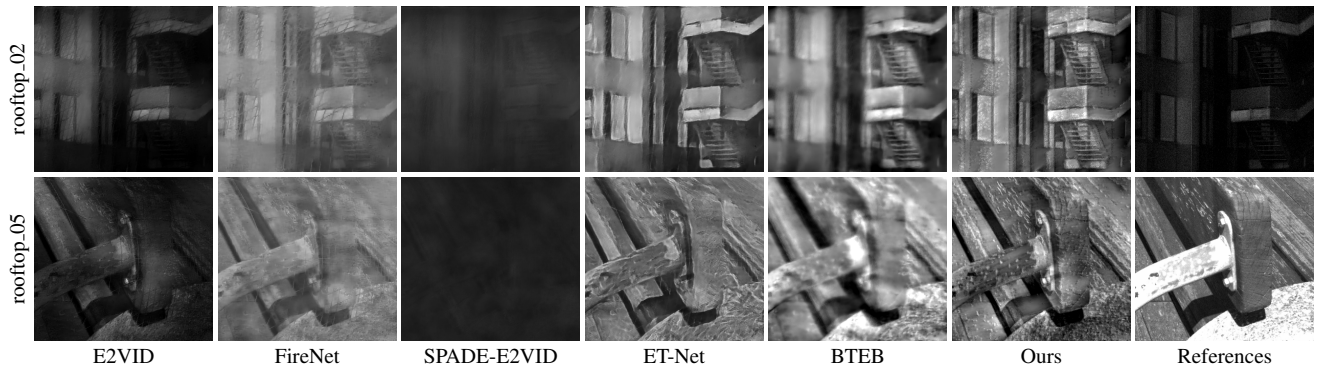


Figure 8. *Additional qualitative comparison of image intensity reconstruction*. Enlarged regions indicated by red boxes in Fig. 7.