Supplementary: ESS: Learning Event-based Semantic Segmentation from Still Images

Zhaoning Sun* Nico Messikommer* Daniel Gehrig Davide Scaramuzza

Robotics and Perception Group, University of Zurich, Switzerland

1. DSEC-Semantic

The existing semantic segmentation labels for DDD17 suffer from artifacts caused by the low-quality and low-resolution grayscale images. Thus, we generate a new dataset with a high quality semantic labels including events and frames. Compared to labels from DDD17, our labels feature much higher quality and more details, as visualized in Fig. 1. Our newly introduced event-based semantic segmentatation dataset, termed *DSEC-Semantic*, is constructed based on sequences of the large-scale DSEC [5] dataset, see Fig. 2. To generate the semantic labels, we first warp the images from the left frame-based camera with a resolution of 1440 × 1080 to the view of the left event camera with a resolution of 640 × 480. The last 40 rows are then cropped since the frame-based camera does not capture these regions. Thus, the the DSEC-semantic labels have a resolution of 640 × 440. In a second step, we apply a state-of-the-art semantic segmentation method [12] to the warped images to generate the labels. We use pre-trained weights provided by the author.

By doing so, we obtain fine-grained labels for 19 classes in the first place, which have the same classes than the Cityscapes labels for evaluation. We then further convert the 19 class labels into 11 classes (background, building, fence, person, pole, road, sidewalk, vegetation, car, wall, and traffic sign) for our experiments. Since frame cameras suffer from image degradation in challenging illumination scenes, we only label a subset of sequences of the DSEC dataset which are recorded during the day to ensure high-quality labels. For the training set, we labeled 8082 frames of the following sequences: 'zurich_city_00_a', 'zurich_city_01_a', 'zurich_city_02_a', 'zurich_city_04_a', 'zurich_city_05_a', 'zurich_city_06_a', 'zurich_city_07_a', 'zurich_city_08_a'. For the test set, we generated labels for 2809 frames of the following sequences: 'zurich_city_13_a', 'zurich_city_14_c', 'zurich_city_15_a'.

2. Event Representation

We convert an event stream $\mathcal{E}$ to a sequence of grid-like representations [4], such as *voxel grids* [13] $\mathbf{V}_k$. Each voxel grid is constructed from non-overlapping windows $\mathcal{E}_k$ each with a fixed number of events

$$\mathbf{V}_k(x, y, t) = \sum_{e_j \in \mathcal{E}_k} p_j \delta(x_j - x) \delta(y_j - y) \max\{1 - |t_j^* - t|, 0\}, \tag{1}$$

where δ is the Kronecker delta and $t_j^* = (B-1)\frac{t_j - t_0}{\Delta T}$ where B is the number of bins, ΔT is the time window of events and t_0 is the time of the first event in the window.

3. Network Architecture

Our network is a fully convolutional network inspired by the U-Net [10] architecture. We use an E2VID encoder E_{E2VID} and an E2VID decoder D_{E2VID} as illustrated in Fig. 4 of [9] with the pre-trained weights provided by the author. The E2VID encoder E_{E2VID} includes a head layer $\mathcal{H}$ and three recurrent encoder layers $\mathcal{E}^i$ with $(i = 0, 1, 2)$. We use the outputs of these three encoder layers as the recurrent, multi-scale embedding $\mathbf{z}_{\text{event}}$. The E2VID decoder D_{E2VID} consists of the remaining two residual blocks $\mathcal{R}^j$, three decoder layers $\mathcal{D}^l$, and the final images prediction layer $\mathcal{P}$. For the image encoder E_{img}, we use the first layers up to the sixth residual block of ResNet-18 [6] without the first max-pooling layer. We use the outputs of the second and fourth residual blocks as skip connections for the task network. The encoder weights are initialized with parameters from ImageNet [11]. The task network T consists of five residual blocks followed by seven convolution layers, and three upsampling layers lie in between. We use concatenation for the skip connection and nearest-neighbor interpolation with an upsampling factor of two for each upsampling layer.

4. Training Details

DDD17 For the experiments on DDD17, we use Cityscapes [3] as the labeled source domain and DDD17

[*] equal contribution.

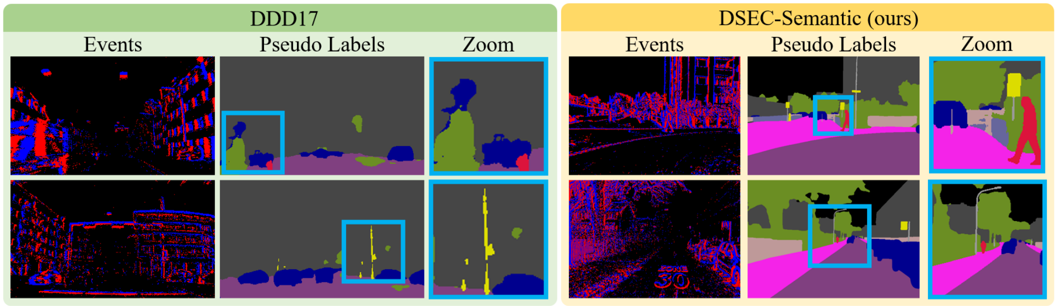

Figure 1. We release a new semantic segmentation dataset for the DSEC [5] dataset. The pseudo labels are constructed based on the RGB images and a state-of-the-art frame-based segmentation network [12]. Compared to DDD17 [1, 2] (left), our labels have a higher level of detail, seen in the zooms. Additionally, our dataset includes more classes (11 classes) compared to [1] (6 classes).

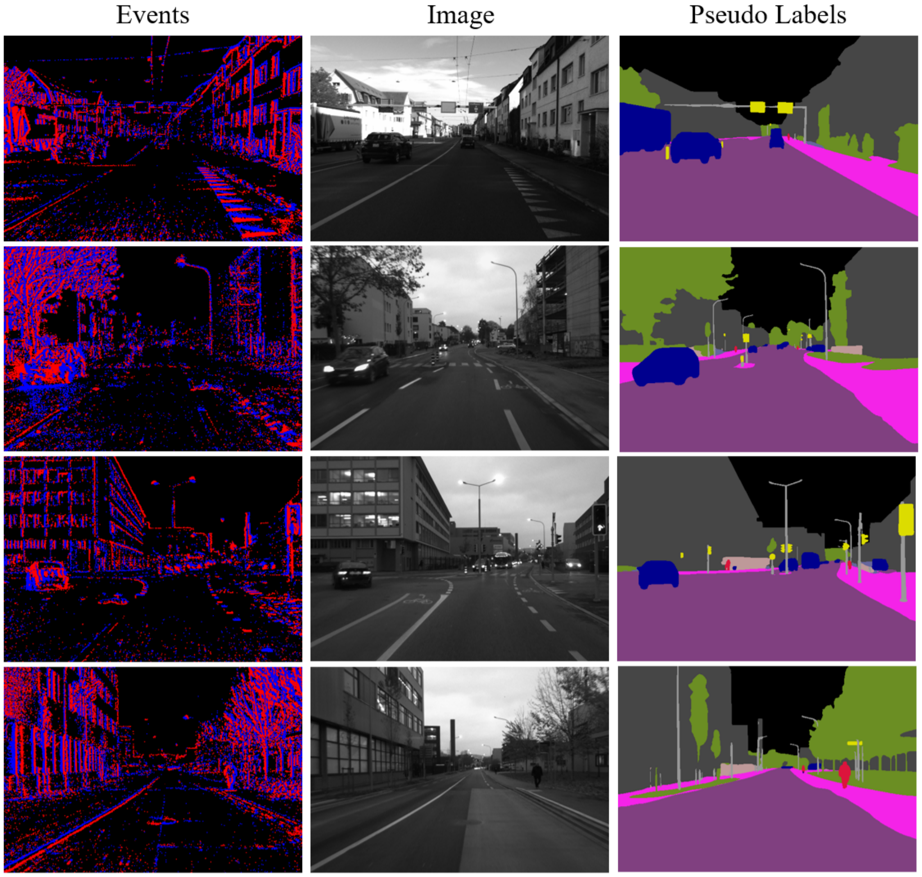

Figure 2. We release a new semantic segmentation dataset for the DSEC [5] dataset containing accurate and fine-grained labels. The pseudo labels are constructed based on the RGB images and a state-of-the-art frame-based segmentation network [12].

as the unlabeled target domain. For each sample, we convert the events into a sequence of 20 voxel grids, each with 32'000 events. The hyper-parameters λ_1, λ_2, λ_3, and λ_4 are set as 1, 0.01, 1, and 0.01, respectively. We set the learning rates as 1×10^{-5} for E_{img} and 1×10^{-4} for T. We empirically found that having a smaller learning rate on E_{img} and activating the accumulation of gradients for E_{img} in the first stage help improve the results. We train our model using the RAdam optimizer [7] with a batch-size of 16 for 50'000 iterations. Additionally, for the comparison with E2VID [9] in the UDA setting, we retrain the image encoder and task network (forming a U-Net) on grayscale images and labels from the Cityscapes dataset [3]. Similar to our method, we train [8] in our UDA setting with the same source and target domains.

DSEC-Semantic Similar to the experiments on DDD17, we leverage the Cityscapes datasets as the labeled source dataset. The difference is that we now use the DSEC-Semantic dataset as the target domain. We increase the number of events per voxel grid to 100'000 due to the higher resolution. To ensure the capturing of enough events at the beginning, we remove the first six samples of each sequence. For computational reasons, we further skip every second sample of a selected sequence, which results in a training set of size 4017 and a test set of size 1395. The hyper-parameters λ_1, λ_2, λ_3, and λ_4 are now set as 1, 1, 1, and 1, respectively. We use the same RAdam optimizer to train our model with a larger learning rate of 5×10^{-4} (for both E_{img} and T), and a smaller batch-size of 8, for 25'000 iterations.

5. E2VID Driving Dataset

To show that our method also works with completely unpaired and unlabeled data, we have applied it to the E2VID dataset [9], which contains driving sequences. The dataset features events recorded with a Samsung DVS Gen3, and images recorded with a Huawei P20. Both cameras were mounted behind a car windshield, however, neither a external calibration nor a time synchronization is available. Thus this dataset contains completely unpaired and unlabeled events. Nevertheless, our method can learn the task on the image of Cityscapes and transfer it to the E2VID dataset, as shown in Fig. 3.

6. Qualitative Results DSEC-Semantic

As stated in the main manuscript, we compare our method against EV-Segnet [1] in the supervised setting. Additionally, we also provide results for our method using both image and event labels during training. Without considering the image labels in the training, our method achieves a performance comparable with EV-Segnet, see Fig. 4 for qualitative samples. However, if we use the full potential of our method by using the image labels as well, we achieve state-of-the-art performance on DSEC-Semantic, outperforming EV-SegNet by 1.53% mIoU.

References

[1] Iñigo Alonso and Ana C Murillo. EV-SegNet: Semantic segmentation for event-based cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2019. 2, 3

[2] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: End-to-end DAVIS driving dataset. In *ICML Workshop on Machine Learning for Autonomous Vehicles*, 2017. 2

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016. 1, 2

[4] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Int. Conf. Comput. Vis. (ICCV)*, 2019. 1

[5] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. In *IEEE Robotics and Automation Letters*, 2021. 1, 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 770–778, 2016. 1

[7] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Int. Conf. Learn. Representations (ICLR)*, 2020. 2

[8] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. In *IEEE Robot. Autom. Lett.*, 2022. 2

[9] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1, 2, 3

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 1

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, Apr. 2015. 1

[12] Andrew Tao, Karan. Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. In *ArXiv*, 2020. 1, 2

[13] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 1

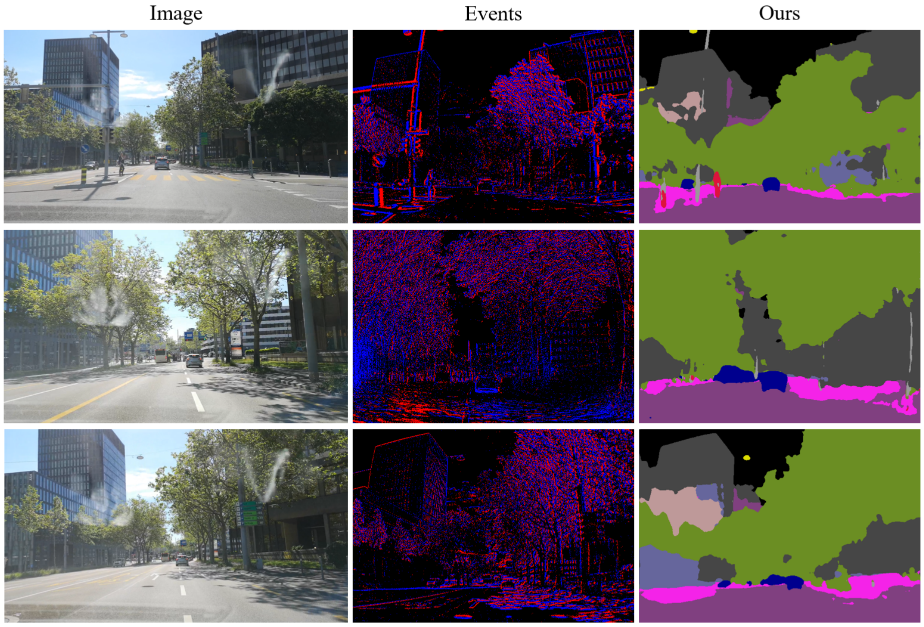

Figure 3. Qualitative samples on E2VID dataset for the UDA setting, i.e., no event labels are available during training. There are no synchronized and aligned image and events available, thus we have selected the image in the dataset closest to the scene captured by the events.

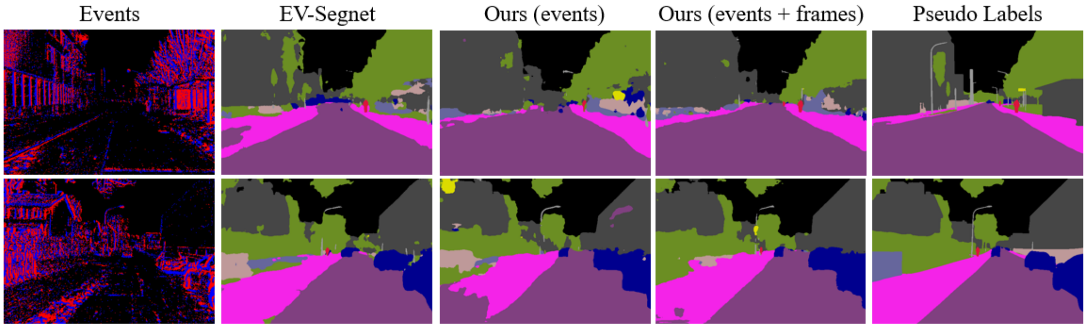

Figure 4. Qualitative samples on DSEC-Semantic in the supervised setting, i.e., event labels are available during training. The combined training on image and event labels improves the semantic predictions. Importantly, at test-time all methods only use events.