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

Zhaoning Sun*, Nico Messikommer*, and Daniel Gehrig, and Davide Scaramuzza

Dept. of Informatics, Univ. of Zurich and
Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich
zhasun@student.ethz.ch {nmessi,dgehrig,sdavide}@ifi.uzh.ch

1 DSEC-Semantic

Our newly introduced event-based semantic segmentatation dataset, termed *DSEC-Semantic*, is constructed based on sequences of the large-scale DSEC [3] dataset, see Fig. 1. 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 [10] 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'.
The dataset and detailed instructions are available at `https://dsec.ifi.uzh.ch/dsec-semantic/`

2 Event Representation

We convert an event stream $\mathcal{E}$ to a sequence of grid-like representations [2], such as *voxel grids* [11] $\mathbf{V}_k$. Each voxel grid is constructed from non-overlapping

* equal contribution

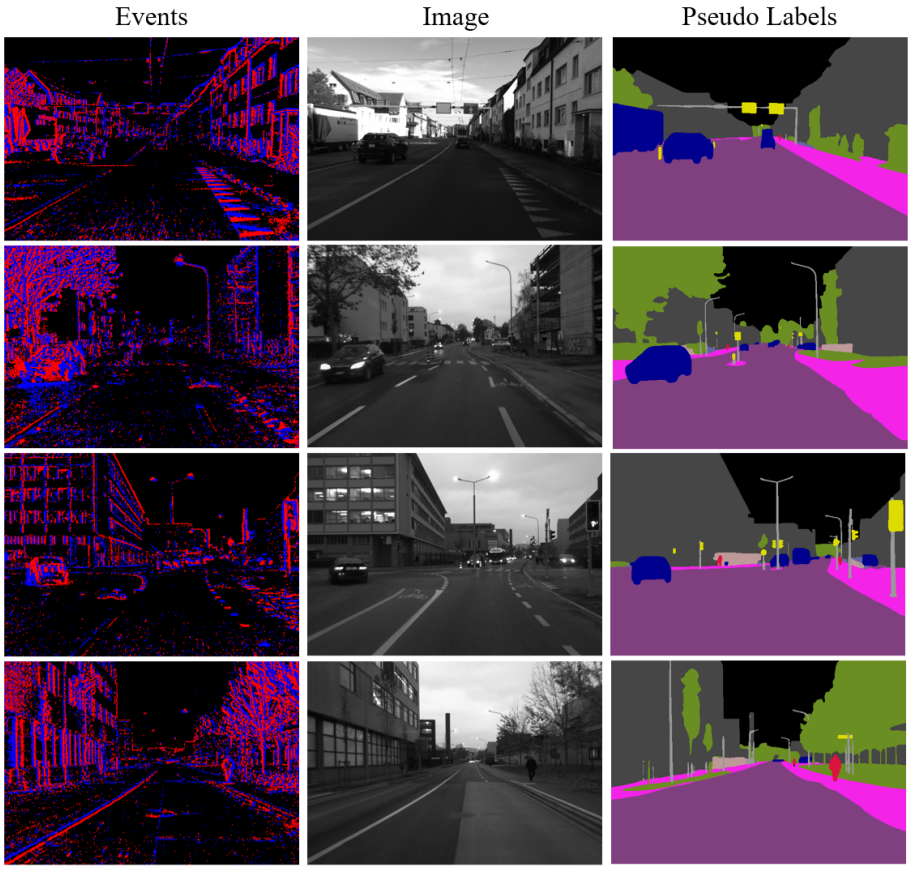

Fig. 1. We release a new semantic segmentation dataset for the DSEC [3] dataset containing accurate and fine-grained labels. The pseudo labels are constructed based on the RGB images and a state-of-the-art frame-based segmentation network [10].

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 [8] architecture. We use an E2VID encoder E_{E2VID} and an E2VID decoder D_{E2VID} as

illustrated in Fig. 4 of [7] 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 [4] 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 [9]. 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 [1] as the labeled source domain and DDD17 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 [5] with a batch-size of 16 for 50'000 iterations. Additionally, for the comparison with E2VID [7] 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 [1]. Similar to our method, we train [6] 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 [7], which contains driving sequences. The dataset features events recorded with a Samsung DVS Gen3, and

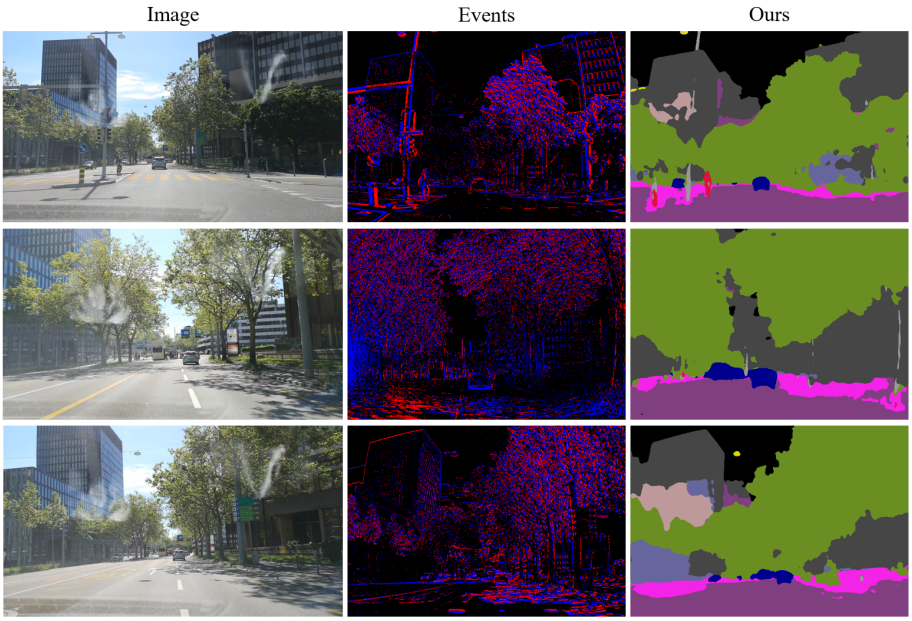

Fig. 2. Qualitative samples on E2VID dataset for the UDA setting, i.e., no event labels are available during training. There are no synchronized and aligned image and events available, thus we have selected the image in the dataset closest to the scene captured by the events.

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. 2.

References

1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2016)
2. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Int. Conf. Comput. Vis. (ICCV) (2019)
3. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. In: IEEE Robotics and Automation Letters (2021). https://doi.org/10.1109/LRA.2021.3068942
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/cvpr.2016.90

5. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: Int. Conf. Learn. Representations (ICLR) (2020)
6. Messikommer, N., Gehrig, D., Gehrig, M., Scaramuzza, D.: Bridging the gap between events and frames through unsupervised domain adaptation. In: IEEE Robot. Autom. Lett. (2022)
7. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Trans. Pattern Anal. Mach. Intell. (2019). https://doi.org/10.1109/TPAMI.2019.2963386
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2015)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Li, F.F.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (Apr 2015). https://doi.org/10.1007/s11263-015-0816-y
10. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. In: ArXiv (2020)
11. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2019)