Finding Meaning in Points: Weakly Supervised Semantic Segmentation for Event Cameras: *Supplementary Materials*

Hoonhee Cho*, Sung-Hoon Yoon*, Hyeokjun Kweon*, and Kuk-Jin Yoon

Visual Intelligence Lab., KAIST
{gnsgnsgml, yoon307, 0327june, kjyoon}@kaist.ac.kr

A Details about DSEC Night-Point

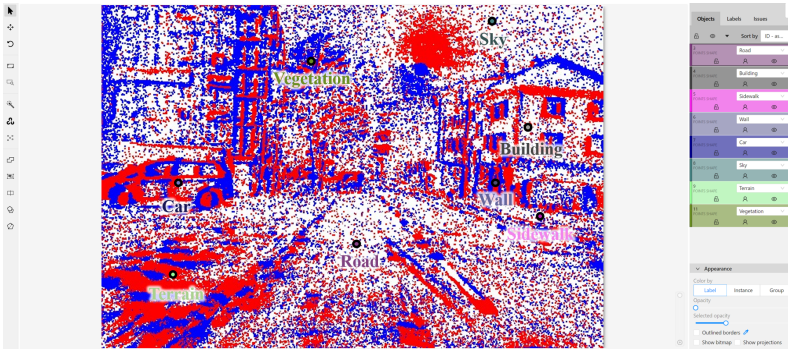

Fig. A1: Demonstration of point annotation process on DSEC-Night dataset.

The proposed Event-based Weakly Supervised Semantic Segmentation (EV-WSSS) method performs semantic segmentation with only one point-level supervision per existing class (referred to as "1-Class-1-Click" or 1C1C) for the given event voxel. We demonstrated the effectiveness of EV-WSSS through the DSEC-Semantic [8] dataset, which is a widely used dataset in event-based semantic segmentation. Furthermore, to emphasize the strength of weakly supervised approaches in situations where dense labeling using images is challenging, we validated our EV-WSSS in nighttime environments. For this, we constructed a novel dataset named DSEC Night-Point, composed of raw event data from DSEC-Night and weak labels newly annotated under our "1C1C" setting.

The detailed protocol for acquiring the DSEC Night-Point is as follows. Since the original DSEC-Night [5,9] dataset is densely annotated only for the *validation*

* Equal contribution.

set, we annotate point labels for the *train* set. The *train* set is divided into five splits according to the DSEC-Night [5,9] as 'Zurich City 09a', 'Zurich City 09b', 'Zurich City 09c', 'Zurich City 09d', and 'Zurich City 09e'. Each split respectively comprises 508, 109, 371, 478, and 226 frames, totaling 1,692 frames.

For the annotation process, we employ five computer vision experts, one per split. Before conducting the annotation, each annotator reviewed the whole visualized event stream, just like watching a video. This enables the annotators to acquire prior knowledge, *e.g.*, overall scene outline or objects existing in the scene, about the given split. Subsequently, the annotators are requested to annotate each event frame with the 1C1C setting. For this, we utilized an online tool (CVAT [7]) as shown in Fig. A1. To ensure labeling consistency between annotators in controversial regions, we incorporated discussions among annotators during the labeling process to reach a consensus. Further, as the accuracy of the point annotation is crucial, we request the annotators to ignore the confusing cases rather than provide a possibly incorrect label. After completing the labeling process, each annotator conducted cross-validation on the other splits.

Note that, throughout the annotation process, only the visualized event streams are exclusively used, unlike the labeling process [1] where paired images were used. However, our annotation process can further utilize images, if it possibly enables more accurate labeling.

We also investigate the burden of our annotation process. In our scenario focused solely on events, obtaining precise labeling can pose challenges, primarily because accurately discerning object boundaries can be difficult. Conversely, our weakly supervised approach with the 1C1C setting necessitates only a single point annotation for each class, involving a straightforward marking within the objects' interiors. Quantitatively, when labeling the DSEC-Night dataset under the 1C1C condition, an average of **50 seconds** per frame is required. Considering that the annotation time for driving scenes with dense pixel-level labels requires more than 1.5 hours [3] and more than 2 hours in the DSEC-Night dataset, it underscores the practicality and importance of weak labels in the context of event-based semantic segmentation.

B Hyperparameters Analysis

Table B1: Results according to different threshold (th).

th	0.3	0.4	0.5	0.6	0.7
mIoU	45.32	45.2	45.55	44.83	44.73

The proposed method involves two notable hyperparameters: the threshold (th) for dual-student learning in Eq. 4 and the temperature (β) for prototype-based contrastive learning in Eq. 7 of the main paper. To check the impact of the change in these parameters on the performance of our framework, we conduct

experiments while adjusting them. First, Table B1 shows the performance of the EV-WSSS framework with various threshold values. The results implied that the proposed method is robust to the threshold for the dual-student learning approach, underscoring the effectiveness of our framework. Further, we tested various temperature values (1.0 and 0.5, where 0.1 is our default). We confirm that the change in temperature does not have much effect on the final performance ($\pm 1\%$ in mIoU).

C Experiments on Incomplete and Noisy Annotation

Unlike the fully-supervised semantic segmentation approaches that can benefit from dense pixel-level GTs, our EV-WSSS only access weak supervision. Therefore, the accuracy of each point label is crucial, similar to the other pointly-supervised approaches in the imagery domain [2, 4, 6]. To verify that our method can perform robustly against the errors innated in weak labels, we conduct an experiment by modeling the possible source of noises involved during the annotation process. We model the noise as (I) absence of annotation and (II) wrongly annotated class.

C.1 Case (I): Incomplete Annotation

Table C2: Experimental result on 1C1C setting with incomplete annotation using DSEC datasets. $\mathcal{W}^{target}$ denotes the weak point-level GT in the target domain. 'Incomplete' indicates that a 10% drop rate was applied to the confusing classes (*wall*, *fence*, *person*, and *traffic sign*).

Method	Used GT Type	mIoU
	Weakly-supervised (1C1C)	
EV-WSSS	$\mathcal{W}^{\mathbf{target}}$	45.55
EV-WSSS	Incomplete $\mathcal{W}^{\mathbf{target}}$	44.10 (-1.4)

As mentioned in Section A, our annotation protocol for the 1C1C setting requests that the annotator ignore the confusing cases rather than provide a possibly incorrect label. Here, the degree of confusion can differ across different categories. For example, the frequent and large classes (*road*, *car*, etc.) are much easier to label than the confusing classes (*wall*, *fence*, *person*, and *traffic sign*). Considering the above, we model the incomplete annotation by discarding some point labels of the confusing classes from each event frame, with a drop rate of 10%. The results shown in Table C2 demonstrate that our EV-WSSS still achieves substantial segmentation performance, even with incomplete annotation. This highlights the practicality of the proposed weakly supervised approach.

C.2 Case (II): Wrong Annotation

Table C3: Experimental result on 1C1C setting with noisy annotation using DSEC.

p	0 (ours)	0.1	0.2
mIoU	45.55	44.73	44.33

Table C3 provides the results of additional experiments regarding case (II). To model the confusion between classes that possibly occurs during the annotation process, We randomly chose two point labels from each event stream and swapped their classes with a probability of p. Although performance declines as p increases, the reduction is not severe, suggesting the robustness of EV-WSSS against the annotation noise.

D Unsupervised Domain Adaptation (UDA) using Weak Labels in the Target Domain

Table D4 provides the results of ESS [8] with weak labels. On DSEC, which has a smaller domain gap with ESS's source, the additional use of weak labels indeed degrades performance compared to ESS alone. This implies that the naive use of weak labels may not provide a meaningful benefit over the gain from the source GT, highlighting the necessity of our approach even from the perspective of UDA. Meanwhile, on DSEC N-P, ESS itself struggles with a severe domain gap. Although the additional use of weak labels mitigates this to some extent, the achieved performance is still far lower than the weak baseline. Conversely, EV-WSSS consistently outperforms all of them. To sum up, all the results clearly demonstrate that our contribution lies in fully utilizing weak labels, rather than simply relying on them.

Table D4: Comparison with weakly supervised UDA methods.

Backbone	ESS [8]	ESS [8] + Weak Sup.	Weak Sup	Ours
DSEC	45.38	42.45	39.06	45.55
DSEC N-P	13.97	26.38	29.35	36.40

References

1. Alonso, I., Murillo, A.C.: Ev-segnet: Semantic segmentation for event-based cameras. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1624–1633 (2018), https://api.semanticscholar.org/CorpusID:54063435
2. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2617–2626 (2022)
3. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
4. Fan, J., Zhang, Z., Tan, T.: Pointly-supervised panoptic segmentation. In: European Conference on Computer Vision. pp. 319–336. Springer (2022)
5. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**, 4947–4954 (2021), https://api.semanticscholar.org/CorpusID:232170230
6. Li, W., Yuan, Y., Wang, S., Zhu, J., Li, J., Liu, J., Zhang, L.: Point2mask: Point-supervised panoptic segmentation via optimal transport. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 572–581 (2023)
7. Sekachev, B., Nikita, M., Andrey, Z.: Computer vision annotation tool: a universal approach to data annotation. Intel [Internet] **1** (2019)
8. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: Ess: Learning event-based semantic segmentation from still images. ArXiv **abs/2203.10016** (2022), https://api.semanticscholar.org/CorpusID:247593948
9. Xia, R., Zhao, C., Zheng, M., Wu, Z., Sun, Q., Tang, Y.: Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 21515–21524 (2023), https://api.semanticscholar.org/CorpusID:260333888