HPL-ESS: Hybrid Pseudo-Labeling for Unsupervised Event-based Semantic Segmentation Supplymentary

Linglin Jing[1,2*], Yiming Ding[1*], Yunpeng Gao[1,4], Zhigang Wang[1†], Xu Yan[3],
Dong Wang[1], Gerald Schaefer[2],Hui Fang[2†], Bin Zhao[1,4], Xuelong Li[1,5]
[1]Shanghai AI Laboratory, [2]Loughborough University, [3]Huawei Noah's Ark Lab,
[4]Northwestern Polytechnical University, [5]Institute of Artificial Intelligence (TeleAI)
l.jing@lboro.ac.uk, wangzhigang@pjlab.org.cn

1. Overview

This document provides supplementary materials to support the main paper. It includes more ablation studies and visualization cases on DSEC-Semantic and DDD17 datasets.

2. Ablation Studies

Other source dataset. We also evaluate our framework with the GTA5 dataset as the source domain, which contains 24,966 synthetic images and each image has 1914×1052 pixels. As a common setup, we resize the images in GTA5 to 1280×720 pixels. As illustrated in Table 1, using these two datasets as source domain exhibits similar event segmentation performance. This experiment demonstrates that our framework is not reliant on the specific characteristics of the source dataset but rather emphasizes learning from unlabeled events. Data visualizations for GTA5 and CityScapes are presented in Figure 1.

Figure 1. The dataset visualization of GTA5 and CityScapes.

Table 1. Ablation study for different source datasets.

Source	Accuracy [%]	mIoU [%]
GTA5	89.95	55.23
CityScapes	89.92	55.19

Warm-up iterations. As stated in our paper, the source data warm-up strategy is crucial for the final performance. We further investigated the effect of the warm-up iterations on performance. As depicted in Figure 2, insufficient warm-up leads to a significant reduction in performance. Once the iterations reach 5,000, the performance tends to stabilize, and further increases may lead to a slight degradation due to overfitting on the source domain.

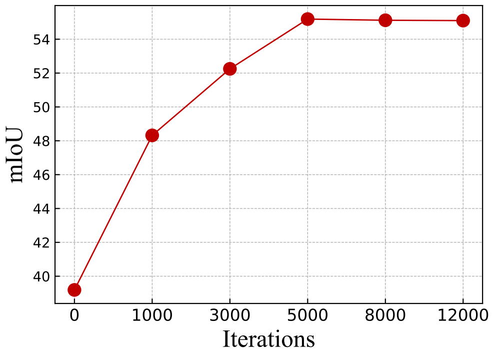

Figure 2. Ablation study on Warm-up iterations.

3. Visualization Analysis

3.1. t-SNE analysis for feature learning

To better develop intuition, we present t-SNE visualizations [1] of the learned representations for HPL-ESS on the DSEC-Semantic dataset. In Figure 3, it is evident that features among different categories are well-separated, indicating that the semantic distributions effectively provide the correct supervision signal for target data.

[*] Equal contribution.
[†] Corresponding author.

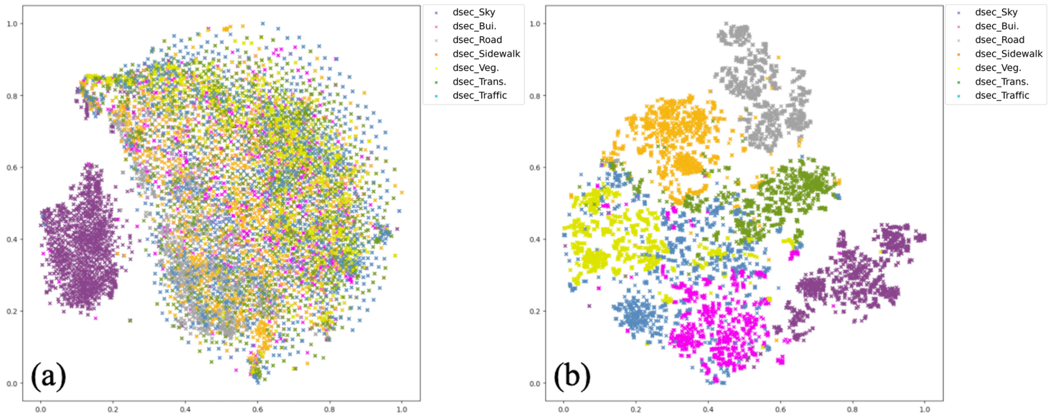

Figure 3. t-SNE analysis of HPL-ESS. (a) at initialization; (b) after training the full model.

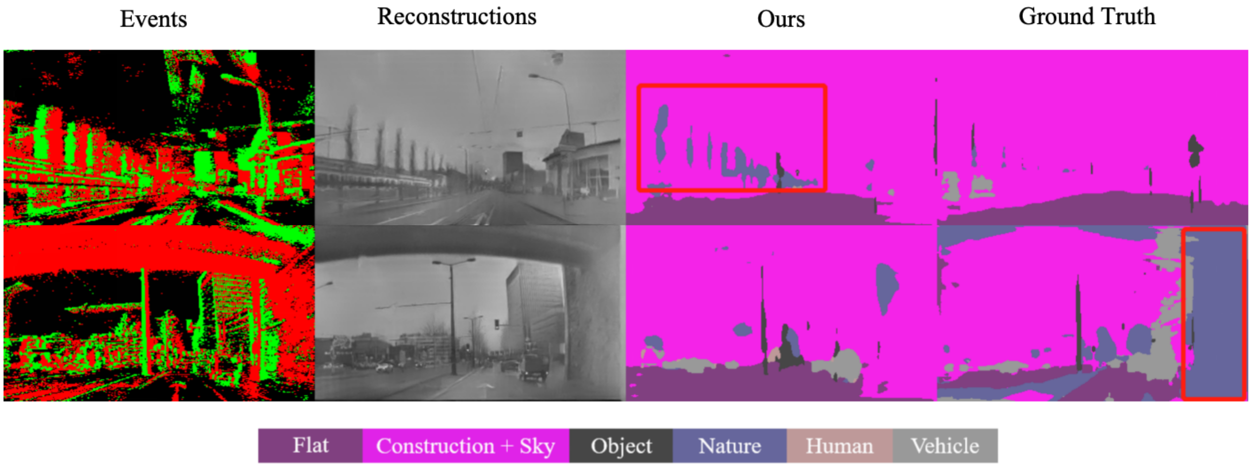

Figure 4. Examples on DDD17 Dataset.

3.2. More Segmentation Results

DDD17. As noted in our paper, the ground truth of DDD17 is impacted by low-quality images, limiting its ability to fully represent the actual scene. As Figure 4 illustrates, our method successfully recovers additional details, such as the plants in the red box of the first line. It's worth highlighting that the ground truth can also misclassify objects into other categories, *e.g.*, bridges in the second line of Figure 4.

DSEC-Semantic. Challenges persist in achieving accurate segmentation for imbalanced categories, notably 'person' and 'pole,' due to the inherent disparities within benchmark datasets. As depicted in Figure 5, these imbalances manifest in limitations when segmenting persons and signage details. Despite these challenges, our results significantly outperform those of other relevant works.

References

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

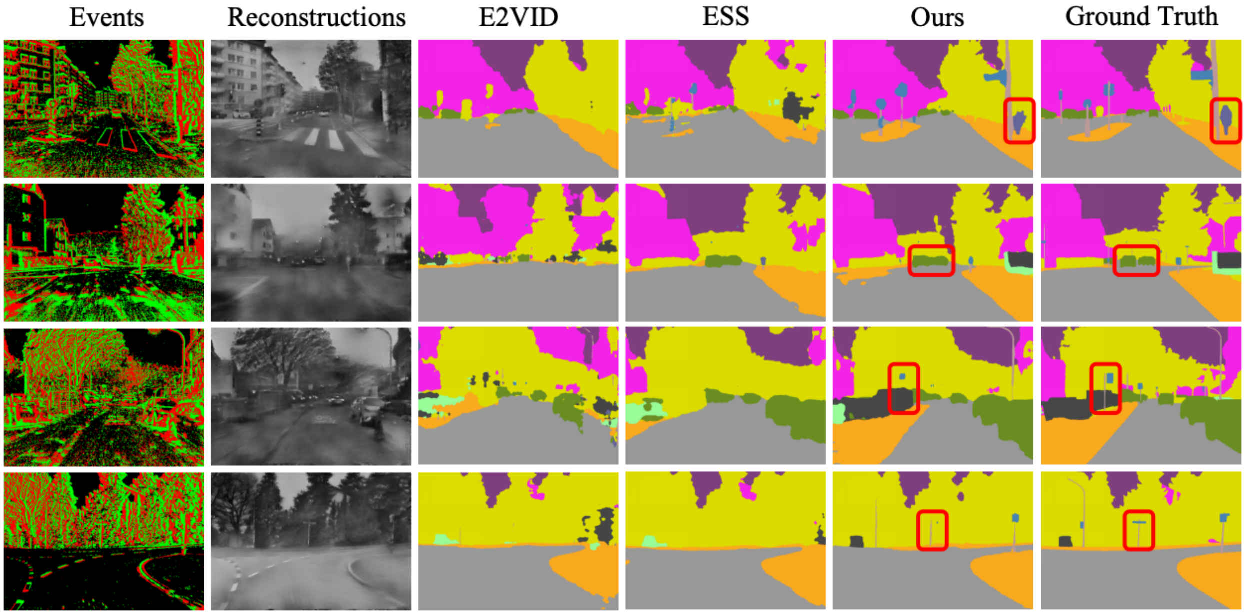

Figure 5. Examples on DSEC-Semantic Dataset.