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

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]SSE & FNII, CUHK-Shenzhen,
[4]Northwestern Polytechnical University, [5]Institute of Artificial Intelligence (TeleAI)
l.jing@lboro.ac.uk, wangzhigang@pjlab.org.cn

Abstract

*Event-based semantic segmentation has gained popularity due to its capability to deal with scenarios under high-speed motion and extreme lighting conditions, which cannot be addressed by conventional RGB cameras. Since it is hard to annotate event data, previous approaches rely on event-to-image reconstruction to obtain pseudo labels for training. However, this will inevitably introduce noise, and learning from noisy pseudo labels, especially when generated from a single source, may reinforce the errors. This drawback is also called confirmation bias in pseudo-labeling. In this paper, we propose a novel hybrid pseudo-labeling framework for unsupervised event-based semantic segmentation, HPL-ESS, to alleviate the influence of noisy pseudo labels. Specifically, we first employ a plain unsupervised domain adaptation framework as our baseline, which can generate a set of pseudo labels through self-training. Then, we incorporate offline event-to-image reconstruction into the framework, and obtain another set of pseudo labels by predicting segmentation maps on the reconstructed images. A noisy label learning strategy is designed to mix the two sets of pseudo labels and enhance the quality. Moreover, we propose a soft prototypical alignment (SPA) module to further improve the consistency of target domain features. Extensive experiments show that the proposed method outperforms existing state-of-the-art methods by a large margin on benchmarks (*e.g.*, +5.88% accuracy, +10.32% mIoU on DSEC-Semantic dataset), and even surpasses several supervised methods.*

1. Introduction

Event cameras are bio-inspired vision sensors that respond to changes in pixel intensity, generating a stream of asynchronous events characterized by exceptionally high temporal resolution. This technology enables the capture of dynamic scenes, providing features of high dynamic range (HDR) and reduced motion blur. Event cameras have been extensively applied in various applications, including object recognition [15, 23], SLAM [8], and autonomous driving systems [19], effectively addressing challenges such as motion blur and overexposure.

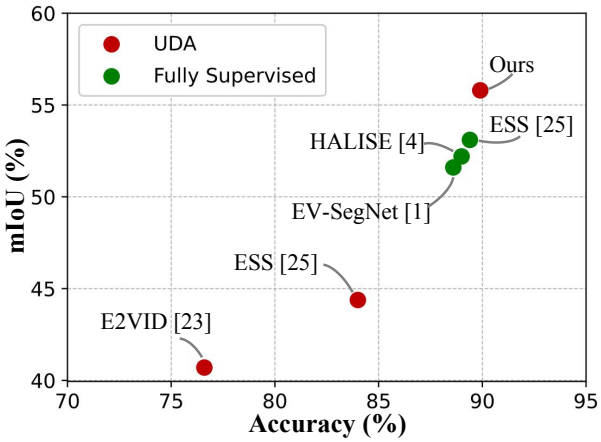

Figure 1. Comparison on the DSEC-Semantic dataset. Our method outperforms other UDA works by a large margin and even surpasses fully supervised methods.

However, event data significantly differ from images, making it difficult to annotate in dense pixel prediction tasks such as semantic segmentation. Previous works [1, 31, 32] require per-pixel paired events and images, and then leverage pre-trained networks on images to generate labels for event data. Although a more precisely paired and sharper image would naturally yield improved results, these methods increase the demands on capture devices. Other methods rely on event-to-image conversion to get rid of the need for ground-truth labels. E2VID [25] is an event-to-image

*Equal contribution.
†Corresponding author.

(ETI) reconstruction method to transform events into images, while VID2E [10] employs image-to-event (ITE) in reverse. Based on the above methods, a feasible strategy is to generate pseudo labels from converted images for event data. ESS [27] further employs unsupervised domain adaptation (UDA) to transfer knowledge from labeled image data (source domain) to unlabeled event data (target domain) through the bridge of event-to-image reconstruction.

Despite improvements, reconstruction-based methods suffer from the limitation that, due to the lack of texture information in event data, the reconstructed images usually have large fuzzy regions, inevitably introducing noise into the generated pseudo labels. Training on noisy pseudo labels has the risk of reinforcing the errors, especially when they are obtained from a single source, a problem that is known as confirmation bias [2] in pseudo-labeling.

To alleviate the bias of single-source pseudo labels, in this paper, we propose HPL-ESS, a hybrid pseudo-labeling framework for unsupervised event-based semantic segmentation. Our method is built upon a modified UDA framework, which executes self-training on the mixture of unpaired images and event data. The framework has the ability to generate a set of pseudo labels by directly predicting the event data. Simultaneously, we introduce offline event-to-image reconstruction into the framework, which generates another set of pseudo labels by predicting the reconstructed images. Through training on these hybrid pseudo labels, the network can progressively improve its ability to directly predict more accurate labels for event data. To gradually mitigate the impact of low-quality reconstructed images during training, we approach this challenge as a noisy label learning (NLL) problem. In this context, we distinguish between noisy data (reconstructed images) and clean data (original events). Then, we introduce a noisy-label adaptation process to further refine pseudo labels at each iteration. In addition, due to the large domain gap between image and event, the network is prone to produce dispersed features in the target domain [10]. To counteract this issue, we also design a soft prototypical alignment (SPA) module to learn the intrinsic structure of the target domain and address the dispersion of target features. As illustrated in Figure 1, the proposed method is very effective, outperforming other state-of-the-art UDA approaches by a large margin and even surpassing several fully supervised methods.

In summary, our contributions in this paper are:

- We propose a hybrid pseudo-labeling framework for unsupervised event-based semantic segmentation. This framework gets rid of event-to-image pairs and is robust to noisy pseudo labels.
- We design a soft prototypical alignment (SPA) module to enforce the network to generate consistent event features under the same class, forming a more compact feature space in the target domain.
- Extensive experiments on two benchmark datasets demonstrate that our method outperforms previous state-of-the-art methods by a large margin.

2. Related Work

2.1. Event-based Semantic Segmentation

Using deep learning, [1] first introduces event cameras to the semantic segmentation task, with an architecture based on an encoder-decoder CNN, pre-trained on the well-known urban environment Cityscapes dataset [6]. An open dataset, DDD17, containing annotated DAVIS driving records for this task is released in [3]. [10] enables the use of existing video datasets by transforming them into synthetic event data, facilitating the training of networks designed for real event data. Despite its capacity to leverage an unlimited number of video datasets, challenges persist due to the sim-to-real gap in many simulated scenarios. [32] employs two student networks for knowledge distillation from the image to the event domain. However, the method heavily depends on per-pixel paired events and active pixel sensor (APS) frames. Consequently, in scenarios where APS frames are unavailable, the application of such a knowledge distillation approach becomes significantly restricted. [31] substitutes the active pixel sensor modality with grayscale images generated by E2VID [25], transferring the segmentation task from the event domain to the image domain. Recently, ESS [27] addresses event-based semantic segmentation by introducing the DSEC-Semantic dataset, which relies on paired high-resolution images and events, thus providing high-quality semantic labels for event streams. ESS also introduces an event-to-image-based UDA method to transfer knowledge from the source image domain to the target event domain.

2.2. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) approaches can be divided into two key methodologies: domain adversarial learning and self-training. Domain adversarial learning focuses on aligning feature distributions across domains [9] but does not inherently ensure the discriminative power of target features [18]. In contrast, self-training capitalizes on a model's high-confidence predictions to bolster the performance within the target domain. This approach significantly alleviates the domain shift issue by iteratively aligning the feature distribution of the target domain to match that of the source domain, which proves to be particularly effective in scenarios where obtaining labels for the target domain is challenging. In this context, strategies such as leveraging domain-invariant features [5, 12, 17], pseudo-labeling [36, 38], intermediate domains [16, 21, 30], and consistency regularisation [13] have been used. We consider that under a similar task and scenario, event data can

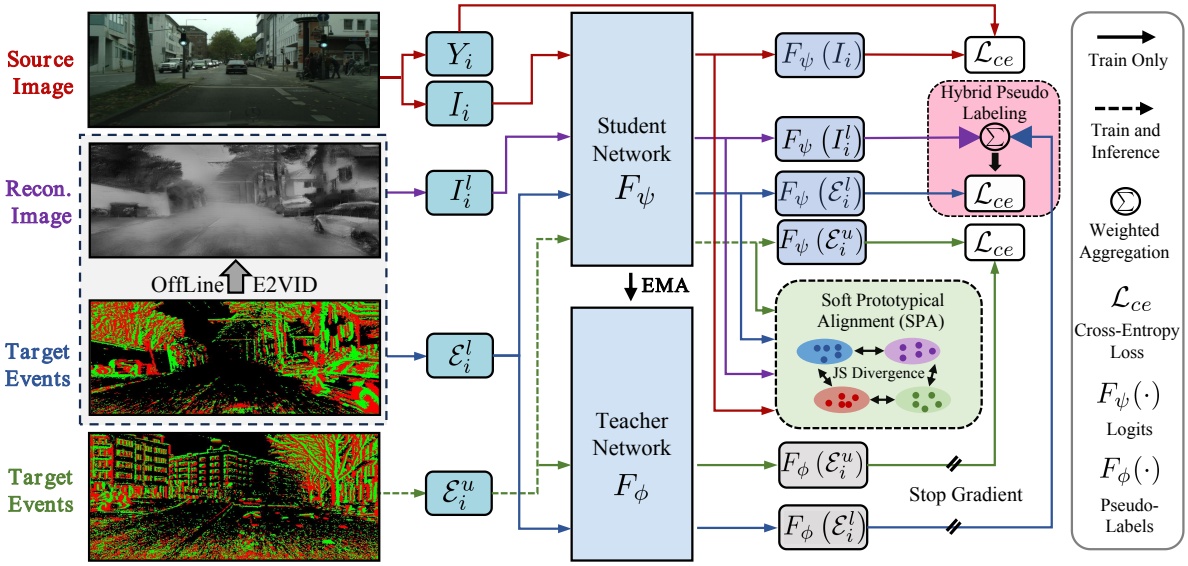

Figure 2. Overview of the HPL-ESS architecture. During training, we introduce offline event-to-image reconstruction as input to our framework. To avoid overfitting noise, we use only a small proportion (5%) of the reconstructions. The network is trained by hybrid pseudo labels from reconstruction and self-prediction. Additionally, a soft prototypical alignment (SPA) module is designed to enhance the consistency of target domain features. In the inference phase, only events are used as input.

also be drawn close to RGB images semantically through the application of UDA methods.

3. Method

As illustrated in Figure 2, the proposed HPL-ESS framework incorporates self-training UDA techniques, described in Section 3.2, and employs offline event-to-image reconstruction to generate hybrid pseudo labels, covered in Section 3.3. To gradually mitigate the impact of low-quality and blurred areas in offline-reconstructed images, we introduce a noisy label learning (NLL) method to refine pseudo labels. We further propose a soft prototypical alignment (SPA) module to explore the intrinsic structure of event data, alleviating the impact of feature divergence as detailed in Section 3.4.

3.1. Definitions and Problem Formulation

In a UDA framework for event-based semantic segmentation, a neural network F is usually trained from labeled source dataset $\mathcal{S} = \{I_i, Y_i\}_{i=1}^{M}$ to transfer to an unlabeled target dataset $\mathcal{T} = \{\mathcal{E}_i\}_{i=1}^{N}$. Specifically, the source domain S consists of images $I_i \in \mathbb{R}^{H\times W}$ and their corresponding labels $Y_i \in \mathbb{R}^{H\times W}$. In contrast, the target domain T consists of numerous continuous and asynchronous event streams $\mathcal{E}_i$ and without having access to the target labels V_i. Each event stream $\mathcal{E}_i$ can be represented as a series of tuples $\{(x_j, y_j, t_j, p_j)\}$, where j denotes the sample index, x, and y denote the spatial co-ordinates, t represents the timestamp, and p indicates the binary polarity (positive or negative) of brightness changes occurring between two timestamps. Due to the high temporal resolution of $\mathcal{E}_i$, we subsample $\mathcal{E}_i$ into a sequence of voxel grid representations [37], where each voxel grid is constructed from non-overlapping temporal windows with a fixed number of events. These are then effectively superimposed to form a static frame.

3.2. UDA Framework Overview

We modify DaFormer [14] as the backbone and baseline for our event-based semantic segmentation UDA method. The framework is composed of two networks: a teacher network F_ϕ and a student network F_ψ. Other modules in DaFormer are eliminated to ensure the simplicity and efficiency of our method. To facilitate knowledge transfer from the source domain to the target domain, the modified baseline is trained using the mixed data of labeled images and unlabeled events. To be specific, in our work, the student network F_ψ first conducts warm-up by being trained with the supervised loss on the source image domain

$$\mathcal{L}_s(F_\psi \mid \mathcal{S}) = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} H\left(F_\psi\left(I_i\right), Y_i\right), \tag{1}$$

where H denotes the cross entropy function. Correspondingly, the parameters of the teacher network are updated using the exponential moving average (EMA) [29] from the student model to maintain stability. After warm-up, the framework follows a self-training strategy, where the

teacher network directly predicts the event data to generate pseudo labels for the training of the student model. This process is repeated until the networks have converged.

In addition, augmentation methods, such as jitter and ClassMix [22], are used on both events and images to improve the method's availability across domains. Although self-training UDA is usually an effective technique, it is challenging to obtain satisfactory results due to the large domain gap between images and events. Furthermore, it suffers from the aforementioned single-source noisy pseudo labels.

3.3. Hybrid Pseudo-Labeling

To address the above issues, we consider the E2VID [26] method to reconstruct the event streams into simulated images, which are then incorporated into our framework as an intermediate domain to narrow down the gap between the source image domain and the target event domain. The reconstructed images also provide another set of pseudo labels to alleviate the bias of single-source pseudo labels. In particular, we randomly sample the event dataset $\mathcal{T} = \{\mathcal{E}_i\}_{i=1}^{N}$ to create two groups, $\mathcal{T}_l = \{\mathcal{E}_i^l\}_{i=1}^{a}$ and $\mathcal{T}_u = \{\mathcal{E}_i^u\}_{i=1}^{b}$, where $a + b = N$. Event streams $\mathcal{E}_i^l$ are reconstructed into simulated images I_i^l as

$$I_i^l = E2VID\left(\mathcal{E}_i^l\right). \tag{2}$$

Now, the inputs to the student network encompass source images I_i, unlabeled events $\mathcal{E}_i^u$, unlabeled events $\mathcal{E}_i^l$ and the corresponding reconstructed images I_i^l. Notably, we do not reconstruct all event streams into simulated images to avoid the network overfitting these noisy data.

As illustrated in Figure 2, the student network F_ψ takes the reconstructed image I_i^l as input and generates the predicted probability map. This map is then utilized as $F_\psi(I_i^l)$, the pseudo-ground-truth for $\mathcal{E}_i^l$. Simultaneously, similar to the self-training backbone, event data $\mathcal{E}_i^u$ and $\mathcal{E}_i^l$ are input to the teacher network F_ϕ to obtain the direct pseudo labels $F_\phi(\mathcal{E}_i^u)$ and $F_\phi(\mathcal{E}_i^l)$. For $\mathcal{E}_i^u$, the student network F_ψ is trained with the supervised loss $\mathcal{L}_u$ calculated as

$$\mathcal{L}_u(F_\psi \mid \mathcal{T}_u) = \frac{1}{|\mathcal{T}_u|}\sum_{i=1}^{|\mathcal{T}_u|} H\left(F_\psi\left(\mathcal{E}_i^u\right), F_\phi(\mathcal{E}_i^u)\right). \tag{3}$$

For $\mathcal{E}_i^l$, $F_\phi(\mathcal{E}_i^l)$ together with the event pseudo-ground-truth $F_\psi(I_i^l)$ constitutes the hybrid pseudo labels.

The event-to-image reconstruction process suffers from limited interpretability and a lack of control, leading to low-quality reconstructed images I_i^l, *e.g.*, incorrect content and blurred areas. Predicting semantic segmentation maps on these images and viewing them as pseudo labels will inevitably introduce significant noise. Directly using them during training may result in sub-optimal performance. Therefore, we treat this as a noisy label learning problem and explicitly regard $F_\psi(I_i^l)$ as a noisy label of the events $\mathcal{E}_i^l$. Inspired by [35], we employ a label correction strategy based on self-prediction to mitigate the noise issue. This strategy adapts the noisy distribution from the pseudo-ground-truth $F_\psi(I_i^l)$ to the view of the event distribution. Specifically, for each $\mathcal{E}_i^l$, we reconstruct the refined pseudo label $\hat{V}_i^{\,l}$ by combining $F_\psi(I_i^l)$ and the $F_\phi(\mathcal{E}_i^l)$ as

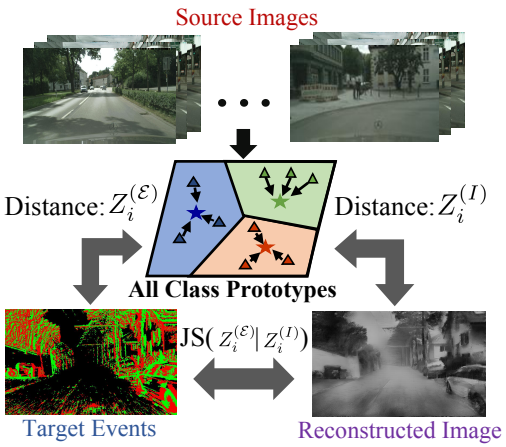

Figure 3. The concept of our SPA module on source domain, reconstructed images, and events.

$$\hat{V}_i^{\,l} = (1-\alpha)F_\psi(I_i^l) + \alpha F_\phi(\mathcal{E}_i^l), \tag{4}$$

with a hyper-parameter α. Then, the modified loss L_l for $\mathcal{E}_i^l$ is

$$\mathcal{L}_l(F_\psi \mid \mathcal{T}_l) = \frac{1}{|\mathcal{T}_l|}\sum_{i=1}^{|\mathcal{T}_l|} H\left(F_\psi\left(\mathcal{E}_i^l\right), \hat{V}_i^{\,l}\right). \tag{5}$$

With the progression of training, the teacher network becomes increasingly potent, thereby gradually generating more accurate $F_\phi(\mathcal{E}_i^l)$ and weakening the impact of $F_\psi(I_i^l)$ in Eq. (4)

3.4. Soft Prototypical Alignment

Although the reconstructed images I_i^l and the event E_i^l belong to the same target domain, we find there is a large distribution gap between them as shown in Figure 4(a). To enhance the consistency of target features, inspired by [36], we propose a soft prototypical alignment (SPA) module to explicitly align the distributions by pulling I_i^l and E_i^l to the same prototype, respectively. As illustrated in Figure 3, we employ each class's mean value $F_\psi(I_i)$ on source images as prototypes η since the source domain can provide much cleaner data. Then, we reduce the soft relative difference between reconstruction-to-source distance and event-to-source distance to achieve our goal. The reconstruction-

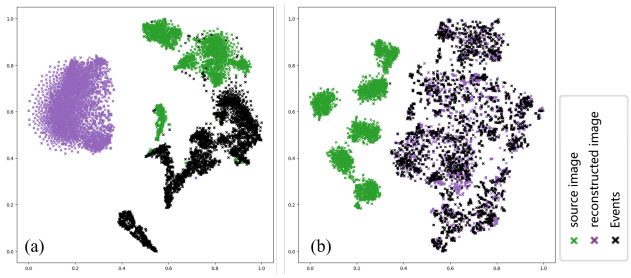

Figure 4. t-SNE analysis of SPA module on the source image, reconstructed image and target events. (a)Without SPA; (b) With SPA.

to-source distance between $F_\psi(I_i^l)$ and η is calculated as

$$Z_i^{(I)} = \frac{\exp\left(-\left\|F_\psi\left(I_i^l\right) - \eta\right\|/\tau\right)}{\sum \exp\left(-\left\|F_\psi\left(I_i^l\right) - \eta\right\|/\tau\right)}, \tag{6}$$

where τ is the coefficient temperature. Similarly, the event-to-source distance between $F_\psi(\mathcal{E}_i^l)$ and η is calculated as

$$Z_i^{(\mathcal{E})} = \frac{\exp\left(-\left\|F_\psi\left(\mathcal{E}_i^l\right) - \eta\right\|/\tau\right)}{\sum \exp\left(-\left\|F_\psi\left(\mathcal{E}_i^l\right) - \eta\right\|/\tau\right)}. \tag{7}$$

We use the Jensen-Shannon (JS) divergence [7] instead of KL divergence used in [24] for distribution alignment due to the symmetry of JS divergence. This ensures an equal pulling effect on the distributions of $F_\psi(I_i^l)$ and $F_\psi(\mathcal{E}_i^l)$. The JS divergence loss is calculated as

$$\mathcal{L}_{JS}^S = \mathrm{JS}\left(Z_i^{(I)} \| Z_i^{(\mathcal{E})}\right), \tag{8}$$

and compels the network to generate consistent event features and image features for $\mathcal{E}_i^l$ and I_i^l under the same class.

Additionally, $\mathcal{E}_i^l$ and $\mathcal{E}_i^u$ are not trained with the same set of pseudo labels, making the target distributions of $\mathcal{E}_i^l$ and $\mathcal{E}_i^u$ more likely to be dispersed. In such a scenario, the network fails to rectify the labels of target data located at the far end of the class cluster. Considering that the distributions of $F_\psi(\mathcal{E}_i^l)$ and $F_\psi(\mathcal{E}_i^u)$ belong to the same scene, their distributions are expected to exhibit similar relative distances. To achieve this, we further employ the mean value of each class in $F_\psi(I_i^l)$ as a prototype. We then bring in the relative distances of $F_\psi(\mathcal{E}_i^l)$ and $F_\psi(\mathcal{E}_i^u)$ to $F_\psi(I_i^l)$, respectively. Employing a methodology akin to Eqns. (6), (7), and (8), we obtain $\mathcal{L}_{JS}^I$ that forms a more compact feature space in the target domain. Figure 4(b) illustrates that SPA effectively minimizes the target domain distribution distance, aligning it with the relative distance observed in the source domain.

The overall loss in our framework is defined as

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u + \mathcal{L}_l + \omega(\mathcal{L}_{JS}^S + \mathcal{L}_{JS}^I), \tag{9}$$

where ω denotes a hyper-parameter.

4. Experiments

4.1. Dataset

As target data, we evaluate the proposed framework on two event-based semantic segmentation datasets, namely DSEC-Semantic [11] and DDD17 [3]. These driving-focussed datasets were captured using automotive-grade event cameras, encompassing a diverse range of urban and rural settings.

The DDD17 dataset comprises per-pixel paired events and frames captured by DAVIS event cameras with a resolution of 346×260. In [1], semantic labels were generated using pre-trained segmentation networks based on DAVIS images, resulting in 15,950 samples for training and 3,890 for testing. Due to the low resolution, several categories in DDD17 have been merged into six classes, namely flat (road and pavement), background (construction and sky), object, vegetation, human, and vehicle.

DSEC-Semantic, a recently introduced dataset for event-based semantic segmentation, extends the comprehensive DSEC dataset [11]. It includes 53 driving sequences captured by an event camera at a resolution of 640×480. [27] used a state-of-the-art image-based segmentation method [28] to generate segmentation labels. This process yields 8,082 labeled training samples and 2,809 testing samples, distributed across 11 classes: sky, building, fence, person, road, pole, sidewalk, vegetation, vehicle, wall, and traffic sign.

As source data, we use the CityScapes street scene dataset [6], which includes 2,975 training and 500 validation images with a resolution of 2048×1024. Following common practice in UDA methods, we resize the CityScape images to 1024×512 pixels.

4.2. Implementation Details

In our experiments, we employ DaFormer [14] as our UDA backbone. The encoder in Daformer uses an MiT-B5 model [34] and is pre-trained on ImageNet-1K. Across all experiments, the batch size is consistently set to 4. We use the AdamW optimizer with a weight decay of 1×10^{-4}. The learning rate is set to 6×10^{-5} and we use a learning rate warm-up for 1,500 iterations, with a linear increase in the learning rate during this period. We additionally warm-up for 5,000 iterations on the source dataset to make the network gain the initial semantic segmentation ability. α in Eq. 4 and ω in Eq. 9 are both set to 0.5. For data augmentation in both source and target domains, we follow [14, 33] and employ techniques such as color jitter, Gaussian blur, and ClassMix [22]. These augmentations are instrumental in training the model to learn more robust features across different domains.

In the event-to-image simulation process, Spade E2VID [25] is employed as our emulator for reconstruction.

Table 1. Performance and necessary number of events on DSEC-Semantic dataset in both UDA and fully supervised learning settings.

Type	Method	No. of Events	Accuracy [%]	mIoU [%]
Supervised	EV-SegNet [1]	-	88.61	51.76
	HALISE [4]	-	89.01	52.43
	ESS [27]	2E6	89.37	53.29
UDA	EV-Transfer [20]	2E6	60.50	23.20
	E2VID [25]	2E6	76.67	40.70
	ESS [27]	2E6	84.04	44.87
	Ours	**1.8E5 (↓ 91.0%)**	**89.92 (+ 5.88%)**	**55.19 (+10.32%)**

This step occurs solely in the offline phase, ensuring that it does not impact the efficiency of our online training and testing process. It is worth noting that E2VID will progressively produce expanding black artifacts if fed with discontinuous event inputs. To mitigate this problem, we reinitialize the E2VID network each time an image is reconstructed, preventing the occurrence of such artifacts. Regarding the event pre-processing on the DDD17 dataset, events are converted into 20 voxel grids, with each grid containing 32,000 events. For the DSEC-Semantic dataset, due to its higher resolution, the number of voxel grids is increased to 40, and each grid comprises 100,000 events.

4.3. Comparison with State-of-the-Art

We compare our method with previous relevant approaches, and use the top-1 accuracy and the mean intersection over union (mIoU) as the common semantic segmentation evaluation metrics. Beyond UDA methods, certain approaches have embraced a fully supervised setting to tackle the challenges. EV-SegNet [1] presents the first baseline for event-based semantic segmentation, which employs an encoder-decoder architecture and takes only events for fully supervised learning. HALISE [4] encodes event frames and source images into a spike stream, representing information in a binarised manner, and aligns the feature distribution in these spike streams. EV-Transfer [20] fabricates the motion of a still image to generate event streams, and then uses source labels and the corresponding synthetic events to conduct training. E2VID [25] converts events in the DSEC-Semantic dataset to reconstruct images, and then predicts semantic segmentation maps using other pre-trained models. E2VID can only perform direct transfer as there is no event label for training. VID2E [10] converts source video frames to synthetic events and trains on the source labels. ESS [27] employs the above E2VID-based process to generate pseudo-labels and attempts to transfer knowledge from the source image domain to the target event domain by the UDA technique. While methods employing supervised learning may achieve superior results compared to traditional UDA approaches, their reliance on labels significantly elevates the demands for dataset collection.

DSEC-Semantic dataset. We employ the CityScape dataset as the labeled source dataset and the DSEC-Semantic dataset as the unlabeled target dataset. This dataset poses additional challenges due to its more fine-grained categories compared to the DDD17 dataset. We report the obtained results for all methods in Table 1.

As we can see from there, our method demonstrates a significant improvement, outperforming the previous state-of-the-art UDA work ESS by 5.87% and 9.65% in terms of accuracy and mIoU, respectively. Notably, our UDA-based method even surpasses the performance of fully supervised approaches by 0.55% in terms of accuracy and 1.9% in terms of mIoU. Since this is a highly imbalanced dataset, the gain in mIoU is more representative in the segmentation task. In addition, by solely utilizing the E2VID reconstruction method offline, our approach avoids dependency on recurrent networks in E2VID during both training and inference, significantly reducing the required input events from 2E6 to 1E5 (a 95% reduction). These enhancements remarkably prove the effectiveness and computational efficiency of our proposed method.

Some example results are visualized in Figure 6. The background of the reconstructed image exhibits fuzzy regions and low resolution, which inevitably poses significant challenges to semantic segmentation Networks. For instance, due to the lack of texture information in events, the reconstructed sky category appears very similar to the

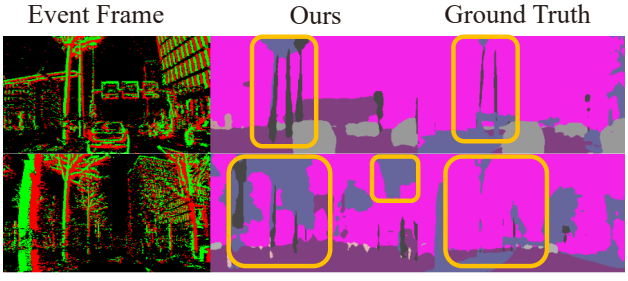

Figure 5. Example results on DDD17 dataset. The DDD17 ground truth lacks details for some objects.

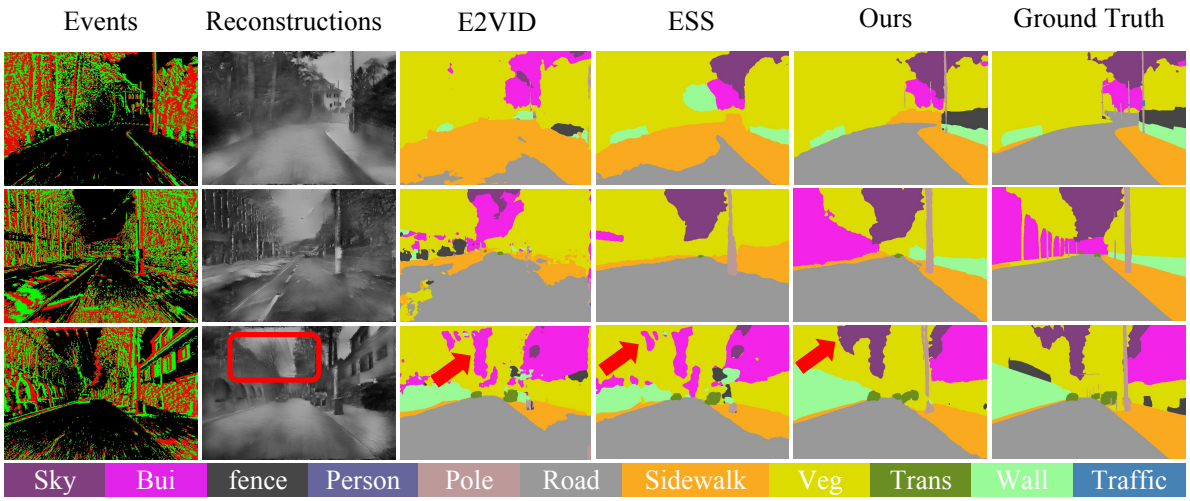

Figure 6. Visualization results on DESC-Semantic dataset. From left to right: event frame, event-to-image reconstruction, the maps predicted by E2VID, ESS, and our proposed HPL-ESS, ground truth.

building category in terms of contrast and edge information, leading to potential misinterpretation of the model's predictions (as indicated by the red arrow). The proposed hybrid pseudo-labeling method effectively mitigates these interference factors in reconstructed images, resulting in improved performance.

DDD17 dataset. Table 2 reports the UDA results on the DDD17 dataset for event-based semantic segmentation. Similar to the DSEC-Semantic dataset, only labeled images from CityScape and unlabeled events from DDD17 are available in this task. Table 2 showcases that our method achieves consistent optimal results, outperforming the previous state-of-the-art work by 1.05% (mIoU) and 0.79% (accuracy), respectively.

Since the event ground truth in DDD17 is derived from the low-quality paired images, this significantly impacts the reliability, especially concerning texture details, as also mentioned in [27]. As Figure 5 illustrates, our predictions even surpass the ground truth in object details. Taking the first line in Figure 5 as an example, in the yellow box, our network better separates the details of the trees and streetlights, which are missing in the DDD17 ground truth. Similarly, in the second line, our method segments more correct trees, while the DDD17 ground truth misclassifies trees with sky. This discrepancy could potentially lower our performance during evaluation. Due to the higher resolution and quality of the DSEC-Semantic dataset, we opted for this dataset to evaluate our method and comparison works.

4.4. Comprehensive Analysis

Since DSEC-Semantic is a higher-quality dataset, all ablation experiments are conducted on DSEC-Semantic.

Table 2. Performance comparison of HPL-ESS with state-of-the-art methods on DDD17 dataset in UDA setting. Only source labels are available.

Method	Accuracy [%]	mIoU [%]
EV-Transfer [20]	47.37	14.91
E2VID [25]	83.24	44.77
VID2E [10]	85.93	45.48
ESS [27]	87.86	52.46
Ours	**88.65 (+0.79%)**	**53.51+(1.05%)**

Design analysis of our framework. We conduct several ablation studies to assess the effectiveness of the proposed framework. As depicted in Table 3, (a) directly applying the UDA baseline alone does not yield satisfactory results, likely due to the substantial domain gap between the image and event domains. Similarly, (b) training directly on the event-to-image (ETI) reconstructed images from E2VID also results in unsatisfactory performance. This result verifies the aforementioned discussion, namely event-to-image-based methods will suffer from the noise brought by the reconstructed image. Both (a) and (b) highlight the unreliability of solely relying on the single-source pseudo labels and emphasize the necessity for hybrid label learning. An intriguing observation is that (c) employing the source data to pre-train the network for a certain number of iterations, *i.e.*, using a warmup phase, significantly enhances the performance. In (d), our method is based on source domain warm-up, and as described in Section 3, the E2VID reconstructed images are introduced on top of the UDA backbone to provide the hybrid pseudo labels for events, leading to considerable performance gains.

Table 3. Ablations study on DSEC-Semantic dataset.

Method	Baseline	ETI	Warmup	NLL	SPA	mIoU [%]
(a)	✓					36.76
(b)		✓				40.70
(c)	✓		✓			44.87
(d)	✓	✓	✓			51.08
(e)	✓	✓	✓	✓		52.23
(f)	✓	✓	✓		✓	52.69
HPL-ESS	✓	✓	✓	✓	✓	**55.19**

Table 4. Ablation study for the proportion of event samples participating in the event-to-image reconstruction.

Proportion	Accuracy [%]	mIoU [%]
0%	82.71	44.87
1%	86.54	46.84
5%	**89.91**	**55.19**
10%	89.89	55.15
50%	89.81	54.96
80%	89.75	54.82
100%	89.63	54.51

We further validate the effectiveness of the proposed NLL strategy and SPA module. As shown in Table 3, NLL reduces the noise of pseudo-labels on reconstructed images through iterative label refinement, making it more adaptive to the event domain and resulting in a performance gain. SPA prioritizes the divergence of various features on the target domain and aligns the labeled and unlabeled events with the source domain prototype, contributing to enhanced evaluation performance. Ultimately, the simultaneous introduction of these two modules in our framework leads to optimal performance.

Proportion of reconstructed event samples. In our framework, we do not transform all event data into reconstructed images to avoid overfitting the reconstruction noise. In fact, as demonstrated in Table 4, the optimal result is achieved when using only 5% of the event data to generate the reconstructed images as the pseudo labels. Performance experiences a slight decline as more reconstructed images are introduced. Particularly, when using 100% of the data, it results in an mIoU drop to 54.51%. The lower dependence on the number of reconstructed images also underscores the remarkable computational efficiency of our method during training. Further reduction of the ratio, below 5%, leads to progressively worse performance, reaching its lowest point at a 0% ratio and reverting back to the UDA backbone.

Online/Offline reconstruction pseudo label. Offline event-to-image reconstruction enables us to directly predict the reconstructed image using a pre-trained network and get pseudo labels for event data, which are named reconstruction pseudo labels here. In this section, we compare the effects of fixing these reconstruction pseudo labels and iteratively repredicting them by our network during training. As shown in Table 5, the online reprediction strategy remarkably surpasses the offline fixing strategy, demonstrating that our method becomes more powerful during training and can predict more accurate reconstruction pseudo labels for event data.

Table 5. Ablation study for online and offline reconstruction pseudo labels.

Method	Accuracy [%]	mIoU [%]
OffLine	83.25	48.45
OnLine	**89.92**	**55.19**

5. Discussion and Limitation

Due to the imbalance issue presented in the benchmark datasets, the accuracy performance of classes with insufficient samples, *e.g.*, 'rider' and 'traffic light,' is comparatively lower than the accuracy of some other classes, *e.g.*, sky and road. These results are illustrated in the visualization examples in **Supplementary Materials**. Despite that our approach yields significant improvement for the classes with a small number of samples when compared to previous methods, we will further consider more strategies to deal with the data imbalance issue in the future work.

6. Conclusion

In this paper, we have proposed a novel hybrid pseudo-labeling framework HPL-ESS for unsupervised event-based semantic segmentation. HPL-ESS effectively alleviates the challenges posed by noisy pseudo labels, a common issue in this field. The proposed method uniquely incorporates self-training unsupervised domain adaptation and offline event-to-image reconstruction to generate high-quality hybrid pseudo labels. The introduction of a noisy label learning strategy further refines the pseudo labels gradually. Moreover, a soft prototypical alignment (SPA) module significantly enhances the consistency and reliability of the target features. The effectiveness of HPL-ESS is evidenced by its superior performance in extensive experiments, where it not only surpasses existing state-of-the-art UDA methods but also exceeds several supervised methods.

7. Acknowledgments

This work is supported by the Shanghai AI Laboratory, National Key R&D Program of China (2022ZD0160101), the National Natural Science Foundation of China (62376222), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001). This work is supported by 111 Project (No. D23006).

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 5, 6

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8, 2020. 2

[3] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 2, 5

[4] Shristi Das Biswas, Adarsh Kosta, Chamika Liyanagedera, Marco Apolinario, and Kaushik Roy. Halsie–hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. *arXiv preprint arXiv:2211.10754*, 2022. 6

[5] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021. 2

[6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5

[7] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34: 30284–30297, 2021. 5

[8] Guillermo Gallego, Jon EA Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza. Event-based, 6-dof camera tracking from photometric depth maps. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2402–2412, 2017. 1

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[10] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 2, 6, 7

[11] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954, 2021. 5

[12] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International conference on machine learning*, pages 222–230. PMLR, 2013. 2

[13] Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Unsupervised domain adaptation with label and structural consistency. *IEEE Transactions on Image Processing*, 25(12):5552–5562, 2016. 2

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 3, 5

[15] Linglin Jing, Yifan Wang, Tailin Chen, Shirin Dora, Zhigang Ji, and Hui Fang. Towards a more efficient few-shot learning-based human gesture recognition via dynamic vision sensors. In *BMVC*, page 938, 2022. 1

[16] Linglin Jing, Ying Xue, Xu Yan, Chaoda Zheng, Dong Wang, Ruimao Zhang, Zhigang Wang, Hui Fang, Bin Zhao, and Zhen Li. X4d-sceneformer: Enhanced scene understanding on 4d point cloud videos through cross-modal knowledge transfer. *arXiv preprint arXiv:2312.07378*, 2023. 2

[17] Huafeng Li, Yiwen Chen, Dapeng Tao, Zhengtao Yu, and Guanqiu Qi. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Information Forensics and Security*, 16:1480–1494, 2020. 2

[18] Fengmao Lv, Jun Zhu, Guowu Yang, and Lixin Duan. Targan: Generating target data with class labels for unsupervised domain adaptation. *Knowledge-Based Systems*, 172: 123–129, 2019. 2

[19] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018. 1

[20] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 6, 7

[21] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1094–1103, 2021. 2

[22] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 4, 5

[23] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman. Hfirst: A temporal approach to object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10): 2028–2040, 2015. 1

[24] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2239–2247, 2019. 5

[25] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1, 2, 5, 6, 7

[26] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1964–1980, 2021. 4

[27] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 2, 5, 6, 7

[28] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020. 5

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[30] Haoxiang Wang, Bo Li, and Han Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *International Conference on Machine Learning*, pages 22784–22801. PMLR, 2022. 2

[31] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021. 1, 2

[32] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 1, 2

[33] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):9004–9021, 2023. 5

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 5

[35] Yu-Chu Yu and Hsuan-Tien Lin. Semi-supervised domain adaptation with source label adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24100–24109, 2023. 4

[36] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 2, 4

[37] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3

[38] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2