This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

HALSIE: Hybrid Approach to Learning Segmentation by Simultaneously Exploiting Image and Event Modalities

Shristi Das Biswas, Adarsh Kosta, Chamika Liyanagedera, Marco Apolinario and Kaushik Roy
Purdue University, West Lafayette, Indiana, USA
{sdasbisw, akosta, cliyanag, mapolina, kaushik}@purdue.edu

Abstract

Event cameras detect changes in per-pixel intensity to generate asynchronous 'event streams'. They offer great potential for accurate semantic map retrieval in real-time autonomous systems owing to their much higher temporal resolution and high dynamic range (HDR) compared to conventional cameras. However, existing implementations for event-based segmentation suffer from sub-optimal performance since these temporally dense events only measure the varying component of a visual signal, limiting their ability to encode dense spatial context compared to frames. To address this issue, we propose a hybrid end-to-end learning framework HALSIE, utilizing three key concepts to reduce inference cost by up to 20× versus prior art while retaining similar performance: First, a simple and efficient cross-domain learning scheme to extract complementary spatio-temporal embeddings from both frames and events. Second, a specially designed dual-encoder scheme with Spiking Neural Network (SNN) and Artificial Neural Network (ANN) branches to minimize latency while retaining cross-domain feature aggregation. Third, a multi-scale cue mixer to model rich representations of the fused embeddings. These qualities of HALSIE allow for a very lightweight architecture achieving state-of-the-art segmentation performance on DDD-17, MVSEC, and DSEC-Semantic datasets with up to 33× higher parameter efficiency and favorable inference cost (17.9mJ per cycle). Our ablation study also brings new insights into effective design choices that can prove beneficial for research across other vision tasks.

1. Introduction

We often see house-flies seamlessly navigating through cluttered spaces, supporting such complex movements with just a few million neurons [1]. Yet, modern autonomous systems with significantly higher compute capability and near unlimited resources [2] still fail to replicate the comprehensive real-time scene understanding achieved by these tiny biological systems with very low power budgets. In this work, we discuss these shortcomings for the task of semantic segmentation which is an important building block in the perception pipeline for autonomous navigation systems.

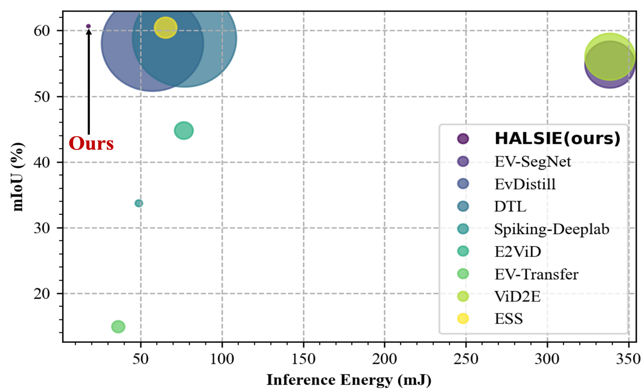
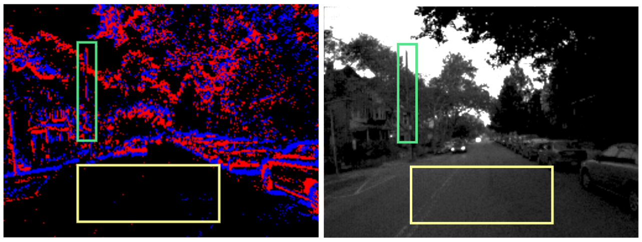


Figure 1. **Segmentation performance vs inference energy** of our HALSIE model on DDD-17 dataset evaluated on 45nm CMOS process. The circle areas are proportional to the model size.

Present-day segmentation falls back on conventional frame cameras that share little with the biological eye in their scene capturing mechanism [3, 4]. They sample intensity frames synchronously at constant and large time intervals. In addition, they fail to capture information in challenging scenarios with HDR and motion blur, leading to the loss of essential scene details. In safety-critical applications like automotive, this may come at the cost of fatalities. Increasing the sampling rate would enable them to capture high speed motion but leads to redundant background information retrieval along with elevated energy consumption.

To circumvent these issues, researchers have explored event cameras [5–7] as an alternate sensing modality. Event sensors asynchronously measure changes in per-pixel intensity to output sparse data streams at high temporal resolution (10µs vs 3ms), higher dynamic range (140dB vs 60dB) and significantly lower energy (10mW vs 3W) compared to frame cameras. These properties make event cameras enticing for high-speed segmentation. However, the event stream only contains information about pixels experiencing intensity changes, rendering the retrieval of dense contex-



(a) (b)

Figure 2. **Complementary feature extraction.** (a) Events capture temporally rich features only where intensity changes occur, (b) lacking the dense spatial information contained in standard frames.

tual information from scenes challenging [8]. As seen in Fig. 2, intensity information is absent at pixels where events are not recorded leading to scenes with incomplete contextual information (highlighted in a yellow box). On the other hand, the electric pole (green box) is concealed in the frame output due to high dynamic range. Both cameras are, therefore, complementary. This complementarity motivates the development of a multi-modal hybrid approach to improve semantic performance in challenging conditions, leveraging advantages of both the frame- and event-domains.

Traditional learning methods based on Artificial Neural Networks (ANNs), originally designed for frame-based images [9–13], are inept at directly handling discrete and asynchronous event streams [14]. Several works propose to convert asynchronous events to standard frames for downstream processing based on handcrafted features [15–17]. Unfortunately, unlike frame-camera images, such accumulated event frames are much sparser and lack dense semantic information. To further aggravate the issue, ANNs discard the temporal ordering of these inputs by representing them as parallel channels and perform sub-optimal stateless computations on the sequential data stream. Recent work has shown that such stateless ANNs show inferior performance on event-based semantic segmentation, only improving when they leverage temporal recurrence [18, 19]. Nevertheless, these methods require large model sizes and incur high inference costs (> 50 mJ per cycle) since they do not rely on efficient processing to exploit the sparse information retrieval aspect of event cameras (Fig. 1).

Recently, Spiking Neural Networks (SNN) [20] have emerged as promising candidates for directly handling event streams. They perform asynchronous sparse event-driven compute [21] and offer an implicit recurrence through their internal states to preserve input temporal information [22] in a stateful manner. Such traits make SNNs ideal for handling sequential inputs efficiently, enabling considerable energy savings on neuromorphic hardware [23–25].

With this in mind, we fundamentally revisit the design of end-to-end-learning frameworks for semantic segmentation using events and frames. In particular, we identify several key components that enable us to alleviate expensive inference cost and high parameter requirements while retaining semantic performance at low latencies. The main contributions of our work are as follows: We propose HALSIE, a simple and efficient composable architecture with (1) novel hybrid spatio-temporal feature extraction scheme to effectively combine events and frames allowing better information retrieval from a scene (compared to these modalities working independently), (2) and multi-scale cue mixing to enable powerful cross-domain feature integration between the aggregated temporal features and current spatial feature. Our method is lightweight, inference-efficient and still offers state-of-the-art performance for semantic segmentation. (3) We evaluate HALSIE on real-world DDD-17 [26], MVSEC [27], and DSEC-Semantic datasets [18] and demonstrate up to 9% improvement over the best performance reported so far with significant energy savings. In addition, we also provide insights into the various components of our method that contribute to these results.

2. Related Works

With event cameras showcasing great potential for semantic segmentation, there have been several efforts in recent years exploring this emerging research direction.

Recent works explore using **stateless ANNs** with dense event representations, discarding temporal correlation across the event window by representing them as channels. Initial work to adapt events for semantic segmentation in [28] used an Xception-type network [29] to achieve robust performance in corner case scenes suffering from overexposure. The authors published the first event-based segmentation dataset with semantic labels [26] generated on synchronised grayscale frames from DAVIS346B [5,6]. Researchers in [30] showed improvements over [28] by training on an augmented dataset comprising real and synthetic events converted from videos. However, they require video datasets, very few of which exist for the task. Improving upon their approach, authors in [31] attempt to exploit knowledge learned from high-quality labeled image datasets such as Cityscapes [32] for unpaired event data, and report better performance. However, their knowledge distillation process leads to much higher compute costs. In contrast, [33] relies on event-to-image transfer but fails to consider any network blocks to address the inherent temporal correlation in events. Instead, [34] reports a method for 'image-to-event transfer' that splits the embedding space into motion-specific features shared by events and images using adversarial learning. However, their method depends on hallucination of motion from images to generate fake events and is prone to mode collapse [35].

A second research direction uses **temporal recurrence in dense neural networks** to achieve better semantic performance with events. Efforts have been made in event-to-image reconstruction methods, with authors in [19] us-

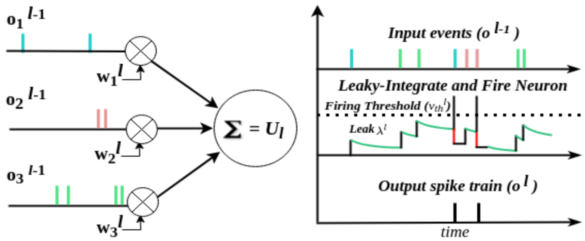


Figure 3. **Dynamics of a Leaky-Integrate and Fire (LIF) neuron.** The firing threshold v_{th}^l and the leak factor λ^l are dynamically updated during training to attain best possible performance.

ing a recurrent network to convert events to images and processing them with standard ANNs. However, not only do they end up discarding temporal ordering in the inputs in their effort to synthesize motion-invariant images but also incur significant parametric overheads and inference-latency during image generation from events. Follow-up work by [18] attempts to transfer the task from labeled image datasets to unpaired events by aligning similar reconstructed event embeddings generated using the recurrent encoder from [19] with standard image embeddings. While mostly suffering from the same bottlenecks as [19], another significant downside of using dense recurrence comes with synchronous networks like RNNs that struggle with efficient processing when sampling rates of the input are highly variable and asynchronous [36,37]. Typical sub-optimal solutions involve padding, copying, or sampling rate conversions of event frames, which comes at the cost of frequency reductions and temporal misalignments of input data [38].

A recent third line of work employed **SNNs** to propagate information sparsely within the network. Authors in [39] were the first to explore a fully-spiking approach, but reported sub-optimal performance compared to state-of-the-art dense networks due to the very sparse event-driven *binary* compute across their deep ANN-inspired architectures [40, 41]. We build on this line of work to instead propose a hybrid method unifying the advantages of SNNs and ANNs in a compact way. Our method differs from existing work in a few key points: (1) we maximize accumulated *analog* temporal context from event streams while still leveraging sparse implicit recurrence. To achieve this, we build a shallow SNN-based Temporal Feature Extractor (TFE) module that does not rely on any of the earlier additional parametric and learning overheads. (2) Enhancing the complementary dense spatial embeddings from the ANN-based Spatial Feature Extraction (SFE) module with aggregated temporal features from the SNN-based TFE module. (3) We combine the rich spatio-temporal embeddings using a Multi-scale Mixer (MMix) to build a fast, lightweight and highly performant end-to-end learning framework deployable on hybrid ANN-SNN neuromorphic chips such as [42].

3. Method

Our segmentation approach is designed to process a stream of events sequentially as they arrive. In every timestep, our network takes a new event bin as input and relies on the accumulated neuronal state from previous inputs to produce temporal feature maps. After mixing the current spatial features and aggregated temporal features, the fused embeddings are used as input to the decoder. Fig. 4 shows an overview of the HALSIE architecture.

3.1. Input Processing

We characterize event data in the Address Event Representation (AER) format as a tuple $e_i = (x_i, y_i, t_i, p_i)$ that occurs at pixel (x_i, y_i) at time t_i, and with polarity $p_i \in 0, 1$. In this work, we employ a simple yet effective pre-processing method to map events into a grid-like presentation. Our preprocessing step starts with discretizing an aggregated event volume as follows: For a set of N input events $\{(x_i, y_i, t_i, p_i)\}_{i\in[1,N]}$ between two consecutive grayscale images and a set of B event bins to be created within this event volume, we generate discretized event bins using bilinear sampling kernels $k_b(a)$ [43]:

$$t_i^* = (B-1)(t_i - t_1)/(t_N - t_1) \quad (1)$$

$$V(x,y,t) = \sum_i p_i k_b(x - x_i) k_b(y - y_i) k_b(t - t_i^*) \quad (2)$$

$$k_b = max[0, 1 - |a|] \quad (3)$$

In words, we generate B temporal bin tensors, each having ON/OFF polarity channels containing the number of positive or negative events within each bin, passed sequentially as timesteps through the TFE module to preserve the inter-bin temporal correlation. The intuition behind a multi-channel representation is to allow the network to learn pixel ownership for moving objects (pixels on the same object will move in the same direction, and generate spatially close iso-polarity events) while capturing short-term temporal correlation over timesteps. We use GT-labels on the latter grayscale image in the event window as done in prior art [18, 28, 31, 33]. Other more complex representations exist [44–46], but their thorough evaluation is not our focus.

3.2. Mixing Spatial and Temporal Features

The HALSIE architecture features a deep hybrid encoder-decoder network for end-to-end learning. To efficiently extract rich spatiotemporal (ST) features from the complementary sensors, we design a dual-path encoder comprising an ANN-based *Spatial Feature Extractor (SFE)* for frames and an SNN-based *Temporal Feature Extractor (TFE)* for events. We enable incorporating a higher effective receptive field by using a 3×3 convolution with overlapping kernels in both encoders that at the same time spatially decimates the feature map from the previous encoder step (number of channels is scaled up by a factor of 2).

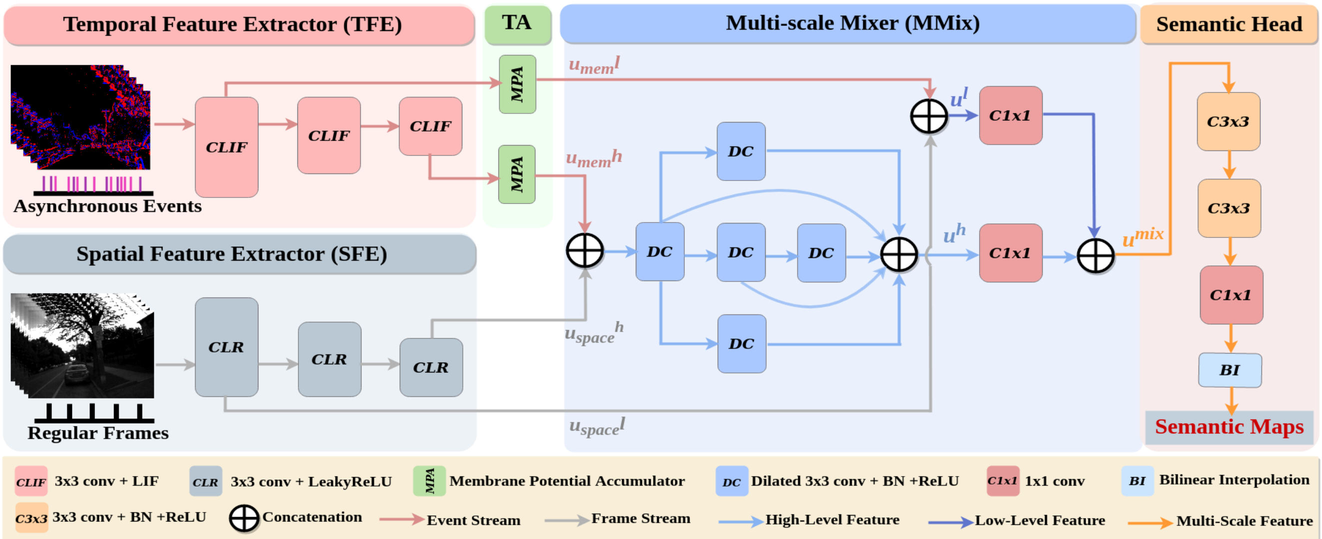


Figure 4. **Overview** of the proposed HALSIE framework. Given a set of inputs, the TFE and SFE blocks extract rich temporal and spatial embeddings. Temporal Accumulor (TA) and Multi-Scale Mixer (MMix) modules combine analog u_{mem} and u_{space} features by specially designed feature-mixing scheme. Finally, the MMix block interfaces with the segmentation head to generate dense semantic maps.

Temporal Feature Extraction We opt for spike-based temporal feature aggregation with Leaky-Integrate-and-Fire (LIF) [47, 48] neurons at each layer in the TFE. LIFs are amongst most widely used bio-inspired spiking neuron models because of their inherent ability to 'remember' and 'recall' past information, skipping computation on neuromorphic hardware if they haven't received any input event (event-driven computation). We characterize the internal dynamics of our LIF neuron model as follows:

$$u^l_{mem}[t] = w^l o^{l-1}[t] + \lambda^l u^l_{mem}[t-1] - v^l_{th} o^l[t-1] \quad (4)$$

$$o[t] = \mathcal{H}(u_{mem}[t] - v_{th}) \quad (5)$$

where $\mathcal{H}$ represents the Heaviside step function [49]. At timestep t, weighted output spikes from the previous neuron $l-1$ are accumulated in the membrane potential $u^l_{mem}[t]$ of the neuron l creating a 'short-term memory'. At the same time, $u^l_{mem}[t]$ of the neuron l decays by a leak factor λ^l to represent 'forgetting'. Once the accumulated membrane potential exceeds the firing threshold v^l_{th}, the neuron generates a binary spike output (o^l). $u^l_{mem}[t]$ is reset using the 'soft reset' strategy [50, 51] after all the B temporal bins are processed. We regard this sparse potential accumulation, decay, and resting process as an efficient temporal memory, motivating us to investigate the SNN layers for temporal feature extraction. We examine this aspect in Sec. 4.5.1. The LIF neural dynamics is shown in Fig. 3.

Binary spike trains are used to transmit temporal features in traditional SNNs. Differently from prior work in Spiking-Deeplab [39], we use the analog membrane potential to transmit neural activations from the TFE instead of the generated output spikes. This enables the aggregated temporal embeddings to have higher expressibility than simple binary spikes while still demanding drastically lower compute compared to traditional RNNs. In other words, applying accumulated analog u_{mem} maps from the Temporal Accumulator (TA) module instead of binary spike maps allows our model to avoid the performance degradation experienced in [39] while still retaining all the benefits of spike-based processing. Dynamic threshold schemes observed in different biological systems play an essential role in their robustness to various external conditions [52, 53]. Specifically inspired by [54], we also learn the v_{th} and λ for each layer dynamically to find the optimal hyperparameters unlike methods in [39]. We find that allowing the network to learn on the go how 'important' a new input is, or how much of the past experiences influence its current state by adjusting its threshold and leak enables our dynamical TFE to attain substantial boosts in dense classification tasks.

Spatial Feature Extraction The ANN-based SFE branch adopts channel-wise dependencies to extract rich texture cues, which we call spatial potential maps u_{space}, from synchronized grayscale images of the DAVIS sensor temporally closest to the event bins. If multiple such images are available over a temporal window, they can be fed as separate channels at the input. Each SFE block comprises a conv. layer with overlapping kernels, batch-norm (BN) [56] and a LeakyReLU activation [57]. We further study the SFE branch as part of our architecture variation in the ablation studies in Sec. 4.5.1.

3.3. Multi-scale Mixer

In the subsequent step, resulting temporal and spatial embeddings, u_{mem} and u_{space}, are mixed using a multi-

Table 1. **Comparison on test set of DDD-17, measured by accuracy and mIoU.** Best results in **bold** and second best <u>underlined</u>. Parameter efficiency and inference energy cost is computed on standard 45nm CMOS process [55] (See suppl. material for details).

Methods	Accuracy [%]	mIoU [%]	Network	Params($\times 10^6$)	#FLOPs$_{\text{ANN}}$($\times 10^9$)	#FLOPs$_{\text{SNN}}$($\times 10^9$)	E$_{\text{Total}}$(mJ)
EV-SegNet [28]	89.76	54.81	ANN	29.09	73.62	-	338.65
EvDistill [31]	-	58.02	ANN	59.34	12.45	-	57.27
DTL [33]	-	58.80	ANN	60.48	16.74	-	77.01
Spiking-Deeplab [39]	-	33.70	SNN	<u>4.14</u>	-	<u>54.34</u>	48.91
E2ViD [19]	83.24	44.77	ANN	10.71	16.65	-	76.59
EV-Transfer [34]	47.37	14.91	ANN	7.37	<u>7.88</u>	-	<u>36.25</u>
ViD2E [30]	90.19	56.01	ANN	29.09	73.62	-	338.65
ESS [18] (E)	<u>91.08</u>	**61.37**	ANN	12.91	14.22	-	65.41
ESS [18] (E+F)	90.37	60.43	ANN	12.91	14.22	-	65.41
Ours (HALSIE)	**92.50**	<u>60.66</u>	Hybrid	**1.82**	**3.84**	**0.267**	**17.89**

scale integrator. Spatial feature maps from the SFE at the highest-level and lowest-level scales (u^h_{space} and u^l_{space}) are combined with the corresponding analog accumulated membrane potential maps (u^h_{mem} and u^l_{mem}) from the Temporal Accumulator (TA) module to obtain high- and low-level mixed potential maps. High-level mixed maps (i.e, u^h) from the last encoder layer pass through a *Multi-scale Mixer (MMix)* block with each branch of the cell employing decoupled rate 3×3 dilated convolutions [13]. By enabling different sampling rates $r_h \times r_w$ for each dilated conv. cell, we capture object scales with different aspect ratios, and create a more diverse feature space with each branch of the block building local multi-scale contextual information through parallel or cascaded representations. More details regarding the decoupled sampling rates for the model can be found in the supplementary. As a next step, the low-level u^l maps and multi-scale high-level u^h maps are concatenated after channel-mixing (1×1 or pointwise convolutions) corresponding to global, dilated mixed features u^{mix}.

3.4. Semantic Head

For the semantic head, we adopt a lightweight task decoder consisting of 2 [(3×3 conv) $\rightarrow$ (BN) $\rightarrow$ (ReLU)] blocks followed by a (1×1 conv) and upsampling layer to predict the segmentation mask. We examine the u^{mix} feature maps as a toolkit to visualise and interpret why such a simple decoder design works well for our method and discuss results in Sec. 4.5.1.

4. Experiments and Evaluation

4.1. Setup

Our models are trained 100 epochs with the ADAM optimizer [58] using a MultiStepLR learning rate schedule to scale the learning rate by 0.7 every 10 epochs. We use a weighted pixel-wise cross entropy loss to examine each pixel individually. Unlike standard backpropagation in ANNs, gradient computation in SNNs is not straightforward since LIF neurons have a spiking mechanism that generates non-differentiable threshold functions. We enable learning with surrogate gradients to approximate the gradient of the Heaviside step function during backpropagation [59, 60] in our TFE branch and use the inverse tangent surrogate gradient function with width $\gamma = 100$ (to allow sufficient gradient flow) since it is computationally inexpensive. To construct event representations, we discretize the event window between consecutive frames into $B = 10$ temporal bins bins and pass them along with the synchronized grayscale frames to the TFE and SFE branches respectively. To estimate energy costs for a single inference, we use the number of floating point operations (FLOPs) performed by the network per inference cycle. For details on computing approximate inference energy, refer to the supplementary material.

4.2. Evaluation on DDD-17 Dataset

Dataset and Training Details: We use the publicly available driving scene dataset DDD-17 [26], containing 40 different driving sequences of synchronized grayscale images and event streams. Due to the low resolution of the DAVIS camera, several classes are fused to create labels for six merged classes. From the provided sequences, this work uses a training set of $15,584$ frames and a test set of $3,584$ frames. Maintaining parity with prior art [28,30,31], we use constant integration time event bins with $T = 50ms$. Our data augmentation includes random flips and rotations on inputs and cropping them to 192×192 size images. We train batch sizes of 32 on an initial learning rate of $8e-4$. and report accuracy and mean intersection over union (mIoU) on our semantic maps to evaluate performance.

Results: Quantitative results are reported in Table 1 and visualized in Fig. 5. We compare our approach with existing works such as [28, 30, 31, 33, 34] that do not leverage temporal correlation between events, and find that our hybrid framework leverages the complementary events and frames with efficient spatio-temporal learning to consistently outperform them and achieve new state-of-the-

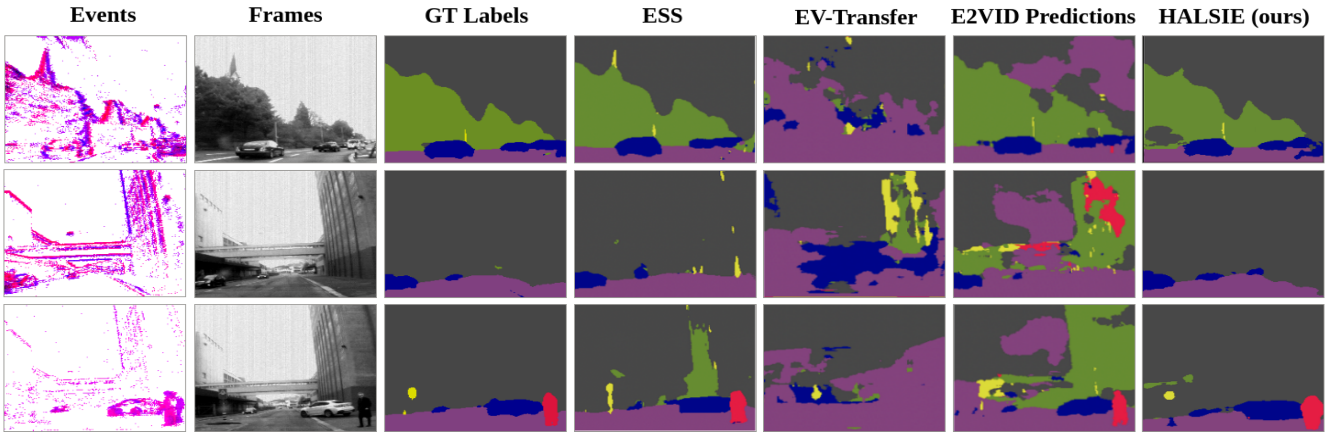


Figure 5. **Samples on DDD-17.** Compared to prior art, our highly lightweight hybrid framework generates more reliable predictions with upto 72.7% lower inference energy (gray: background; green: vegetation; blue: vehicle; violet: street; yellow: object; red: person)

art performance of 60.66% mIoU and 92.50% accuracy. ESS [18] which uses a recurrent event encoder to hallucinate motion-invariant event embeddings in the event only (E) and event+frame (E+F) settings claims comparable results, albeit at the cost of 85.9% lower parameter efficiency. This can be mainly attributed to the inherent self-recurrence in SNNs which are more suitable to denoise and extract sparse cues compared to traditional RNNs which are not designed for sparse, asynchronous or irregular data (also pointed out in [61]). We will verify this later in Sec. 4.5.1 and the supplementary. Our lightweight model is the smallest in our comparison with up to a staggering 73% lower inference cost than existing approaches. Still, our efficient multi-scale cross-domain feature mixing allows our method to generate the most reliable predictions without compromising on qualitative performance, visualized in Fig. 1.

We also observe that the fully-spiking approach in [39] using reconfigured LIF neurons to include atrous convolutions reports very low performance, in line with our intuition regarding the need for multi-modal hybrid networks to leverage complementary information from both sensors. In the first row in Fig. 5, our semantic maps were unable to predict the tower peak above the vegetation and classifies it as vegetation itself. However, since the tower peak is not a crucial element in the scene compared to the presence of a nearby traffic pole, an incoming vehicle, or a person, we posit that the error is not critical. Our method makes more reliable predictions in the scenes in the second and third rows compared to its counterparts.

4.3. Evaluation on MVSEC Dataset

Dataset and Training details: As events in the DDD-17 dataset are very sparse and noisy, we present experimental results on the MVSEC dataset [27] comprising of various driving scenes for stereo estimation. Due to the poor quality of frames in the 'outdoor day1' sequence, we mainly use the

Table 2. **Comparison on MVSEC dataset**. Existing approaches [31] and [33] fail to report their accuracy metrics.

Methods	Accuracy [%]	mIoU [%]	Params($\times 10^6$)	$E_{\text{Total}}(mJ)$
EvDistill [31]	-	55.09	<u>59.34</u>	<u>101.84</u>
DTL [33]	-	<u>60.82</u>	60.48	136.89
Ours (HALSIE)	**92.13**	**66.31**	**1.82**	**31.39**

Table 3. **Results on DSEC-Semantic.** Our method shows comparable performance to ESS [18] with a staggering 74% lower inference cost, making it a prime candidate for edge-deployment.

Methods	Accuracy [%]	mIoU [%]	Params($\times 10^6$)	$E_{\text{Total}}(mJ)$
EV-Transfer [34]	60.50	23.20	<u>7.37</u>	<u>197.48</u>
E2ViD [19]	76.67	40.70	10.71	416.99
EV-SegNet [28]	88.61	51.76	29.09	1863.83
ESS [18] (E+F)	**89.37**	**53.29**	12.91	356.32
Ours (HALSIE)	<u>89.01</u>	<u>52.43</u>	**1.82**	**94.41**

'outdoor day2' sequence and divide data into training and testing sets [31, 33]. We also remove redundant sequences such as vehicles stopping at traffic lights, etc. and train with batch sizes of 32 and an initial learning rate of $8e-4$.

Results: We summarize the results in Table 2. HALSIE outperforms the two existing methods [31] and [33] by around 20.4% and 9.1% respectively, while using 33× fewer parameters and up to a significant 77% lower inference energy. Effectiveness of our method can also be verified from Fig. 6 where our segmentation results are able to detect very fine details such as a poles or traffic signs (8^{th} column), where our contemporaries fail. The results validate that our highly lightweight framework is able to efficiently leverage spatiotemporal context from both modalities to predict dense semantic maps with fine predictions. In several examples in Fig. 6, we found that our method predicts objects which were not present in the GT labels but were clearly visible in the images, causing misleading reductions in our detection score. We also use HDR scenes to

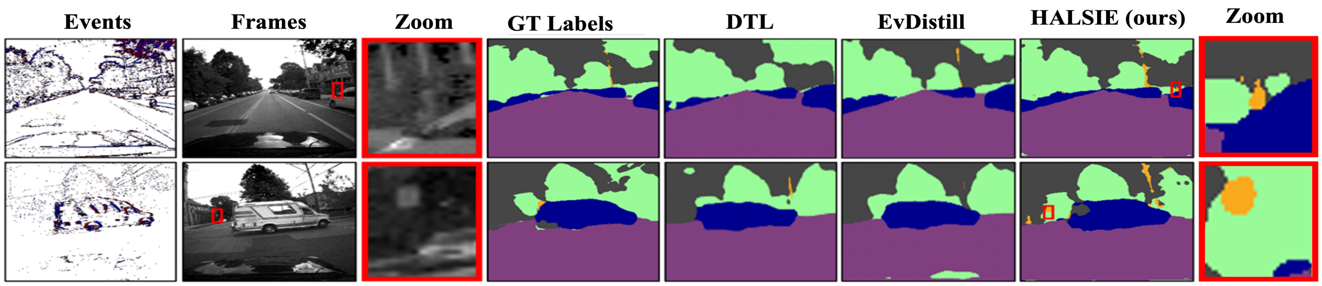


Figure 6. **Results on MVSEC.** Small objects are sometimes missed in the low-quality semantic labels (zoomed-in patch in the 3^{rd} column). This leads to a lower detection score on MVSEC even though our predictions provide more fine-grained detections (see 8^{th} column). (gray: background; green: vegetation; blue: vehicle; violet: street; yellow: object)

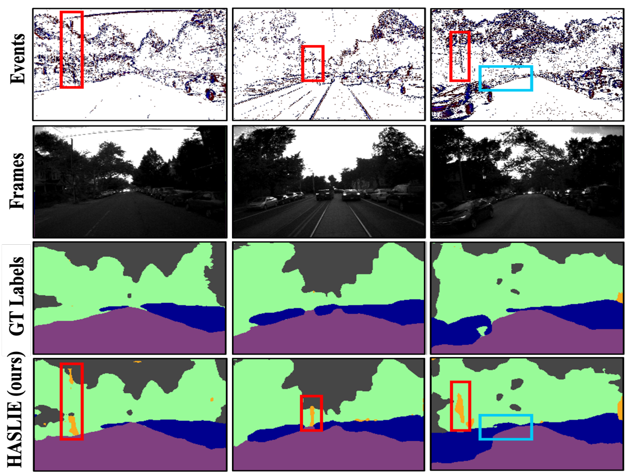


Figure 7. **Results on challenging scenes.** Qualitative samples of HDR scenes in the MVSEC dataset. Best viewed in color.

show how our method provides reliable performance when grayscale frames are ill-exposed. HALSIE is able to efficiently extract information from events and shows promising performance in such challenging conditions where fine details are almost invisible in the grayscale frames (highlighted in colored boxes). See Fig. 7 for qualitative results.

4.4. Evaluation on DSEC-semantic Dataset

Dataset and Training Details: We further evaluate our method on the recently released DSEC-semantic [18] dataset containing 4017 training and 1395 testing samples with 11 semantic classes [62]. The dataset was collected in a variety of urban and rural environments using automotive-grade standard cameras and high-resolution event cameras. Similar to [62], we generate $B = 10$ event bins with a constant event density of $100K$ events/bin to be passed sequentially to the TFE block and associate event bins with labels using the provided semantic timestamps. The SFE module is fed with images from the left frame-camera corresponding to the same semantic timestamp as the events. We train for 100 epochs with an initial learning rate of $5e-4$.

Results: The performance of our method evaluated on the test set is reported in Table 3. We compare our approach with [19, 34] that do not leverage temporal recurrence and find that HALSIE significantly improves segmentation results, surpassing existing methods with around 29% increase in mIoU while using a significant 77.4% lower inference cost. ESS [18] in the events+frames (E+F) setting claims comparable results, while using a much larger network and exorbitant inference costs ($\sim$ 73.5% higher). Finally, our model is the smallest amongst existing literature by a large margin and still achieves 1.3% higher mIoU than EV-Segnet [28] while using 16× fewer parameters, making it a top-bidder for energy-efficient edge-applications. Refer to the suppl. for qualitative samples from our method.

Table 4. **Architectural and input modality variations.** We denote the single encoder setting as 'SE' and dual encoder setting as 'DE'. All models are approximately iso-parameterized.

Methods	Training Data	Accuracy [%]	mIoU [%]
(A) SE: only SFE	frames	88.15	56.38
(B) SE: only SFE	events	83.42	46.01
(C) SE: only TFE	events	83.59	46.09
(D) DE: both SFE	events + frames	90.16	58.28
(E) DE: both TFE	events + frames	86.01	54.43
(F) w/o MSFI	events + frames	90.31	58.83
(G) SNN → LSTM	events + frames	90.25	58.67
(H) (ours) HALSIE	events + frames	**92.50**	**60.66**

4.5. Ablation studies

4.5.1 Mixing Spatial and Temporal Cues

We conduct the following experiments on DDD-17 to demonstrate effects of spatial and temporal cues on the task: *(A)* and *(B)* Removing the TFE branch but keeping the SFE branch; *(C)* Removing SFE, but keeping TFE; *(D)* Dual encoder setting with SFE branches for both frames and events; *(E)* Same setting but with TFE branches instead; *(F)* Removing the MSFI module; *(G)* Replacing the SNNs in SFE with an LSTM [63] with one conv. layer per cell. Comparing the original HALSIE *(H)* to *(A)*, *(B)*, *(C)*, *(D)* or *(E)*, we

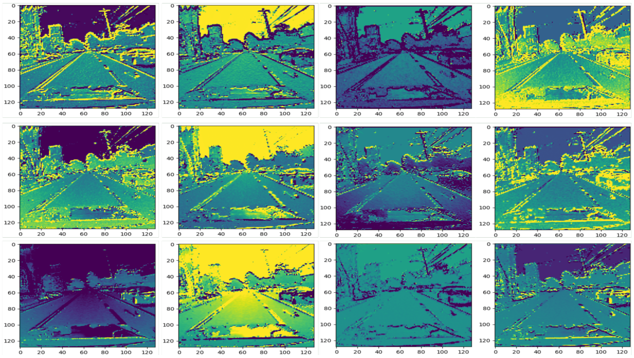


Figure 8. **Mixed feature maps on MVSEC.** Four channels of the u^{mix} maps generated from processing Top row: Events in TFE and Frames in SFE. Middle row: Both inputs using SFE. Bottom row: Both inputs using TFE. Best viewed with zoom in.

Table 5. **Event representation strategy.** Constant event density (CED) bins lead to best results on the DSEC-Semantic dataset.

Event	CED			CIT		
Representation	10k	100k	1000K	10ms	50ms	100ms
Accuracy [%]	88.78	89.01	88.24	87.76	88.02	87.23
mIoU [%]	51.58	52.43	51.01	50.14	50.67	49.89

witness a significant performance degradation, validating that effectively extracting and mixing temporal and spatial cues is essential for boosting performance. Note that setting *(A)* draws comparison of our method with pure frame-based segmentation. Unsurprisingly, when trained with both the modalities, we can detect small objects like traffic poles and people. In several examples, we find that our method segments objects that were not present in the labels, but were clearly visible in the images. For visualization, refer to the supplementary material.

We use the u^{mix} feature maps as a toolkit to visualise and interpret how our simple decoder benefits from the rich mixed feature representation. We study Fig. 8 and find that mixed feature maps generated from processing events with TFE and frames with SFE leads to the sharpest and cleanest features (row 1) even with such simple encoders. In contrast, when using SFE for both inputs (row 2), the feature maps appear less discriminative due to the non-sparse processing of events which does not contribute to sharp edge-extraction. Note that these features are however generated at the highest inference cost amongst the three variants due to dense processing in both encoder branches. While applying TFE to both inputs (row 3) offers the most energy-efficient sparse processing paradigm, the feature maps appear relatively noiseless but suffer from loss of information. As such, our simple decoder design does not work as well on the only SFE or only TFE encoder approaches since it is unable to take advantage of the powerful representation induced by complementary SNN-ANN processing.

We also notice performance degradation in the *(F)* setting. This results reflects the effectiveness of the proposed MSFI block. Compared to *(G)*, we still achieve better performance. This is to some degree surprising because LSTMs can also extract temporal cues. Notably, the spiking mechanism of SNNs acts not only as temporal memory but also as a natural noise filter, which is beneficial to robust predictions. We further examine the denoising aspect of our TFE module in the supplementary material.

4.5.2 Event representation and event density

Thorough evaluation of event representations is not our focus and hence we only study the influence of simple event representations which may not leverage the full potential of event data [64] on our method's performance. Efficient low-level encoding of event data is still an open research problem that we have not addressed in this work. Ablation results in Table 5 and visualisations in the supplementary on the DSEC-Semantic dataset suggest that maintaining a moderately dense bin with constant event density (CED) shows better semantic performance compared to high density bins with trailing artifact events from fast moving objects, or low density bins with minimal contribution to the segmentation performance. We also find that having CED bins consistently helps the network learn the end-task better than with constant integration time (CIT) bins.

5. Conclusion

We introduce HALSIE, a lightweight yet powerful hybrid end-to-end framework for semantic segmentation that is capable of effectively mixing temporal and spatial cues encoded in events and frames. The proposed network relies on several novel modules. We devise an SNN-based temporal feature extractor and an ANN-based spatial feature extractor, which efficiently exploits statistical cues of spatial and temporal information for robust predictions. We also introduce a novel multi-scale mixer for compactly combining embeddings from the two domains. Effectiveness of our design choices is evidenced by the strong performance of our method in detecting finer details on DDD-17, MVSEC and DSEC-Semantic benchmarks while offering sizeable benefits in terms of inference cost and parameter efficiency. The resulting design is deployable for resource-constrained edge applications, and paves the way for low-energy semantic segmentation with event cameras without compromising on performance. Nonetheless, we hope that this work also inspires novel designs in future hybrid systems.

6. Acknowledgement

This work was supported in part by the Center for Brain-inspired Computing (C-BRIC), a DARPA sponsored JUMP center, the Semiconductor Research Corporation (SRC), the National Science Foundation, the DoD Vannevar Bush Fellowship, and IARPA MicroE4AI.

References

[1] A. Borst, J. Haag, and D. F. Reiff, "Fly motion vision," *Annual review of neuroscience*, vol. 33, pp. 49–70, 2010. 1

[2] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017. 1

[3] M. Srinivasan, S. Zhang, M. Lehrer, and T. Collett, "Honeybee navigation en route to the goal: visual flight control and odometry," *Journal of Experimental Biology*, vol. 199, no. 1, pp. 237–244, 1996. 1

[4] E. Baird, N. Boeddeker, M. R. Ibbotson, and M. V. Srinivasan, "A universal strategy for visually guided landing," *Proceedings of the National Academy of Sciences*, vol. 110, no. 46, pp. 18686–18691, 2013. 1

[5] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128$\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008. 1, 2

[6] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240$\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014. 1, 2

[7] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch, "Activity-driven, event-based vision sensors," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 2426–2429, IEEE, 2010. 1

[8] C. Lee, A. K. Kosta, A. Z. Zhu, K. Chaney, K. Daniilidis, and K. Roy, "Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks," in *European Conference on Computer Vision*, pp. 366–382, Springer, 2020. 2

[9] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020. 2

[10] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019. 2

[11] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context for semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2375–2398, 2021. 2

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017. 2

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018. 2, 5

[14] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020. 2

[15] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "Hots: a hierarchy of event-based time-surfaces for pattern recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1346–1359, 2016. 2

[16] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. 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*, pp. 5419–5427, 2018. 2

[17] H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization," 2017. 2

[18] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza, "Ess: Learning event-based semantic segmentation from still images," *arXiv preprint arXiv:2203.10016*, 2022. 2, 3, 5, 6, 7

[19] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 1964–1980, 2019. 2, 3, 5, 6, 7

[20] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, 2019. 2

[21] S. Wang, T. H. Cheng, and M. H. Lim, "A hierarchical taxonomic survey of spiking neural networks," *Memetic Computing*, vol. 14, no. 3, pp. 335–354, 2022. 2

[22] W. Ponghiran and K. Roy, "Spiking neural networks with improved inherent recurrence dynamics for sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 8001–8008, 2022. 2

[23] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018. 2

[24] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 34, no. 10, pp. 1537–1557, 2015. 2

[25] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014. 2

[26] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd17: End-to-end davis driving dataset," *arXiv preprint arXiv:1711.01458*, 2017. 2, 5

[27] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera

dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018. 2, 6

[28] I. Alonso and A. C. Murillo, "Ev-segnet: Semantic segmentation for event-based cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019. 2, 3, 5, 6, 7

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017. 2

[30] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3586–3595, 2020. 2, 5

[31] L. Wang, Y. Chae, S.-H. Yoon, T.-K. Kim, and K.-J. 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*, pp. 608–619, 2021. 2, 3, 5, 6

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016. 2

[33] L. Wang, Y. Chae, and K.-J. 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*, pp. 2135–2145, 2021. 2, 3, 5, 6

[34] N. Messikommer, D. Gehrig, M. Gehrig, and D. Scaramuzza, "Bridging the gap between events and frames through unsupervised domain adaptation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3515–3522, 2022. 2, 5, 6, 7

[35] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in gans," in *2020 international joint conference on neural networks (ijcnn)*, pp. 1–10, IEEE, 2020. 2

[36] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019. 3

[37] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased lstm: Accelerating recurrent network training for long or event-based sequences," *Advances in neural information processing systems*, vol. 29, 2016. 3

[38] S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*. Association for Computing Machinery and Morgan & Claypool, 2018. 3

[39] Y. Kim, J. Chough, and P. Panda, "Beyond classification: directly training spiking neural networks for semantic segmentation," *Neuromorphic Computing and Engineering*, 2022. 3, 4, 5, 6

[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 3

[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014. 3

[42] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, *et al.*, "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019. 3

[43] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015. 3

[44] L. Wang, Y.-S. Ho, K.-J. Yoon, *et al.*, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10081–10090, 2019. 3

[45] R. Baldwin, R. Liu, M. M. Almatrafi, V. K. Asari, and K. Hirakawa, "Time-ordered recent event (tore) volumes for event cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[46] S. Barchid, J. Mennesson, and C. Djéraba, "Bina-rep event frames: A simple and effective representation for event-based cameras," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3998–4002, IEEE, 2022. 3

[47] L. F. Abbott, "Lapicque's introduction of the integrate-and-fire model neuron (1907)," *Brain research bulletin*, vol. 50, no. 5-6, pp. 303–304, 1999. 4

[48] P. Dayan, L. F. Abbott, *et al.*, "Theoretical neuroscience (vol. 806)," 2001. 4

[49] E. W. Weisstein, "Heaviside step function," *https://mathworld. wolfram. com/*, 2002. 4

[50] E. Ledinauskas, J. Ruseckas, A. Juršėnas, and G. Buračas, "Training deep spiking neural networks," *arXiv preprint arXiv:2006.04436*, 2020. 4

[51] B. Han, G. Srinivasan, and K. Roy, "Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13558–13567, 2020. 4

[52] B. Fontaine, J. L. Peña, and R. Brette, "Spike-threshold adaptation predicted by membrane potential dynamics in vivo," *PLoS computational biology*, vol. 10, no. 4, p. e1003560, 2014. 4

[53] K. Pozo and Y. Goda, "Unraveling mechanisms of homeostatic synaptic plasticity," *Neuron*, vol. 66, no. 3, pp. 337–351, 2010. 4

[54] N. Rathi and K. Roy, "Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 4

[55] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 10–14, 2014. 5

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015. 4

[57] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015. 4

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015. 5

[59] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019. 5

[60] J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training deep spiking neural networks using backpropagation," *Frontiers in neuroscience*, vol. 10, p. 508, 2016. 5

[61] L. Deng, Y. Wu, X. Hu, L. Liang, Y. Ding, G. Li, G. Zhao, P. Li, and Y. Xie, "Rethinking the performance comparison between snns and anns," *Neural networks*, vol. 121, pp. 294–307, 2020. 6

[62] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021. 7

[63] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015. 7

[64] J. Li, J. Li, L. Zhu, X. Xiang, T. Huang, and Y. Tian, "Asynchronous spatio-temporal memory network for continuous event-based object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2975–2987, 2022. 8