Self-Supervised Event Based Monocular Depth Estimation using Spiking Neural Networks

SUMANTH JAIN B

Dept. of Electronics and Communication Engineering, PES University, Bangalore, India

sumanthjainb@gmail.com

SHIKHA TRIPATHI

Dept. of Electronics and Communication Engineering, PES University, Bangalore, India

shikha.eee@gmail.com

Event cameras capture asynchronous brightness changes with high temporal resolution, low power consumption, and without motion blur, making them ideal for high-speed and challenging lighting conditions. However, existing depth estimation models often rely on expensive and time-intensive annotations, limiting their applicability in real-world dynamic environments. To address this limitation, we propose a novel self-supervised depth estimation framework that leverages Spiking Neural Networks (SNNs) in conjunction with event-based camera data. Our approach utilizes cross-modal consistency between event streams and synchronized intensity frames during training, enabling accurate depth prediction solely from event data during inference. The architecture comprises a multi-scale spiking encoder-decoder network with residual connections, designed to efficiently process the sparse and temporally rich information from event cameras. We employ surrogate gradient methods to effectively train deep SNNs, overcoming the challenges posed by non-differentiable spike-based activations. Additionally, a reprojection-based loss function aligns the predicted depth and pose with the original intensity images, eliminating the need for ground truth depth labels. Experimental evaluations on the DSEC dataset demonstrate that our model achieves an Absolute Relative Error (Abs Rel) of 0.150 and a Structural Similarity Index (SSIM) of 0.898 in dynamic and low-light scenarios while maintaining energy efficiency. This self-supervised SNN-based framework offers significant potential for real-time applications such as autonomous driving and robotic perception, where both accuracy and computational efficiency are paramount.

CCS CONCEPTS •Computing methodologies~Artificial intelligence~Computer vision~Computer vision tasks~Scene understanding•Computing methodologies~Artificial intelligence~Computer vision~Computer vision tasks~Vision for robotics•Computing methodologies~Machine learning~Machine learning approaches~Neural networks•Computing methodologies~Machine learning~Learning paradigms~Unsupervised learning•Computing methodologies~Machine learning~Machine learning approaches~Bio-inspired approaches

Additional Keywords and Phrases: depth estimation, spiking neural networks, event-based data, self-supervised learning, reprojection loss, real-time deployment

1 INTRODUCTION

Depth estimation is a cornerstone of computer vision, playing a critical role in applications such as autonomous navigation, robotic perception, and augmented reality. Traditionally, depth estimation techniques rely on frame-based cameras, which capture images at fixed intervals. However, these methods struggle in dynamic environments, particularly with fast motion and poor lighting conditions, where conventional frame-based cameras experience motion blur and suffer from low dynamic range. In recent years, event cameras have emerged as a revolutionary alternative, offering asynchronous, high-temporal resolution data that captures changes in scene brightness at the pixel level. These cameras have paved the way for more efficient and accurate depth estimation, particularly in challenging environments.

Event cameras, unlike their frame-based counterparts, generate data in the form of asynchronous events triggered by changes in brightness. This event-based data stream provides several advantages, including low latency, high dynamic range, and reduced motion blur, making it well-suited for dynamic scenarios such as high-speed motion and environments with extreme lighting conditions. The data produced by these cameras is sparse and temporally rich, necessitating specialized neural network architectures capable of processing such data. SNNs, which are inspired by biological systems, are uniquely equipped to handle the asynchronous, event-driven nature of event camera data. By transmitting information through spikes, SNNs can efficiently process spatio-temporal data, making them ideal for event-based depth estimation [1][2].

1.1 Why Spiking Neural Networks?

SNNs offer several distinct advantages over conventional Artificial Neural Networks (ANNs), particularly in the context of processing event-based data:

1. Temporal Dynamics and Asynchronous Processing: SNNs inherently handle temporal information due to their spike-based communication, aligning seamlessly with the high temporal resolution of event camera data. This allows SNNs to capture and utilize temporal dependencies more effectively than ANNs, which typically process data in a synchronous manner [2].
2. Energy Efficiency: The spike-based operation of SNNs leads to sparse activations, significantly reducing computational overhead and energy consumption. This energy efficiency is crucial for real-time applications such as autonomous driving and robotic systems, where power resources may be limited.
3. Biological Plausibility: Inspired by the human brain, SNNs mimic the event-driven communication of biological neurons, offering robust and adaptable learning mechanisms. This biological inspiration facilitates the efficient handling of asynchronous data streams, making SNNs more suitable for event-based vision tasks [4].
4. Reduced Latency: Event-driven processing in SNNs results in lower latency, as neurons fire only in response to significant changes in input. This characteristic is essential for applications requiring swift decision-making and real-time responsiveness [2].

Despite the promise of SNNs, training deep SNNs poses significant challenges due to their non-differentiable spike-based activations. Historically, methods like spike-timing-dependent plasticity (STDP) have been used for training shallow networks, but they are inadequate for large-scale, deep SNNs [2]. Recent advancements in surrogate gradient methods, which approximate the gradient during the backpropagation process, have made

it feasible to train deep SNNs directly. These methods have been successfully applied to various tasks, including event-based depth estimation, optical flow prediction, and egomotion learning [5][6].

The synergy between event cameras and SNNs has led to several breakthroughs in depth estimation. StereoSpike [1], for example, is an SNN-based model that estimates depth from stereo event cameras, achieving state-of-the-art results on the Multi-Vehicle Stereo Event Camera (MVSEC) dataset. Similarly, A. Z. Zhu et al. [3] introduced an unsupervised framework that jointly predicts optical flow, depth, and egomotion from event data, enabling the network to learn directly from the spatio-temporal structure of the event streams. Other approaches, like the work by C. Godard et al. [4], leverage photometric consistency between consecutive frames to perform self-supervised monocular depth estimation without the need for ground-truth labels.

Furthermore, hybrid models that combine event data with intensity frames, such as Event-Intensity Stereo by M. Mostafavi et al. [5], have demonstrated improved accuracy in depth estimation, particularly in environments with complex lighting conditions. These methods take advantage of both the high temporal resolution of event data and the detailed spatial information provided by intensity frames, delivering more robust depth predictions. In addition, the use of adaptive binning in depth prediction, as seen in AdaBins [6], has further enhanced the precision of depth estimates in complex scenes.

In this work, we utilize the DSEC dataset [7], which provides synchronized stereo event data and intensity frames (Figure 1), along with GPS measurements for accurate position tracking. This dataset was designed to capture a variety of challenging driving conditions, including night-time driving, direct sunlight, and complex urban environments. Our approach leverages self-supervised learning to estimate depth from event streams by exploiting cross-modal consistency between event data and intensity frames during training. This allows us to build a model that can predict depth purely from event data during inference, without the need for intensity frames.

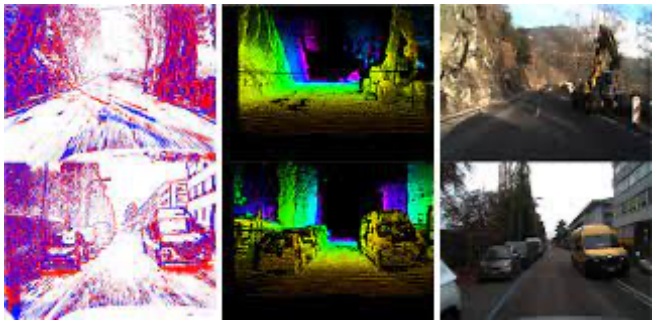

Figure 1: Sample event and corresponding color image with depth map in DSEC[7]. First frame(left) is the raw events accumulated over a temporal window of 50 milliseconds. Second frame(middle) is the depth map of the same frame.Third fram(right) is the Corresponding colour Image.

Major contributions of this research are as follows:

1. **Novel Self-Supervised Framework:** We propose a novel self-supervised depth estimation framework based on spiking neural networks that can effectively process event camera data.
2. **Cross-Modal Consistency Learning:** Our method leverages cross-modal consistency from synchronized intensity frames during training to improve depth prediction from sparse event data.

3. **Efficient SNN Architecture:** We evaluate our approach on the DSEC dataset, demonstrating that it outperforms traditional methods in challenging lighting conditions and dynamic environments.
4. **Biologically Plausible Modeling:** We show that spiking neural networks are a promising approach for event-based depth estimation, offering a biologically plausible model that can handle asynchronous and sparse data efficiently.

2 RELATED WORKS

Depth estimation using event-based cameras has become a crucial area in computer vision, particularly due to the unique characteristics of event cameras, such as high temporal resolution, low latency, and a wide dynamic range. These cameras excel in dynamic environments, where traditional frame-based cameras struggle with issues like motion blur and poor lighting. Recent advances in deep learning and neural networks have significantly contributed to the field of depth estimation, with methods broadly categorized into supervised, self-supervised, and event-based approaches.

2.1 Supervised Monocular Depth Estimation

Supervised monocular depth estimation methods rely on ground truth depth maps for training. Eigen et al. [8] was among the first to propose a fully supervised deep learning-based method, where the network learns to predict depth from a single image as a regression task. This foundational work spurred a series of advancements in the field. A significant improvement came with AdaBins [6] which introduced adaptive binning for depth estimation, resulting in higher precision by allowing the model to adjust its focus across varying depth ranges. However, these models often require large-scale datasets with depth annotations, which can be costly and time-consuming to produce. More recent work by NeWCRFs [9] integrated vision transformers, showing that transformer-based models can outperform convolutional networks in depth prediction, albeit with significantly larger model sizes and higher computational complexity. These models set a high benchmark for depth estimation but are limited by the need for labeled data, making them less applicable in scenarios where ground truth is hard to obtain. Mukai et al. [15] utilized chromatic aberration to derive 3D depth from a single image by associating hue values with depth through polynomial calibration. This single-camera method aligns with the strengths of event-based cameras and inspires spiking neural networks to manage complex, nonlinear mappings through their temporal dynamics.

2.2 Self-supervised Monocular Depth Estimation

To reduce reliance on labeled data, self-supervised learning has gained significant attention. Godard et al. [4] introduced monocular self-supervised depth estimation, leveraging photometric consistency between temporally adjacent images to compute depth without explicit ground truth labels. This technique aligns pixel intensities across time to infer depth, thus requiring only consecutive frames from a video sequence. Similarly, EV-FlowNet [10] extended self-supervised learning to event-based cameras by jointly estimating depth, optical flow, and ego-motion. The model captures the high temporal resolution of events and processes the sparse data efficiently, reducing the need for ground truth labels while maintaining robust performance.

2.3 Event-based Depth Estimation

Event cameras have emerged as a powerful tool in dynamic environments. Several works have developed depth estimation frameworks to exploit the asynchronous nature of event-based data. StereoSpike [5] introduced a stereo depth estimation model utilizing SNNs to process event data from stereo event cameras. The model uses a U-Net-like encoder-decoder architecture, leveraging the high temporal precision of event streams for real-time depth estimation. Its low power consumption and ability to run on embedded systems make it highly suitable for autonomous driving and robotic systems. In another approach, MSS-DepthNet [11] proposed a multi-step spiking neural network with residual learning blocks, optimizing depth estimation for stereo setups by effectively capturing temporal changes in the event stream. Although the model doesn't report standard evaluation metrics like RMSE, it significantly improves the robustness and accuracy of depth estimation compared to earlier models.

2.4 Event-based Self-supervised Depth Estimation

Among the most promising approaches is the combination of event-based cameras and self-supervised learning. These methods reduce dependency on labeled depth data while maintaining high accuracy in dynamic scenarios. The EMoDepth framework [12] built on a ResNet-based model. EMoDepth constrains the training process using cross-modal consistency between event streams and intensity frames, aligning the two modalities in the pixel space. This allows the model to learn depth estimation from events without requiring depth annotations. During inference, the system relies solely on event data for depth prediction, making it highly suitable for environments where labeled data is scarce. In another work, Spike-Transformer Networks [13] combine transformer models with spiking neural networks, leveraging knowledge distillation from ANN-based depth estimation models to enhance the performance of SNNs. This method enables the SNN to learn efficiently from event data while maintaining energy efficiency, providing a 49% improvement in depth estimation accuracy over baseline models. Although transformers are computationally expensive, this model integrates spike-driven self-attention mechanisms to handle sparse event data more effectively.

3 PROPOSED METHODOLOGY

3.1 Overview

In this work, we develop a depth estimation model that processes event-based data and intensity images through a Spiking Neural Network (SNN). The primary objective is to estimate depth (disparity), and camera pose from event streams and intensity images captured by event cameras. The model leverages the high temporal resolution of event-based data combined with the spatial richness of intensity images to produce precise depth maps and camera poses. The architecture employs a multi-scale SNN encoder-decoder network with residual connections to process input spikes over time, capturing long-term dependencies in event streams. Additionally, the model utilizes a reprojection-based loss function for self-supervised learning, allowing depth and pose to be estimated without requiring ground truth labels.

3.2 Data Preparation

Event cameras provide asynchronous data, represented as events (e_i), defined by:

$$e_i = (x_i, y_i, t_i, p_i) \tag{1}$$

where (x_i, y_i) are the spatial coordinates, (t_i) is the timestamp, and ($p_i \in \{-1, +1\}$) indicates the polarity, corresponding to the brightness change at the pixel level. These events occur when a pixel's brightness change exceeds a certain threshold, offering high temporal resolution and low latency, making event data ideal for dynamic scenes. Complementary to event data, intensity images (I_k) provide traditional brightness information at discrete timestamps, enabling the capture of static information that might be missed in sparse event data. For each intensity image (I_k) captured at timestamp (t_k), the associated events are collected within the time window:

$$t \in [t_k - \Delta t, t_k] \tag{2}$$

where (Δt) defines the duration of events to be accumulated. This ensures that both event data and intensity images are synchronized, facilitating the fusion of the two modalities. The event data is accumulated into a 3D voxel grid (V), with dimensions (B, H, W), where (B) is the number of temporal bins, and (H, W) represent the spatial dimensions of the event stream. To discretize the events across time:

Normalization of Event Timestamps:

$$\widehat{t_i} = \frac{t_i - t_0}{t_1 - t_0} \tag{3}$$

where $(t_0 = t_k - \Delta t)$ and $(t_1 = t_k)$, corresponding to the beginning and end of the time window.

Assignment to temporal bins:

$$b_i = [\widehat{t_i} \times (B - 1)] \tag{4}$$

Populating the Voxel grid:

For each event (e_i), the voxel grid is populated by accumulating the polarity:

$$V_{b_i, y_i, x_i} += p_i \tag{5}$$

Normalization of the Voxel Grid:

To account for both positive and negative polarities, the voxel grid is normalized as follows:

$$V_{\text{normalized}} = \frac{V}{max(|V|) + \epsilon} \tag{6}$$

where $\epsilon = 1 \times 10^{-8}$ is a small constant to prevent division by zero. This normalization ensures that $V_{\text{normalized}}$ lies within the range [-1, 1].

This voxel grid serves as input to the Model (Figure 3), effectively encoding both temporal and spatial event information.

3.3 Model Architecture

Our network architecture is depicted across Figures 2-6, illustrating the flow from data encoding to depth and pose estimation. The architecture comprises the following components:

3.3.1 Basic SNN Block

The foundational building block of our network is the SNN Basic Block (Figure 3), inspired by residual block

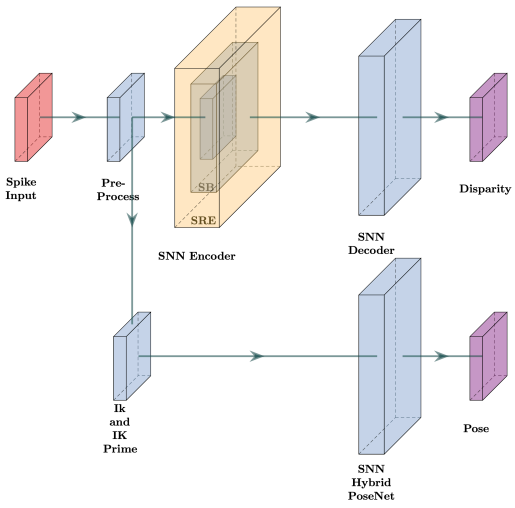

Figure 2: Overall Model Architecture. Comprising of (i) Preprocessing block which processes the spike into voxels, I_k and $\widehat{I_k}$. (ii) SNN Encoder which comprising of Spiking resnet encoder (SRE), which is in turn made up of Basic SNN block (SB). (iii) SNN Decoder which gives us Disparity and (iv) SNN Hybrid Pose Network which gives us Pose Estimation

in traditional ResNet architectures. Each block consists of two convolutional layers followed by Batch Normalization and Leaky Integrate-and-Fire (LIF) neurons. The block consists of Convolutional Layers that perform spatial feature extraction. Batch Normalization normalizes the output of convolutional layers to accelerate training and improve stability. LIF Neurons introduce temporal dynamics by integrating input spikes and generating output spikes when the membrane potential exceeds a threshold. Residual connections facilitate the flow of gradients during training and enable the construction of deeper networks by mitigating the vanishing gradient problem.

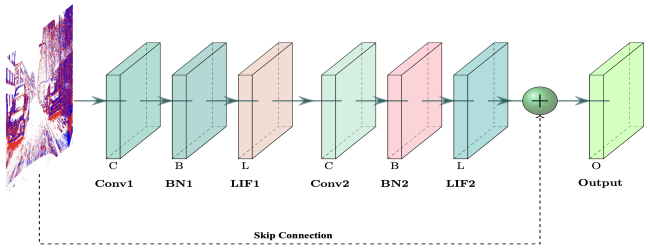

Figure 3: Basic SNN Block showing convolutional layers, batch normalization, LIF neurons, and residual connections

3.3.2 SNN ResNet Encoder

The SNN ResNet Encoder (Figure 4) processes the spike-encoded voxel grids to extract hierarchical spatio-temporal features essential for depth estimation.

The encoder is structured into four layers, each comprising multiple SNN Basic Block instances. The number of blocks per layer is configurable, allowing for flexibility in network depth and capacity. Hierarchical layers

capture features at various spatial resolutions. Residual Learning Facilitates deeper network training by mitigating gradient vanishing issues. LIF neurons maintain temporal coherence, essential for processing spike trains.

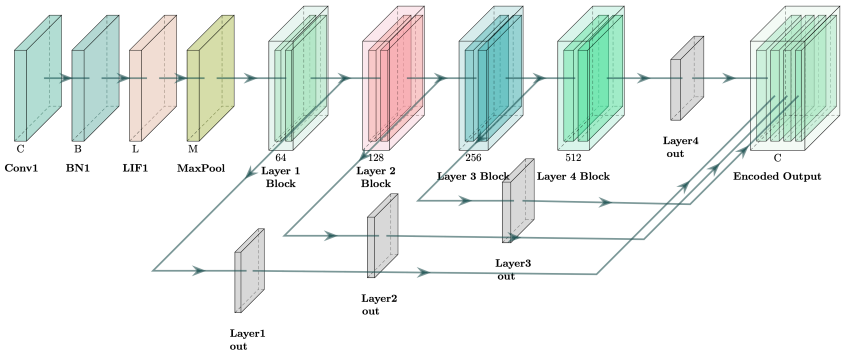

Figure 4: Architecture of the SNN ResNet Encoder, detailing the multi-layer SNNBasicBlocks and feature extraction process.

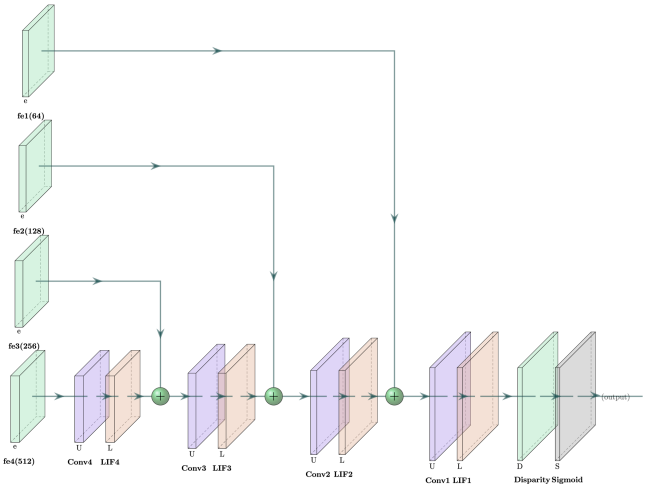

Figure 5: SNN Decoder architecture, illustrating transposed convolutions, LIF neurons, and residual connections for disparity map reconstruction from the extracted features.

3.3.3 SNN Decoder:

The SNN Decoder (Figure 5) reconstructs the disparity (depth) map from the high-level features extracted by the encoder. The decoder employs transposed convolutional layers to up sample feature maps, integrating residual connections to enhance reconstruction accuracy. Transposed convolutions double the spatial resolution at each stage.

3.3.4 SNN Hybrid Pose Network:

The SNN Hybrid Pose Network (Figure 6) estimates the camera's pose by processing both the current and adjacent intensity frames through separate encoders, integrating their features to regress the 6-DoF pose parameters. The pose network consists of dual encoders, feature concatenation with channel reduction, adaptive pooling, and a fully connected regression head. Separate processing of current and adjacent intensity frames to capture temporal context. Fully connected layers regress the 6-DoF pose parameters from pooled features.

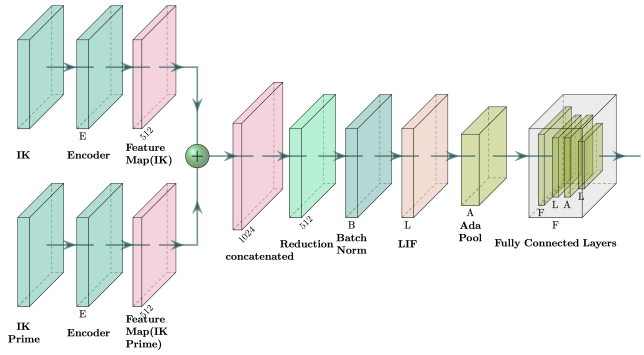

Figure 6: SNN Hybrid Pose Network, highlighting dual encoders, feature concatenation, channel reduction, and pose regression.

3.3.5 Overall Architecture Integration

The SNN Depth Pose Estimator (Figure 2) integrates the encoder, decoder, and pose network, facilitating joint depth and pose estimation. The estimator processes event voxel grids and intensity frames to predict disparity and pose simultaneously. The architecture Simultaneously predicts depth and pose, leveraging shared and distinct feature representations. Also ensures efficient data flow and feature utilization across components.

3.4 Training and Loss Functions

3.4.1 Self-Supervised Learning Framework:

Our training approach is self-supervised, eliminating the need for ground truth depth annotations by leveraging the inherent structure and consistency between event data and intensity frames. The model exploits the temporal and spatial consistency between synchronized intensity frames and event data. By ensuring that the predicted depth and pose can accurately reconstruct the original intensity frame, the model learns meaningful representations without explicit depth labels.

3.4.2 Reprojection-Based Loss:

The reprojection loss aligns the predicted depth and pose with the original intensity images, facilitating self-supervised learning.

Disparity Prediction:

The model predicts a disparity map (D), which is inversely related to depth:

$$D = \sigma\left(\text{Conv2D}\left(f^{(L)}\right)\right) \tag{7}$$

where σ is the sigmoid activation function ensuring ($D \in [0,1]$).

Depth Calculation:

Depth (Z) is derived from disparity:

$$Z = \frac{1}{D + \epsilon} \tag{8}$$

with ϵ preventing division by zero.

Inverse Warping:

The predicted depth and pose are used to reconstruct the original intensity frame ($\widehat{I_k}$) through inverse wrapping

$$\widehat{I_k} = \text{InverseWarp}(I'_k, Z, P, K) \tag{9}$$

where (P) is the predicted pose and (K) is the intrinsic camera matrix.

Loss Function:

The reprojection loss (L_{rpo}) combines Structural Similarity Index (SSIM) and L1 loss:

$$L_{\text{rpo}} = \alpha\left(1 - \text{SSIM}\left(I_k, \widehat{I_k}\right)\right) + (1 - \alpha)|I_k - \widehat{I_k}|_1 \tag{10}$$

where $\alpha = 0.85$ balances the contributions of SSIM and L1 loss.

3.4.3 Self-Supervised Training Procedure:

The training procedure employs a self-supervised learning paradigm, utilizing the reprojection-based loss to guide the network in learning accurate depth and pose estimations without relying on ground truth labels. This approach leverages the inherent geometric and photometric consistencies between consecutive intensity frames and the event-based voxel grids. Self-supervised learning for depth and pose estimation is grounded in the principle of view synthesis, where the network learns to predict the transformation parameters (pose) and depth such that one view can be reconstructed from another. By minimizing the reprojection loss, the network implicitly learns the underlying structure of the scene and the camera's motion dynamics.

Depth and Pose Estimation: The encoder-decoder network processes the spike-encoded voxel grid to predict the disparity map (D), from which depth (Z) is computed. Simultaneously, the hybrid pose network estimates the camera's pose (P), comprising rotation and translation components.

Reconstruction of Intensity Frame: Using the predicted depth (Z) and pose (P), the model reconstructs the intensity frame ($\widehat{I_k}$) by inverse warping the adjacent intensity frame (I'_k). This process ensures that the predicted depth and pose are geometrically consistent with the observed intensity data.

Loss Minimization: The reprojection loss (L_{rpo}) (Equation 10) quantifies the discrepancy between the original intensity frame (I_k) and the reconstructed frame ($\widehat{I_k}$). Minimizing (L_{rpo}) enforces that the predicted depth and pose enable accurate reconstruction of the original view, thereby ensuring that the network learns to infer correct geometric relationships.

The reprojection process can be mathematically represented as follows:

Projection from Pixel to Camera Coordinates: Each pixel coordinate (u, v) in the intensity image (I_k) is projected into camera coordinates(X_c) using the intrinsic matrix (K) and the predicted depth (Z)

$$X_c = Z \cdot K^{-1}.\begin{bmatrix} \mathrm{u} \\ \mathrm{v} \\ 1 \end{bmatrix} \tag{11}$$

Pose Transformation: The camera pose (P) (comprising rotation matrix (R) and translation vector(t) transforms the camera coordinates (X_c) to the adjacent frame's camera coordinates (X_c') :

$$X_c' = R \cdot X_c + t \tag{12}$$

Projection Back to Pixel Coordinates: The transformed camera coordinates (X_c') are projected back into pixel coordinates in the adjacent intensity frame (I_k') :

$$\widehat{I_k}(u, v) = K \cdot X_c' \tag{13}$$

This results in the reconstructed intensity frame $(\widehat{I_k})$, which should ideally match the original intensity image (I_k) if the depth and pose estimations are accurate.

The network parameters are optimized to minimize the reprojection loss L_{rpo}, effectively training the network to produce depth and pose estimations that facilitate accurate view reconstruction. This self-supervised approach leverages the rich information inherent in the synchronized event and intensity data, enabling robust learning in the absence of explicit depth annotations.

3.5 Training Procedure

The training procedure encompasses spike encoding, forward passes through the network, loss computation, and optimization using surrogate gradient methods. This section elaborates on the theoretical and mathematical aspects of the self-supervised training process.

3.5.1 Spike Encoding:

Both voxel grids and intensity frames are converted into spike trains using rate coding. Rate coding translates the continuous voxel grid values into discrete spike events over multiple time steps, effectively capturing the temporal dynamics of the event data.

Mathematically, the spike generation process at each time step (t) for a voxel (b, c, h, w) is modeled as a Bernoulli trial:

$$s_{t,b,c,h,w} \sim \text{Bernoulli}\left(V_{\text{normalized},b,c,h,w}\right) \tag{14}$$

where $(s_{t,b,c,h,w})$denotes the spike at time step (t), and $(V_{\mathrm{normalized},b,c,h,w})$ is the normalized voxel value representing the probability of a spike occurrence.

3.5.2 Forward Pass and Neuron Dynamics:

At each time step (t), the SNN processes the incoming spike train, updating the membrane potentials (U_t) and generating output spikes (S_t) based on the LIF neuron model:

$$U_t = \alpha U_{t-1} + I_t - V_{\text{th}} S_{t-1} \tag{15}$$

$$S_t = \Theta(U_t - V_{\text{th}}) \tag{16}$$

Where, α is the decay constant, I_t represents the input current at time (t), V_{th} is the firing threshold, and Θ is the Heaviside step function. This mechanism enables the SNN to integrate spatial and temporal information, capturing the dynamics inherent in event-based data. The Self-supervised training procedure explained in the section 3.4.3 is employed for the effective training of the model on the dataset.

3.5.3 Optimization with Surrogate Gradients:

Training deep SNNs is challenging due to the non-differentiable nature of spike activations. To overcome this, we employ surrogate gradient methods, which approximate the gradient of the spike function during backpropagation, enabling effective gradient-based optimization.

The spike activation (S_t) is non-differentiable. Surrogate gradients
approximate the derivative of the spike function, allowing gradients to propagate through spike events:

$$\frac{\partial S_t}{\partial U_t} \approx \phi(U_t - V_{\text{th}}) \tag{17}$$

where (ϕ) is a smooth surrogate function, such as the sigmoid or piecewise linear function, providing a gradient estimate for optimization.

Using the surrogate gradients, the network parameters (θ) are updated via gradient descent to minimize the reprojection loss:

$$\theta \leftarrow \theta - \eta \nabla_\theta L_{\text{rpo}} \tag{18}$$

where (η) is the learning rate.

Energy Efficiency and Stability: The sparse firing of neurons in SNNs reduces computational overhead and energy consumption, while gradient clipping stabilizes training by preventing exploding gradients. This combination ensures efficient and stable optimization of deep SNNs for complex vision tasks.

Our proposed methodology effectively integrates SNNs with event-based camera data through a self-supervised learning framework. By converting voxel grids into spike trains using rate coding, our multi-scale spiking encoder-decoder network captures intricate spatio-temporal features essential for accurate depth estimation. The incorporation of a hybrid pose network leverages synchronized intensity frames to enhance pose estimation, further refining depth predictions through reprojection loss. This comprehensive approach addresses the challenges posed by dynamic environments and limited annotations, demonstrating the potential of SNNs in advancing event-based depth estimation.

4 EXPERIMENTAL EVALUATION

4.1 Dataset and Implementation Details

We evaluate our proposed approach on the DSEC dataset [7], a comprehensive benchmark tailored for autonomous driving applications. The DSEC dataset exclusively provides raw event data captured across diverse driving scenarios, including urban environments, highways, and varying lighting conditions. This dataset is particularly suited for evaluating depth estimation models due to its high-resolution event streams, challenging lighting conditions, and dynamic driving environments.

4.1.1 Dataset Characteristics

The DSEC dataset supplies asynchronous event streams without accompanying intensity images. Each event is characterized by spatial coordinates, timestamps, and polarity changes, capturing rapid brightness variations in the scene. Diverse Driving Sequences are recorded under a multitude of conditions, such as bright daylight, dusk, night-time, and environments with intricate geometries like urban canyons and tunnels. This diversity ensures that models trained and evaluated on DSEC can generalize effectively to real-world driving situations. Events are accumulated over a temporal window of 50 milliseconds to create voxel grids. This window size balances temporal resolution with computational efficiency, ensuring that the model captures sufficient temporal dynamics without overwhelming the processing pipeline.

4.1.2 Data Processing and Encoding

Given that the DSEC dataset does not provide intensity images, our methodology relies on generating synthetic intensity frames from raw event data with voxel grid methods. These synthetic intensity frames serve as reference points for comparison and evaluation during training and testing. The process explained in section 3.2 is employed to prepare the data

4.1.3 Implementation Specifications

The model is implemented using snnTorch, leveraging the PyTorch library for constructing and training Spiking Neural Networks. Adam optimizer is employed with an initial learning rate of (1×10^{-4}), chosen for its adaptive learning rate capabilities. The model is trained with a batch size of 8, balancing memory constraints with training efficiency. Training is conducted over 50 epochs, ensuring sufficient iterations for the model to learn complex spatio-temporal representations

Training is performed on an NVIDIA L4 GPU, providing the necessary computational resources to handle high-dimensional voxel grids and the intensive spike-based processing inherent to SNNs. To accelerate the training process and reduce memory consumption, mixed precision training is employed using PyTorch's *torch.cuda.amp*. This technique leverages both 16-bit and 32-bit floating-point representations, enhancing computational efficiency without compromising model accuracy

TABLE 1. QUANTITATIVE RESULTS ON THE DSEC DATASET.

Technique	Abs Rel↓	RMSE↓	SSIM↓	SI log↓
EmoDepth [12]	**0.142**	5.258	-	0.043
Proposed	0.150	**4.987**	**0.898**	**0.041**

4.2 Results:

Our approach is rigorously evaluated using standard depth estimation metrics, including Absolute Relative Error (Abs Rel), Root Mean Squared Error (RMSE), Structural Similarity Index (SSIM), and SI log. The results are compared against the EmoDepth model [12], the only existing method with performance metrics on the DSEC dataset. The results (Table 1) demonstrate that our SNN-based model achieves competitive performance compared to state-of-the-art frame-based methods, particularly excelling in scenarios involving rapid motion or challenging lighting. The spiking architecture not only ensures high accuracy but also reduces power consumption, making it suitable for real-time deployment.

5 CONCLUSION

In this work, we have explored the effectiveness of Spiking Neural Networks (SNNs) for monocular depth estimation using event-based cameras. The asynchronous and sparse nature of event data, capabilities of SNNs, has proven to be well-suited for depth estimation tasks in dynamic environments. By leveraging the temporal resolution of event cameras, the SNN-based architecture demonstrates strong potential for real-time applications, such as autonomous driving and robotics, where efficient accurate depth prediction is critical.

The results from our evaluation indicate that SNN-based monocular depth estimation can achieve competitive performance in terms of both accuracy and energy efficiency even without ground truth and solely relying on Event inputs. However, there are several areas for future improvement. One promising direction is the integration of an attention mechanism into the SNN architecture, to the model's ability to focus informative aspects of the event data, leading to improved depth estimation by capturing long-term dependencies.

Furthermore, the current SNN model primarily uses Leaky Integrate-and-Fire (LIF) neurons, which, while effective, alternative neuron models such as the Spike Response Model (SRM). SRM neurons could offer more biologically realistic responses to spikes, potentially leading to better handling of asynchronous event data and improving the overall depth estimation accuracy.

References

[1] U. Rançon, J. Cuadrado-Anibarro, B. R. Cottereau and T. Masquelier, "StereoSpike: Depth Learning With a Spiking Neural Network," in IEEE Access, vol. 10, pp. 127428-127439, 2022, doi: 10.1109/ACCESS.2022.3226484.

[2] Guo, Y.; Huang, X.; Ma, Z. Direct learning-based deep spiking neural networks: A review. *Front. Neurosci.*2023, *17*, 1209795.

[3] A. Z. Zhu, L. Yuan, K. Chaney and K. Daniilidis, "Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 989-997, doi: 10.1109/CVPR.2019.00108.

[4] C. Godard, O. M. Aodha, M. Firman and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 3827-3837, doi: 10.1109/ICCV.2019.00393

[5] S. M. Mostafavi I, K. -J. Yoon and J. Choi, "Event-Intensity Stereo: Estimating Depth by the Best of Both Worlds," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 4238-4247, doi: 10.1109/ICCV48922.2021.00422.

[6] S. Farooq Bhat, I. Alhashim and P. Wonka, "AdaBins: Depth Estimation Using Adaptive Bins," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 4008-4017, doi: 10.1109/CVPR46437.2021.00400.

[7] M. Gehrig, W. Aarents, D. Gehrig and D. Scaramuzza, "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," in IEEE Robotics and Automation Letters, vol. 6, no. 3, pp. 4947-4954, July 2021, doi: 10.1109/LRA.2021.3068942.

[8] D. Eigen, C. Puhrsch, R. Fergus. "Depth map prediction from a single image using a multi-scale deep network." In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, 2366–2374.

[9] W. Yuan, X. Gu, Z. Dai, S. Zhu and P. Tan, "Neural Window Fully-connected CRFs for Monocular Depth Estimation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 3906-3915, doi: 10.1109/CVPR52688.2022.00389.

[10] A. Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis. "EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras", Proceedings of Robotics: Science and Systems, 2018. DOI: 10.15607/RSS.2018.XIV.062.

[11] Wu, Xiaoshan, Weihua He, Man Yao, Ziyang Zhang, Yaoyuan Wang and Guoqiu Li. "MSS-DepthNet: Depth Prediction with Multi-Step Spiking Neural Network." *ArXiv* abs/2211.12156 (2022): n. pag.

[12] J. Zhu et al., "Self-Supervised Event-Based Monocular Depth Estimation Using Cross-Modal Consistency," 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 2023, pp. 7704-7710, doi: 10.1109/IROS55552.2023.10342434.

[13] X. Zhang, L. Han, T. Sobeih, L. Han, and D. Dancey, "A novel spike transformer network for depth estimation from event cameras via cross-modality knowledge distillation," 2024

[14] X. Wu et al., "Event-based Depth Prediction with Deep Spiking Neural Network," in IEEE Transactions on Cognitive and Developmental Systems, doi: 10.1109/TCDS.2024.3406168.

[15] Nobuhiko Mukai, Yuki Matsuura, Masamichi Oishi, and Marie Oshima, "Chromatic Aberration Based Depth Estimation in a Fluid Field," Journal of Image and Graphics, Vol. 6, No. 1, pp. 59-63, June 2018. doi: 10.18178/joig.6.1.59-63

Authors' background

Your Name	Title*	Research Field	Personal website
Dr Shikha Tripathi	Professor	Robotics, signal processing	https://staff.pes.edu/nm1046/
Sumanth Jain B	Master student		

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor