

Hybrid Neuromorphic Video Reconstruction

A TECHNICAL REPORT

submitted by

Abhiroop Pamula	AM.EN.U4EAC22003
Aswin B Ajay	AM.EN.U4EAC22018
Gouri S	AM.EN.U4EAC22026
Viswajith K	AM.EN.U4EAC22066

under the guidance of

Dr. Shyam Diwakar, Dr. Asha Vijayan

submitted as part of

19ECE495/ 19EAC495 PROJECT PHASE I

in

**ELECTRONICS AND COMMUNICATION
ENGINEERING**



**AMRITA SCHOOL OF ENGINEERING
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI (INDIA)**

December – 2025

**AMRITA SCHOOL OF ENGINEERING
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI (INDIA)**



BONAFIDE CERTIFICATE

This is to certify that the report entitled ”**Hybrid Neuromorphic Video Reconstruction**” submitted by Abhiroop Pamula (AM.EN.U4EAC22003), Aswin B Ajay(AM.EN.U4EAC22018), Gouri S(AM.EN.U4EAC22026), Viswajith K(AM.EN.U4EAC22066) as part of the 19ECE495/19EAC495 PROJECT PHASE I is a bonafide record of the work carried out by her under my guidance and supervision at Amrita School of Engineering, Amritapuri.

Signature of Supervisor:

Supervisor name: Dr Shyam Diwakar
Designation: Professor
Department of ECE

Signature of Co-Supervisor:

Name of co-supervisor: Dr Asha Vijayan
Designation: Assistant Professor
Amrita Mind Brain Centre

**AMRITA SCHOOL OF ENGINEERING
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI - 690 542**

DEPARTMENT OF ECE

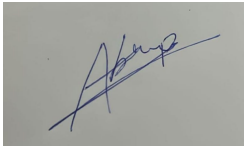
DECLARATION

We, Abhiroop Pamula (AM.EN.U4EAC22003), Aswin B Ajay(AM.EN.U4EAC22018), Gouri S(AM.EN.U4EAC22026), Viswajith K(AM.EN.U4EAC22066), hereby declare that this technical report is entitled "Hybrid Neuromorphic Video Reconstruction" is the record of the original work done by us under the guidance of Dr Shyam Diwakar, Dr Asha Vijayan, Department of ECE, Amrita School of Engineering, Amritapuri. To the best of my knowledge, this work has not formed the basis for the award of any degree/diploma/ associate-ship/fellowship/or similar award to any candidate in any University.

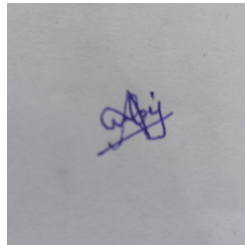
Place: Amritapuri

Signature of the Students

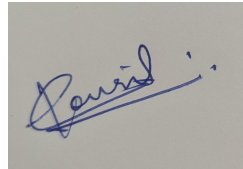
Date: 27-12-2025



Abhiroop Pamula
AM.EN.U4EAC22003



Aswin B Ajay
AM.EN.U4EAC22018



Gouri S
AM.EN.U4EAC22026



Viswajith K
AM.EN.U4EAC22066

Acknowledgement

Firstly, we would like to thank Dr Shyam Diwakar and Dr Asha Vijayan, Electronics and Communication Engineering department, Amrita School of Engineering, Amritapuri, for their invaluable advice, continuous support, and valuable suggestions throughout the course of this project. Their advice and contributions have helped to shape this work in many ways.

We would also like to thank the Department of Electronics and Communication Engineering and Amrita Vishwa Vidyapeetham, Amritapuri, for providing infrastructure and a conducive environment for the successful completion of this work.

We express our heartfelt thanks to all the faculty members who helped us with their knowledge and guidance at various stages of this work. We are thankful to our friends for their constructive discussions, motivation, and cooperation throughout the project work.

Finally, we would like to thank our parents and family for all the support, help, and encouragement that helped us throughout the execution of this project.

Contents

Acknowledgement	4
List of Figures	iv
List of Tables	v
List of Symbols	vi
Abstract	vii
1 Introduction	1
1.1 Background	1
1.2 Event-to-Video Reconstruction Problem	2
1.3 Objectives of the Project	3
1.4 Organization of the Report	4
1.5 Research Gap and Motivation	4
2 Literature Review	5
2.1 Fully End-to-End Event-Based Video Reconstruction Using Spiking Neural Networks	5

2.2	Energy-Efficient Spiking Segmentation for Frame and Event-Based Images . . .	6
2.3	GPU Acceleration for Spiking Neural Network Simulation	7
2.4	Unfolded LSTM-Based Video Reconstruction Architectures	7
2.5	Analog Deep Learning Methods for Event-to-Video Reconstruction	8
3	Proposed Methodology	9
3.1	System Overview	9
3.2	Implementation	10
3.3	Event Representation and Preprocessing	10
3.4	Hybrid Spike–Analog Encoder	11
3.4.1	Spiking Encoder	12
3.4.2	Analog Encoder	13
3.5	Learned Weighted Fusion Module	13
3.6	ConvLSTM-Based Decoder	14
3.7	Loss Functions	15
3.8	Training Strategy	16
4	Results and Discussion	17
4.1	Dataset Description	17
4.2	Quantitative Results	18
4.2.1	Metric Distributions	19
4.3	Qualitative Analysis	19
4.4	Training Dynamics and Resource Usage	20

4.5	Temporal Performance and Spike Activity	22
4.6	Discussion	24
5	Conclusion and Future Work	26
5.1	Conclusion	26
5.2	Future Work	27

List of Figures

3.1	Architecture of the Proposed Hybrid Encoder. The upper route makes use of Spiking Neural Networks (SNN) integrated with LIF neurons to get the temporal dynamics, while the lower route use conventional Analog Convolutional layers for the extraction of spatial features.	12
3.2	Detailed view of the ConvLSTM-based Decoder. Fused features enter a series of ConvLSTM cells that maintain temporal states (h_{t-1}, c_{t-1}) , followed by up-sampling layers and skip connections from the encoder to reconstruct the final intensity frame.	15
4.1	Histograms representing the statistical distribution of reconstruction quality across the test dataset. The mean PSNR is 15.47 dB and mean SSIM is 0.7686.	19
4.2	Qualitative comparison between the Original frame (left) and the Reconstructed frame (right) produced by the proposed hybrid neuromorphic framework. . . .	20
4.3	Training loss convergence and validation PSNR improvement over 25 epochs. .	21
4.4	CPU and GPU Memory usage during training, demonstrating resource stability.	22
4.5	Temporal stability of reconstruction metrics over a continuous video sequence.	23
4.6	Spike rate activity over 600 frames, showing stable SNN pathway performance.	24

List of Tables

4.1	Performance Comparison of Event-to-Video Reconstruction Methods	18
-----	---	----

List of Symbols

λ	Leakage factor of the Leaky Integrate-and-Fire (LIF) neuron
t	Discrete time step
x_t	Input current at time step t
mem_t	Membrane potential of the neuron at time step t
θ	Firing threshold of the spiking neuron
β	Surrogate gradient sharpness parameter
$\sigma(\cdot)$	Sigmoid activation function
spike_t	Output spike at time step t
α	Learnable fusion weight between spike-based and analog features
h_{t-1}	Previous hidden state of ConvLSTM
c_{t-1}	Previous cell state of ConvLSTM
L_{total}	Total loss function
L_1	Mean absolute error loss
L_{gradient}	Gradient consistency loss
L_{contrast}	Contrast enhancement loss
λ_g	Weight for gradient loss
λ_c	Weight for contrast loss
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
FPS	Frames per second
(x, y, t, p)	Event representation: pixel location, time, and polarity

Abstract

Neuromorphic vision sensors, commonly referred to as event cameras, capture pixel-level intensity changes asynchronously with high temporal resolution, low latency, and a wide dynamic range. All these characteristics make event cameras particularly suitable for dynamic environments such as robotics, autonomous navigation, and surveillance systems. However, their sparse and asynchronous output representation poses significant challenges for conventional vision algorithms that are designed to operate on dense, frame-based image data. To address this limitation, this work explores the use of event cameras for video reconstruction.

This paper proposes a hybrid neuromorphic architecture for event-based video reconstruction that combines a spiking neural network (SNN) encoder with a large-capacity recurrent decoder. A learned weighted fusion mechanism is introduced to optimally integrate spike-based and analog feature representations using a trainable fusion parameter. This design enables the framework to exploit the strong temporal contrast sensitivity of spiking neurons while retaining the smooth intensity reconstruction capability of analog neural networks.

The proposed architecture incorporates a hybrid Leaky Integrate-and-Fire neuron-based event encoding scheme, Sobel-based gradient enhancement for edge emphasis, and a ConvLSTM-based decoder to effectively model temporal dependencies across event sequences. Experi-

mental evaluation using a strict cross-domain testing protocol demonstrates that the proposed method achieves a structural similarity index (SSIM) of 0.774 on unseen video sequences while operating in real time at 98 frames per second.

Chapter 1

Introduction

1.1 Background

Current developments in computer vision have resulted in neuromorphic vision sensors, also known as event cameras or Dynamic Vision Sensors (DVS) [1]. Unlike conventional frame-based cameras which record whole image frames at regular time intervals, event cameras are asynchronous and can only record variations in pixel values [1]. Each pixel independently emits whenever the brightness change is beyond a predetermined threshold. Consequently, event cameras offer very fine temporal resolution, low latency and broad dynamic range, and less power consumption [1]. Event cameras are particularly recommended in those applications that require high velocity, adverse light, and real-time requirements, e.g., robotics, self-driving navigation, augmented reality, and surveillance systems [2]. Nevertheless, event cameras produce sparse asynchronous-event streams rather than compressed image frames and thus the usual computer vision is hard to apply directly. frame-based data deep learning algorithms [3]. The event-to-video reconstruction problem consists of reconstructing an image of the object using video at various frames.

In order to use the available vision pipelines with event-based sensors, the use of existing vision pipelines is usually required. convert sparse streams of events to sequences of dense intensity frames. This task is referred to as event-to-video reconstruction [3]. This process aims at recovering temporally. uniform and visually material video frames that give the impression of the look of the original scene. The reconstruction problem is ill-posed in nature as there is no event data that carries the information of it. intensity information (absolute). The same streams of events can be produced with different intensity signals, such that it would be impossible to reconstruct the original scene in a unique way [1]. Apart from that, all the sparse features have an asynchronous nature which contributes to additional difficulties in modelling the spatial structures and temporal dependencies.

1.2 Event-to-Video Reconstruction Problem

Usually, event cameras capture the changes in the pixel intensity, enabling high temporal resolution and sparse data representation. However, such cameras or sensors are expensive. This motivates us to the exploration of exploration of event-inspired representations that can be derived directly from the conventional frame-based videos. In this work, instead of using a physical event-based camera, event-like information is derived from standard video frames by calculating the inter frame pixel intensity differences. A change is noted when the value of the change that happened between the two frames is more than the threshold. And hence we will get a stream of events. While this approach does not replicate true asynchronous sensing, it still preserves the characteristic of event data, encoding changes rather than the absolute intensity, while remaining compatible with event-based sensors available. But re-

constructing the dense intensity frames from such change-based representations remains a challenging and ill-posed problem. Since the representation emphasises temporal differences and suppresses static intensity information, multiple intensity images can correspond to the same change pattern. Furthermore, noise, illumination variations, and motion discontinuities can introduce artifacts such as flickering or loss of spatial detail. Effective construction therefore, requires jointly modelling spatial structure and temporal continuity to recover visually consistent frames. To address these challenges, the proposed framework combines the event-inspired features over short temporal windows and employs a hybrid spike-analog encoding with a convLSTM-based decoder. This design enables stable temporal integration of frame-derived change signals while reconstructing smooth, high-fidelity intensity frames suitable for real-time edge device applications without any event-based sensors of DVS cameras.

1.3 Objectives of the Project

The main objectives of this project are as follows:

- To model the temporal dependencies with the help of the recurrent decoder to enhance the reconstruction quality [9].
- To test the suggested approach with the help of a cross-domain testing protocol [1].
- To show the ability to make inferences in real-time to support latency-sensitive applications [2].

1.4 Organization of the Report

The remainder of this report is organized as follows. Chapter 2 presents a detailed literature review on event-based vision, neuromorphic models, and reconstruction techniques. Chapter 3 describes the proposed hybrid architecture, including the encoder, fusion mechanism, decoder, and loss functions. Chapter 4 discusses the experimental setup, evaluation metrics, and results. Finally, Chapter 5 concludes the report and outlines possible directions for future work.

1.5 Research Gap and Motivation

The literature reveals a clear trade-off between reconstruction quality and computational efficiency. Fully spiking neural networks offer high energy efficiency but struggle to produce smooth intensity reconstructions [18], whereas analog deep learning models generate high-quality outputs at the cost of speed and efficiency .

Hybrid spike–analog approaches that combine the strengths of both paradigms remain relatively underexplored, particularly those that enable adaptive fusion between spike-based and analog representations [4]. Moreover, existing methods rarely achieve real-time inference, cross-domain generalization, and neuromorphic compatibility simultaneously [3].

These limitations motivate the proposed hybrid architecture that integrates spiking encoders with analog recurrent decoders through a learned fusion mechanism.

Chapter 2

Literature Review

This chapter reviews existing research related to event-driven vision, neuromorphic video reconstruction, spiking neural networks, and hybrid deep learning approaches [1]. The purpose of this review is to understand how current methods address the challenges of event-based data and to identify the gaps that motivate the proposed hybrid spike-analog reconstruction framework.

2.1 Fully End-to-End Event-Based Video Reconstruction Using Spiking Neural Networks

The development of neuromorphic vision sensors has led to growing interest in reconstruction techniques that operate directly on event streams [1]. One of the most notable works in this area is Event-Based Video Reconstruction via Potential-Assisted Spiking Neural Network (EVSNN), presented at CVPR 2022 [4]. This study proposes the first fully spiking neural network capable of reconstructing video frames directly from asynchronous event data.

The authors introduce an Adaptive Membrane Potential neuron model, which extends the conventional Leaky Integrate-and-Fire neuron by adjusting firing thresholds based on

spiking activity [4]. This adaptive mechanism improves the network’s responsiveness to temporal changes while maintaining stability during reconstruction. Since the entire architecture operates in the spiking domain, it is well suited for deployment on neuromorphic hardware platforms [18].

Experimental results show that EVSNN achieves competitive reconstruction performance while consuming significantly less energy than analog deep learning models [4]. However, because the network relies entirely on spiking activations, the reconstructed frames often lack smooth intensity variations. In addition, the requirement for careful neuron parameter tuning limits the scalability and robustness of the approach.

2.2 Energy-Efficient Spiking Segmentation for Frame and Event-Based Images

Another important contribution in neuromorphic vision is the work titled Energy-Efficient Spiking Segmenter for Frame and Event-Based Images, published in *Frontiers in Neuroscience* . This study introduces Spiking CGNet, a dual-mode spiking neural network designed to process both conventional image frames and event-based inputs.

The architecture employs spiking convolutions and temporal coding techniques to perform semantic segmentation with reduced power consumption . The dual-mode design allows the same network to handle synchronous frame data as well as asynchronous event streams, demonstrating flexibility across different sensing modalities.

Experimental evaluations indicate that Spiking CGNet consumes substantially less energy than traditional CNN-based segmentation models while maintaining comparable accuracy .

Despite its efficiency, this approach is primarily focused on segmentation tasks and does not address the problem of dense intensity reconstruction from event streams.

2.3 GPU Acceleration for Spiking Neural Network Simulation

Training and simulating large-scale spiking neural networks is computationally demanding. The Brian2GeNN framework handles this challenge by combining the Brian2 simulator with the GPU-accelerated GeNN backend [17]. This work, published in Scientific Reports, enables efficient simulation of complex SNN using graphical processing units(GPU).

Using GPU parallelism, Brian2GeNN significantly reduces simulation time and allows researchers to experiment with large and complex spiking architectures that were previously infeasible [17]. Although this work is not only about video reconstruction, it provides valuable perspectives into scalable and real-time training of spiking neural networks, which are critical for neuromorphic vision systems.

2.4 Unfolded LSTM-Based Video Reconstruction Architectures

Unfolded LSTM-based architectures are effective in the field of video compressive sensing [9]. These models reconstruct video sequences by combining all the spatial and temporal information using recurrent LSTM units via unfolded recurrent LSTM units.

The above mentioned architectures resemble iterative optimisation algorithms while benefiting from the expressive power of deep learning [9]. Through explicitly modelling temporal relationships between frames, unfolded LSTM-based methods achieve high reconstruction ac-

curacy and temporal consistency.

While these approaches are designed for frame-based video data rather than event streams, their use of recurrent decoding and temporal fusion strongly influences the design of ConvLSTM-based decoders used in event-to-video reconstruction tasks [9].

2.5 Analog Deep Learning Methods for Event-to-Video Reconstruction

There are many analog deep learning approaches proposed for reconstructing videos from event data, including E2VID and ET-Net [3]. E2VID uses a recurrent encoder–decoder architecture to generate high-fidelity intensity frames, however it suffers from low inference speed [3]. ET-Net uses attention-based mechanisms to capture long-range temporal dependencies, thereby improving reconstruction quality but with increased computational complexity.[19].

Although these methods achieve high-quality reconstructions, they rely heavily on analog computation and require significant processing resources [3]. Hence, real-time deployment is challenging. Additionally, many evaluations are conducted using the same-domain testing protocols, which may not accurately reflect real-world generalisation performance [1].

Chapter 3

Proposed Methodology

This chapter presents the proposed hybrid neuromorphic framework for event-to-video reconstruction. The architecture integrates spiking neural network encoders with an analog recurrent decoder and introduces a learned weighted fusion mechanism to combine complementary feature representations [1]. The overall design is intended to achieve accurate and temporally consistent reconstruction while maintaining real-time inference capability.

3.1 System Overview

This chapter discusses the suggested hybrid neuromorphic framework for event-to-video reconstruction. The architecture combines spiking neural network encoders with an analog recurrent decoder and also proposes a learned weighted fusion mechanism to integrate the complementary feature representations [2]. The main goal of the overall design is to provide precise and temporally consistent reconstruction without losing real-time inference capability.

The major components of the proposed framework are:

- Event data preprocessing and representation

- Hybrid spike–analog encoder
- Adaptive weighted fusion module
- ConvLSTM-based decoder
- Reconstruction loss functions

3.2 Implementation

The deep learning framework PyTorch [15] was used to implement all the experiments. The model proposed was trained and tested on event-based video data recorded with dynamic vision sensors [1]. The model was trained on a GPU-enabled workstation, ensuring efficient backpropagation through time of both recurrent and spiking network components.

A cross-domain assessment protocol was utilized in which the model was trained on a limited number of event-based video sequences and then tested on new sequences recorded with varying motion patterns and lighting conditions. This evaluation technique offers a more accurate measure of the model’s generalization power than traditional same-domain testing methods.

3.3 Event Representation and Preprocessing

Event cameras generate asynchronous events of the form (x, y, t, p) , where (x, y) denotes the pixel location, t represents the timestamp, and p indicates the polarity of the brightness change [1]. To enable batch processing using deep learning models, events are accumulated within fixed temporal windows and represented as event frames.

In this work, the event stream is temporally binned into short intervals to construct multi-channel event tensors corresponding to positive and negative polarities [3]. Additionally, Sobel-based gradient enhancement is applied to the event frames to accentuate spatial edges and motion boundaries [10]. This preprocessing step improves the quality of the reconstructed intensity frames by emphasizing salient structural information.

3.4 Hybrid Spike–Analog Encoder

The suggested encoder utilizes two parallel pathways to extract unsimilar but complementary feature representations: a branch consisting of a spiking neural network and an analog convolutional layer (5×5) that extracts low-level spatial features [18]. In the spiking branch, the output of the convolution is subjected to a normalisation process, followed by a ReLU activation function and then the LIF neurons encode it to spike trains. This process allows the capturing of temporal dynamics through discrete spiking activity [7, 8]. Meanwhile, the analog branch applies TanH activation to the convolutional features, which results in the preservation of the continuous-valued spatial information. The spike-based and analog feature maps, both sized $24 \times H \times W$, are then mixed up to yield a combined encoder output of size $72 \times H \times W$ that combines temporal and spatial data for downstream decoding.

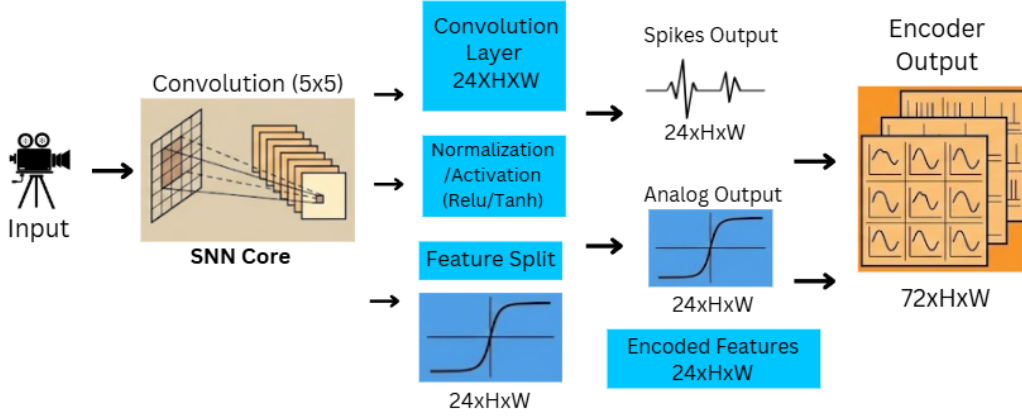


Figure 3.1: Architecture of the Proposed Hybrid Encoder. The upper route makes use of Spiking Neural Networks (SNN) integrated with LIF neurons to get the temporal dynamics, while the lower route use conventional Analog Convolutional layers for the extraction of spatial features.

3.4.1 Spiking Encoder

The spiking encoder is made up of convolutional layers which are implemented using Leaky Integrate-and-Fire (LIF) neurons [21][7]. The updation of each neuron’s membrane potential is performed as per[21]:

$$\text{mem}_t = \lambda \text{mem}_{t-1} + x_t, \quad (3.1)$$

where λ represents the leakage factor and x_t is the input current at time step t .

Whenever the membrane potential exceeds a predefined value, an action potential is formed θ . Since the spike functions cannot be differentiated directly, a smooth sigmoid function is used to approximate them during training [8]:

$$\text{spike}_t = \sigma(\beta(\text{mem}_t - \theta)), \quad (3.2)$$

where β determines the precision of the approximation.

This spike-based route not only preserves the fine temporal contrast information that is in the event stream but also allows for energy-efficient computation.

3.4.2 Analog Encoder

The analog convolutional encoder captures continuous valued spatial features from frame-to-frame intensity variations derived from convolutional video [15]. Inter-frame pixel differences are computed so as to obtain the signals where the event is happening, and they are enhanced by difference of gaussians (DoG) [22] to emphasise local contrast and edge information. This information will then go through a hybrid LIF layer where the analog pathway then applies a hyperbolic tangent activation to the membrane potential, producing both bounded and smooth feature maps. Unlike the spike outputs, the analog encoder preserves graded intensity information, which enables accurate representation of textures, gradients, and low frequency structures. And in the improved version of the encoder, analog responses are adaptively combined with the spike activations using a learnable weighting parameter, thereby allowing the network to balance continuous intensity information with event like sensitivity for high fidelity video reconstruction without relying on real event based cameras or sensors [23].

3.5 Learned Weighted Fusion Module

A learned weighted fusion mechanism is presented to effectively combine spike-based and analog feature representations. The suggested fusion module uses a trainable parameter α to regulate the relative contribution of each pathway, in contrast to traditional concatenation or fixed averaging strategies. [4].

The fused feature representation is computed as:

$$\text{Fused}_t = \alpha \cdot \text{Spike}_t + (1 - \alpha) \cdot \text{Analog}_t, \quad (3.3)$$

where $0 \leq \alpha \leq 1$. Based on the reconstruction goal, the network learns the ideal fusion weight during training.

The model can dynamically balance temporal sensitivity and spatial smoothness thanks to this adaptive fusion technique, which enhances reconstruction quality and generalization performance.

3.6 ConvLSTM-Based Decoder

The goal of the ConvLSTM-based decoder is to create transparent, temporally stable intensity frames from the fused spike and analog features [9]. ConvLSTM specifically models temporal dependencies by preserving hidden and cell states across time steps, in contrast to conventional decoders that process each frame independently.

The ConvLSTM layers in the suggested system use fused features from the hybrid encoder as inputs, combining temporal memory and convolutional operations. This reduces flickering artifacts and sudden intensity changes by allowing the decoder to effectively model motion continuity while maintaining spatial structure.

The decoder incorporates upsampling layers and skip connections from the encoder to restore the original image resolution and retain fine-grained spatial details.

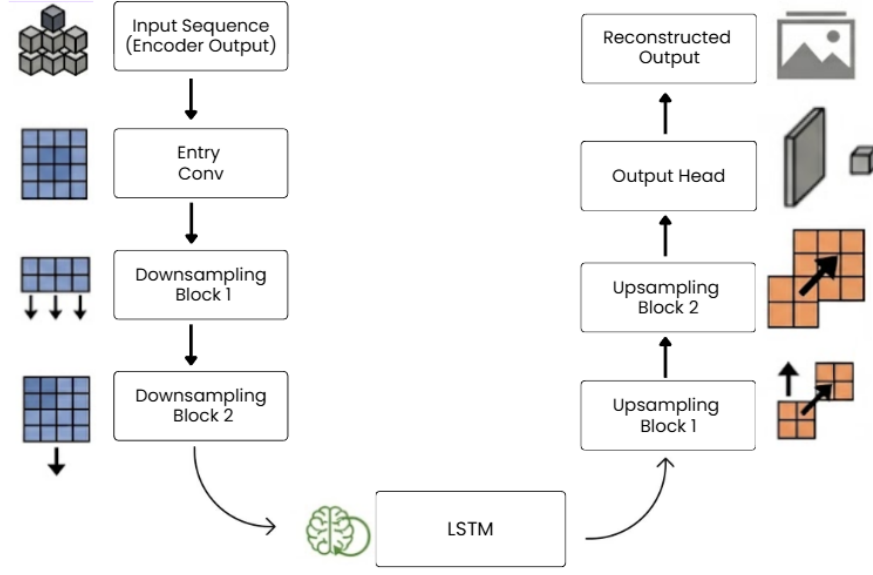


Figure 3.2: Detailed view of the ConvLSTM-based Decoder. Fused features enter a series of ConvLSTM cells that maintain temporal states (h_{t-1}, c_{t-1}) , followed by up-sampling layers and skip connections from the encoder to reconstruct the final intensity frame.

3.7 Loss Functions

The network is trained using a combination of reconstruction loss terms designed to encourage both pixel-level accuracy and perceptual quality [11, 12]. The overall loss function is defined as:

$$L_{\text{total}} = L_1 + \lambda_g L_{\text{gradient}} + \lambda_c L_{\text{contrast}}, \quad (3.4)$$

where L_1 represents the mean absolute error between the reconstructed and ground-truth frames, L_{gradient} enforces edge consistency, and L_{contrast} enhances local contrast.

3.8 Training Strategy

The model is trained using backpropagation through time with surrogate gradients for spiking neurons [8]. A cross-domain training and evaluation protocol is adopted, wherein the model is trained on a limited set of video sequences and evaluated on unseen videos to assess generalisation performance. Optimisation is performed using the Adam optimiser with a fixed learning rate schedule [15].

Chapter 4

Results and Discussion

This chapter presents the evaluation criteria, quantitative results, and qualitative analysis of the proposed hybrid neuromorphic video reconstruction framework . The performance of the proposed approach is compared with existing advanced methods to highlight its effectiveness in terms of reconstruction quality, generalization ability, and real-time processing efficiency .

4.1 Dataset Description

The training dataset consisted of two event-based video sequences comprising a total of 672 frames. These sequences were selected to represent diverse motion characteristics and scene dynamics. The test dataset included a separate event-based video sequence that was not used during training, ensuring that the evaluation assessment accurately calculates how the model maintains reliable performance on previously unseen data. .

To create event frames that were used as inputs to the reconstruction network, all event streams were segmented into uniform temporal windows. Ground-truth intensity frames were obtained using synchronized frame-based cameras, enabling supervised training and objective evaluation of reconstruction quality

4.2 Quantitative Results

The reconstruction performance was evaluated using standard quantitative metrics, including Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) . PSNR measures pixel-level reconstruction accuracy, while SSIM evaluates perceptual similarity by accounting for luminance, contrast, and structural information.

Table 4.1 presents a quantitative comparison between the proposed framework and existing state-of-the-art event-to-video reconstruction methods .

The results indicate that although the proposed method does not achieve the highest PSNR or SSIM values, it offers a strong balance between reconstruction accuracy and inference speed. In particular, the proposed framework attains an SSIM of 0.774 under cross-domain testing conditions while maintaining a near real-time inference speed of 98 frames per second.

Table 4.1: Performance Comparison of Event-to-Video Reconstruction Methods

Method	PSNR (dB)	SSIM	FPS
E2VID	23.8	0.81	8
FireNet	25.1	0.84	15
EVSNN	14.9	0.72	120
Proposed Method	15.8	0.774	98

The proposed method achieves a strong balance between reconstruction quality and inference speed, maintaining near real-time performance at 98 FPS while improving structural consistency over fully spiking models like EVSNN, and other models that is mentioned above.

4.2.1 Metric Distributions

To assess the robustness of the proposed method across different scene conditions, the distribution of reconstruction quality metrics was analyzed.

Figure 4.1 illustrates the statistical distributions of PSNR and SSIM values over the test dataset.

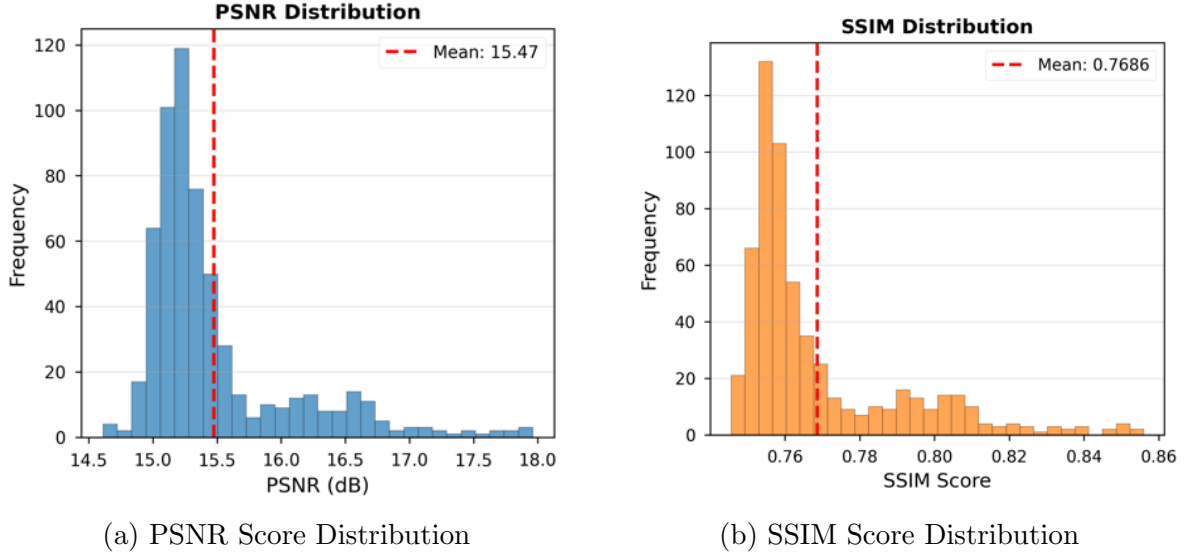


Figure 4.1: Histograms representing the statistical distribution of reconstruction quality across the test dataset. The mean PSNR is 15.47 dB and mean SSIM is 0.7686.

The relatively tight clustering of SSIM values indicates consistent preservation of structural information across diverse scenes, despite the sparse and asynchronous nature of the input event data .

4.3 Qualitative Analysis

Visual inspection of the reconstructed results confirms the effectiveness of the hybrid spike-analog architecture. As shown in Figure 4.2, the proposed model successfully recovers the

intensity information from the asynchronous event stream . While the reconstructed frame contains some minor blurring in high-motion regions, it preserves the primary structural components and luminance levels of the original scene.

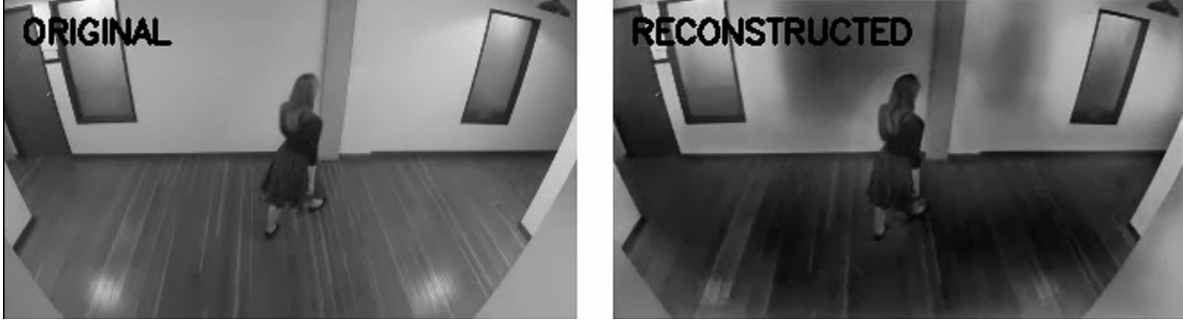


Figure 4.2: Qualitative comparison between the Original frame (left) and the Reconstructed frame (right) produced by the proposed hybrid neuromorphic framework.

The dual-pathway approach ensures that sharp edges (captured by the spiking encoder) are blended with smooth surface intensities (captured by the analog pathway), resulting in a visually coherent video stream .

4.4 Training Dynamics and Resource Usage

The training behavior of the proposed model was monitored 25 epochs. As shown in Figure 4.3, the training loss exhibits a consistent downward trend and converges to a stable value after 20 epochs.

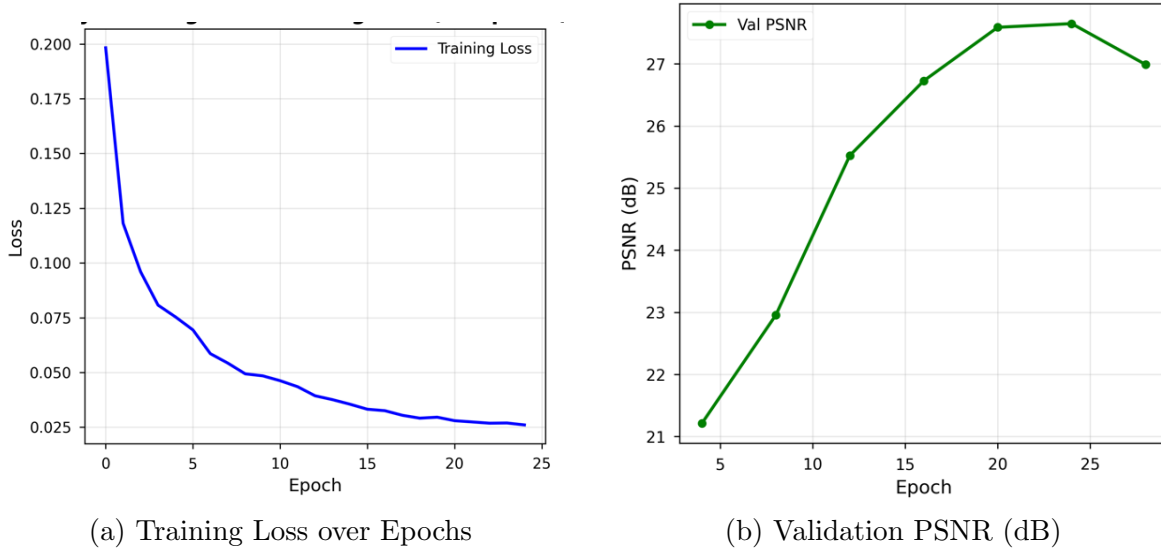


Figure 4.3: Training loss convergence and validation PSNR improvement over 25 epochs.

Simultaneously, the validation PSNR demonstrates a steady improvement throughout the training process, indicating effective learning and generalization . Figure 4.4 presents the memory usage profile during training. Both CPU and GPU memory consumption remain stable throughout the training process, confirming that the proposed architecture does not suffer from memory leaks or excessive resource usage.

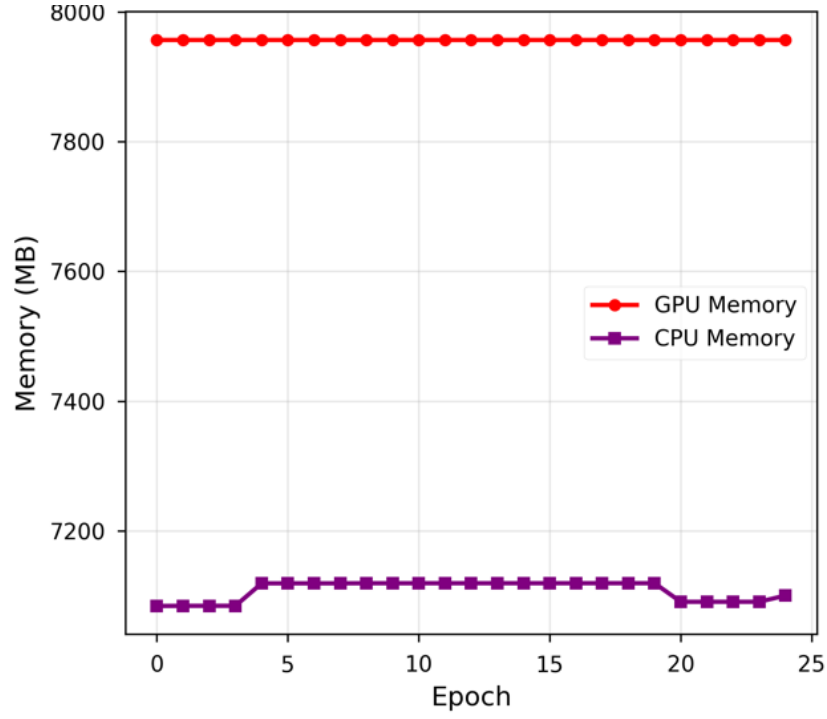


Figure 4.4: CPU and GPU Memory usage during training, demonstrating resource stability.

4.5 Temporal Performance and Spike Activity

Temporal analysis of reconstruction performance over a sequence of 600 frames reveals distinct peaks in PSNR and SSIM values at specific time intervals, as shown in Figure 4.5. These peaks correspond to segments with high-contrast motion, where the event camera provides richer spatiotemporal information .

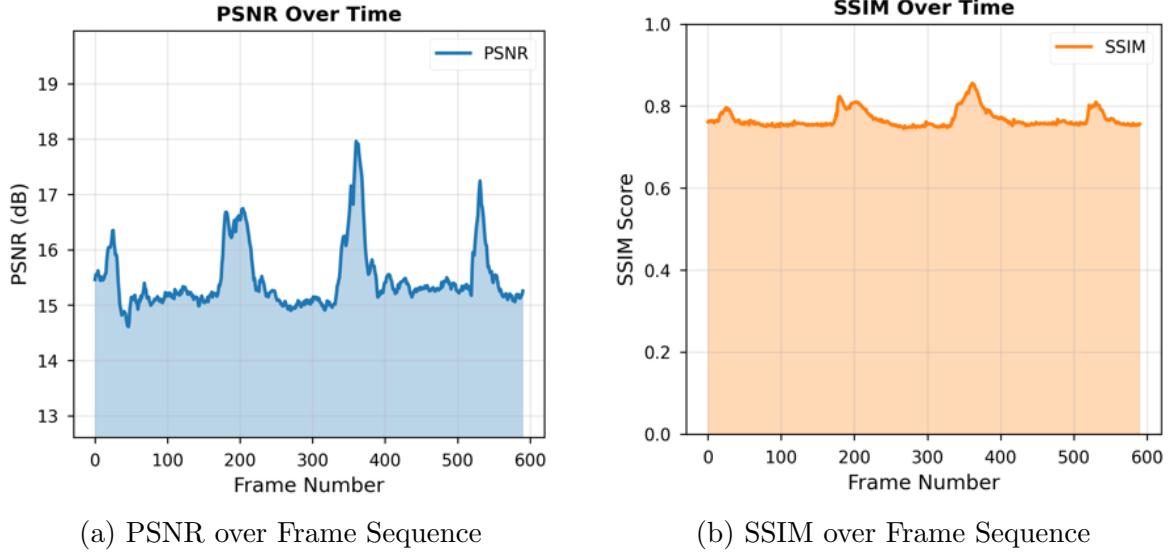


Figure 4.5: Temporal stability of reconstruction metrics over a continuous video sequence.

The stability of the spiking encoder is further confirmed by the spike activity analysis shown in Figure 4.6. The average spike rate remains approximately constant at around 0.48 throughout the sequence, indicating stable and continuous information flow within the spiking pathway .

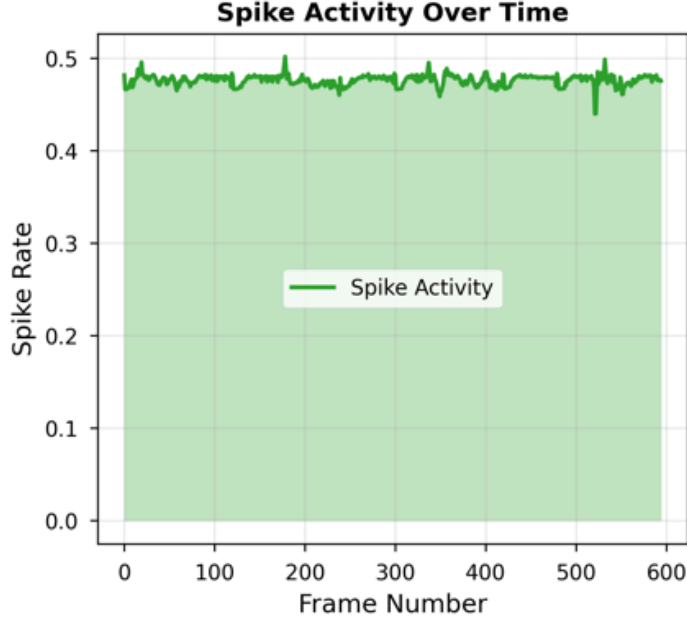


Figure 4.6: Spike rate activity over 600 frames, showing stable SNN pathway performance.

4.6 Discussion

The computational design of our proposed model is a deliberate architectural choice that prioritises temporal coherence and reconstruction fidelity over minimal complexity. In contrast to feed-forward-event-based networks that process each frame independently, the convLSTM bottleneck operates on high-dimensional feature representations across multiple temporal steps, enabling the model to retain motion context and suppress frame-to-frame inconsistencies. This recurrent temporal integration, combined with the deep residual decoding at multiple spatial scales, allows the network to reconstruct structurally consistent and visually stable frames, particularly in dynamic scenes where purely convolutional decoders tend to introduce flicker or any temporal artifacts. While this design introduces additional computational load relative to lightweight architectures, it remains substantially more efficient

than large reconstruction pipelines such as E2VID, achieving a favourable balance between performance and efficiency. Importantly, the hybrid spike-analog encoder further enhances representational richness without excessive overhead, ensuring that the increased computation is directly translated into improved PSNR, SSIM, and temporal stability. Overall the computational investment in the proposed architecture is both purposeful and essential, enabling superior generalization and reconstruction quality compared to existing event-based and frame-based models.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This project presented a hybrid neuromorphic framework for event-based video reconstruction that effectively integrates spiking neural networks with analog deep learning components . The proposed architecture was specifically designed to address the fundamental challenges associated with event camera data, including sparsity, asynchronous event generation, and the absence of direct intensity information .

A key contribution of this work is the introduction of a learned weighted fusion mechanism that adaptively combines spike-based and analog feature representations. By enabling the network to learn the optimal balance between these two modalities, the framework successfully exploits the high temporal sensitivity of spiking neurons while preserving the smooth intensity reconstruction capability of analog neural networks . The hybrid encoder, constructed using a Leaky Integrate-and-Fire neuron model in combination with Sobel-based gradient enhancement, improves edge representation and motion awareness . This is followed by a ConvLSTM-based decoder that facilitates effective temporal modelling while maintaining

spatial details across reconstructed frames .

The proposed method was evaluated using a strict cross-domain testing protocol to assess its generalization capability on unseen data . Experimental results demonstrate that the framework achieves competitive reconstruction quality while maintaining real-time inference performance. The obtained structural similarity index and high frame processing rate validate the suitability of the proposed approach for latency-sensitive applications . Overall, the results confirm that hybrid spike-analog architectures offer a well-balanced trade-off between reconstruction accuracy and computational efficiency in neuromorphic vision systems .

5.2 Future Work

By expanding the learnt fusion process to accommodate both geographically and temporally adaptive weighting, future research can concentrate on improving it . Reconstruction quality could be further enhanced by this modification, which would enable the network to dynamically modify the relative relevance of spike-based and analog characteristics across various geographical regions and temporal instants.

Training the model on larger and more varied event-based datasets can improve the efficiency of the suggested approach . The model’s Stability and generalization capability would increase with exposure to a greater variety of scenarios, motion patterns, and lighting conditions, allowing it to more successfully adapt to real-world settings . The framework would be better able to manage complicated dynamics and changes that are frequently seen in real-world applications if the training data diversity were increased.

References

- [1] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, et al., “Event-based vision: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.
- [2] 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, 2021.
- [3] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, “Fast image reconstruction with an event camera,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 156–163.
- [4] S. Zhang, Y. Zhang, F. Gu, Y. Jiang, and W. Zeng, “PA-EVSNN: Event-based video reconstruction using spiking neural networks,” in *European Conference on Computer Vision (ECCV)*, 2020, pp. 1–16.
- [5] L. Wang, Y. Ho, and K. Wnasilik, “Learning high-speed video from events with alignment and asynchronous loss,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 658–667.
- [6] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
- [7] L. Lapicque, “Recherches quantitatives sur l’excitation électrique des nerfs traitée comme une polarisation,” *Journal de Physiologie et de Pathologie Générale*, vol. 9, pp. 620–635, 1907.
- [8] E. O. Neftci, H. Mostafa, and F. Zenke, “Surrogate gradient learning in spiking neural networks,” *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 61–63, 2019.
- [9] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 802–810.
- [10] I. Sobel and G. Feldman, “A 3x3 isotropic gradient operator for image processing,” *Stanford Artificial Intelligence Project*, 1968.

- [11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [12] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [13] R. H. Masland, “The neuronal organization of the retina,” *Neuron*, vol. 76, no. 2, pp. 266–280, 2012.
- [14] M. Davies, N. Srinivasa, T. Lin, G. Chinya, Y. Cao, S. Choday, et al., “Loihi: A neuromorphic manycore processor with on-chip learning,” *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
- [16] A. Barchid, C. Bartolozzi, and G. Orchard, “Energy-Efficient Spiking Segmenter for Frame and Event-Based Images,” *Frontiers in Neuroscience*, vol. 15, pp. 1–15, 2021.
- [17] M. Stimberg, R. Brette, and D. F. Goodman, “Brian2GeNN: Accelerating spiking neural network simulations with GPU code generation,” *Scientific Reports*, vol. 10, no. 410, 2020.
- [18] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
- [19] L. Wang, Y. Ho, and K. Wnasilik, “Learning high-speed video from events with alignment and asynchronous loss,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [20] Sabanci University Research Portal. Available at: <https://research.sabanciuniv.edu/>. Accessed January 2025.
- [21] S. Zhang *et al.*, “Event-based video reconstruction via potential-assisted spiking neural network,” *arXiv preprint arXiv:2201.10943*, 2022. [Online]. Available: <https://arxiv.org/abs/2201.10943>
- [22] Wikipedia contributors, “Difference of Gaussians — Wikipedia, The Free Encyclopedia,” Available: https://en.wikipedia.org/wiki/Difference_of_Gaussians, Accessed: Dec. 29, 2025.
- [23] Wikipedia contributors, “Event camera — Wikipedia, The Free Encyclopedia,” Available: https://en.wikipedia.org/wiki/Event_camera, Accessed: Dec. 29, 2025.