ST-FlowNet: An Efficient Spiking Neural Network for Event-Based Optical Flow Estimation

Hongze Sun[a], Jun Wang[a], Wuque Cai[a], Duo Chen[a,b], Qianqian Liao[a], Jiayi He[a], Yan Cui[a,c], Dezhong Yao[a,d,*], Daqing Guo[a,d,*]

[a]*Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for NeuroInformation, China-Cuba Belt and Road Joint Laboratory on Neurotechnology and Brain-Apparatus Communication, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China.*
[b]*School of Artificial Intelligence, Chongqing University of Education, Chongqing 400065, China.*
[c]*Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, Chengdu 610072, China.*
[d]*Research Unit of NeuroInformation (2019RU035), Chinese Academy of Medical Sciences, Chengdu 611731, China.*

Abstract

Spiking neural networks (SNNs) have emerged as a promising tool for event-based optical flow estimation tasks due to their capability for spatio-temporal information processing and low-power computation. However, the performance of SNN models is often constrained, limiting their applications in real-world scenarios. To address this challenge, we propose ST-FlowNet, a novel neural network architecture specifically designed for optical flow estimation from event-based data. The ST-FlowNet architecture integrates ConvGRU modules to facilitate cross-modal feature augmentation and temporal alignment of the predicted optical flow, thereby improving the network's ability to capture complex motion patterns. Additionally, we introduce two strategies for deriving SNN models from pre-trained artificial neural networks (ANNs): a standard ANN-to-SNN conversion pipeline and our proposed BISNN method. Notably, the BISNN method alleviates the complexities involved in selecting biologically inspired parameters, further enhancing the robustness of SNNs for optical flow estimation tasks. Extensive evaluations on three benchmark event-based datasets demonstrate that the SNN-based ST-FlowNet model outperforms state-of-the-art methods, achieving superior accuracy in optical flow estimation across a diverse range of dynamic visual scenes. Furthermore, the energy efficiency of models also underscores the potential of SNNs for practical deployment in energy-constrained environments. Overall, our work presents a novel framework for optical flow estimation using SNNs and event-based data, contributing to the advancement of neuromorphic vision applications.

Keywords: Spiking neural networks, Optical flow estimation, Event-based images, Training methods.

1. Introduction

As the third generation of neural networks (Maass, 1997), spiking neural networks (SNNs) have garnered increasing attention in recent years (Zheng et al., 2021). Unlike conventional artificial neural networks (ANNs), SNNs employ bio-inspired neurons and discrete spike trains to mimic the complex spatiotemporal dynamics of the human brain (Zheng et al., 2024; Sun et al., 2024). These characteristics enable SNNs to achieve competitive performance across a wide range of neuromorphic intelligent tasks (Ussa et al., 2024; Yu et al., 2023a). Furthermore, the event-driven information coding and transmission mechanisms ensure lower power consumption, thus enhancing the feasibility of SNNs for hardware implementation (Yao et al., 2023).

Optical flow estimation is a fundamental topic in the field of computer vision and has extensive applications (Ilg et al., 2017; Vandal and Nemani, 2023), particularly in motion-related intelligent tasks. For instance, by clustering motion patterns in different regions, optical flow can assist object segmentation models in separating the foreground and background (Zitnick et al., 2005). The motion vector of the target object is critically important for predicting the search space in object tracking tasks (Ussa et al., 2024). Optical flow also serves as a compensation tool for frame insertion-based image reconstruction and enhancement (Fan et al., 2021; Wang et al., 2020). However, prevalent research has predominantly concentrated on optical flow estimation from frame-based images captured by conventional cameras, resulting in significant performance degradation in challenging scenarios characterized by high-speed motion or unfavorable lighting conditions (Zhai et al., 2021). Event-based neuromorphic cameras, capable of asynchronously recording changes in light intensity within a high dynamic range of illumination (Han et al., 2020), respond to these challenging scenarios effectively and exhibit potential for minimizing energy consumption, thereby presenting an attractive prospect for deployment on edge devices (Yu et al., 2023a). Furthermore, drawing inspiration from inherent imaging mechanisms, event-based images can circumvent errors introduced by assumptions related to the conservation of pixel intensity in optical flow estimation (Gaur, 2022). Thus, the advantages of event-based optical flow estimation are significant when compared with conventional approaches (Zhu et al., 2018b; Mueggler et al., 2017;

*Corresponding authors: dyao@uestc.edu.cn (Dezhong Yao) and dqguo@uestc.edu.cn (Daqing Guo).

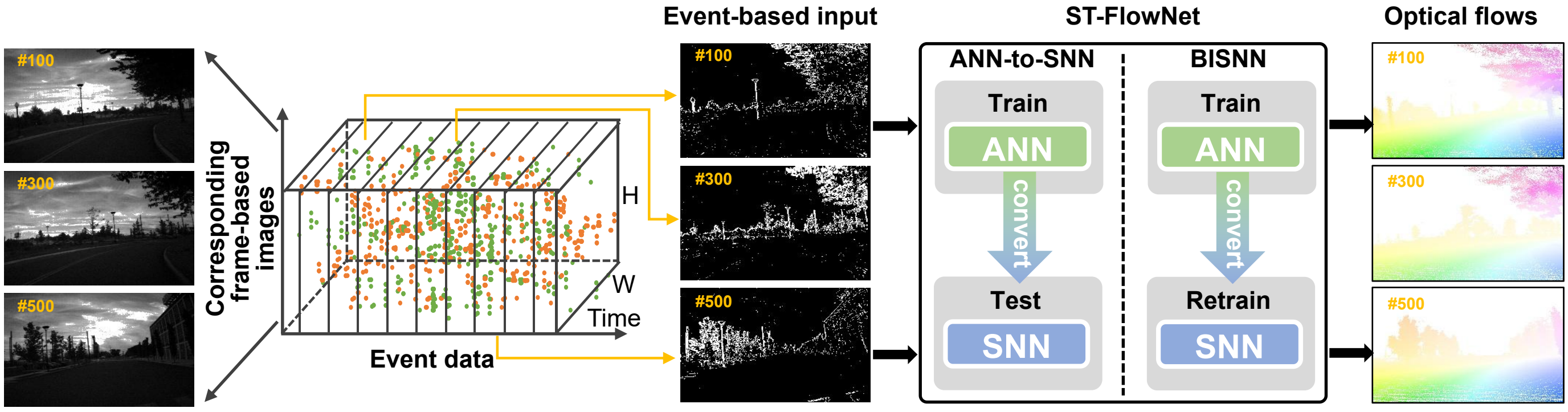

Figure 1: The framework of our proposed optical flow estimation method is illustrated. The ST-FlowNet (both ANN and SNN) models utilize event-based images as input data. Following training on the ANN model, an SNN ST-FlowNet model is derived through the A2S conversion or BISNN method. Optical flow prediction is achievable using both ANN and SNN models. Additionally, for reference, the corresponding frame-based images are presented in the left black box. Conventional frame-based images exhibit abundant spatial texture information indiscriminately, while event-based images emphasize motion-related objects by leveraging spatio-temporal cues simultaneously.

Scheerlinck et al., 2020).

Owing to the structural specificity inherent in the event modality, current approaches typically involve the reconstruction of event data into frame-based images for the estimation tasks handled by conventional ANN models (Zhu et al., 2018a; Ilg et al., 2017). Despite some notable progress, these approaches often neglect potential errors inherent in the reconstruction process. Representatively, the resulting reconstructed images frequently exhibit pronounced motion blur in spatial visual cues, particularly as the sampling time window lengthens (Stoffregen et al., 2020; Tian and Andrade-Cetto, 2022). Concurrently, the valuable dynamic features embedded in the temporal domain remain underutilized. In contrast, an SNN model, comprising spiking neurons as fundamental units, accepts spiking events as input and uses firing spikes as a medium for information propagation and presentation (Cai et al., 2024). The SNN is an essential tool for event-based feature extraction (Sun et al., 2024). Consequently, a natural solution to address the aforementioned challenges involves developing an SNN model specifically tailored for event-based optical flow estimation.

Unlike conventional ANN models, training SNN models with standard backpropagation (BP) is challenging due to the non-differentiable nature of spike signals (Wu et al., 2018), leading to suboptimal performance in real-world applications. To address this challenge, several methods have been proposed, which can be classified into four main categories: (1) unsupervised learning inspired by biological neuronal plasticity (Diehl and Cook, 2015), (2) indirect training via ANN-to-SNN conversion (A2S) (Deng and Gu, 2021), (3) spatio-temporal backpropagation (STBP) employing surrogate gradient approximation (Wu et al., 2018), and (4) hybrid training strategies (Sun et al., 2024). Among these, STBP has emerged as the most widely adopted approach, enabling SNN training with a procedure similar to that of ANNs while maintaining competitive accuracy. However, STBP-based SNN models typically require extended temporal windows for effective training, resulting in significant computational overhead. Moreover, approximation errors introduced by surrogate gradients further constrain their performance in complex real-world tasks. Consequently, A2S methods have gained prominence as an alternative, allowing the derivation of SNN models from pre-trained ANNs. Compared to the STBP method, A2S strategy substantially reduces training complexity, particularly for intricate task-specific models. Additionally, hybrid strategies integrating diverse training paradigms or neuronal plasticity mechanisms have attracted growing interest, further enhancing SNN training efficiency and model performance.

To fully exploit the potential of event data in optical flow estimation, it is imperative to effectively address two key challenges: (1) developing a novel architecture capable of simultaneous spatio-temporal feature extraction based on event data; and (2) proposing a novel model training strategy for achieving superior optical flow estimation performance. Thus far, limited research has delved into these pertinent issues. Therefore, this work introduces a novel method to effectively tackle these challenges (shown in Fig. 1). We present our ST-FlowNet architecture tailored specifically to estimate event-based optical flow. By incorporating the ConvGRU layers (Ballas et al., 2016), ST-FlowNet achieves cross-scale fusion of dynamic optical flow features from event data in the temporal dimension. In contrast to prior models inspired by pyramidal architectures (Ilg et al., 2017), ST-FlowNet employs a more streamlined decoder structure, enabling direct decoding of latent information across the entire multi-scale feature space. Following training on the ANN model, the SNN model for optical flow estimation can be derived through two strategies: the A2S method or our proposed bio-information-fused training strategy (BISNN). Notably, our BISNN approach achieves parameter-free model conversion while preserving the performance of optical flow estimation.

Our key contributions in this paper are summarized as follows:

- We propose ST-FlowNet, a novel architecture designed for efficient optical flow estimation by leveraging spatio-temporal features in event-based data.

- We present the first framework for constructing efficient SNN models for optical flow estimation, utilizing an A2S conversion approach.
- We introduce a novel parameter-free SNN training strategy, which further mitigates training challenges while enhancing overall training efficiency.
- Our experimental results demonstrate that ST-FlowNet attains superior performance when compared with other state-of-the-art models on challenging benchmark datasets such as MVSEC, ECD, and HQF.

2. Related Work

Optical flow estimation is a fundamental task in the field of computer vision (Ilg et al., 2017), garnering significant research attention. In this section, we initially delve into the evolution of event-based optical flow estimation. Subsequently, our focus shifts to the models built by SNNs, which demonstrate notable proficiency in extracting spatio-temporal features from event-based images.

2.1. Event-based Optical Flow Estimation

Considering the distinctive attributes of event images in contrast to the RGB modality, early research focused on innovating paradigms for event-based optical flow computation, but only achieved limited success in several simple scenarios (Benosman et al., 2013). The introduction of large-scale benchmark datasets has facilitated applying deep learning models to event-based optical flow estimation tasks (Zhu et al., 2018b; Mueggler et al., 2017), significantly enhancing the efficacy of optical flow estimation and reducing the challenges associated with model design (Scheerlinck et al., 2020; Tian and Andrade-Cetto, 2022). Previous research attempted to use encoder-decoder networks to decouple spatio-temporal features embedded in event data across multiple resolutions (Zhu et al., 2018a). A series of variants emerged through the refinement of network structures and adjustments to loss functions, achieving reliable and accurate optical flow estimation facilitated by self-supervised and end-to-end learning methods (Scheerlinck et al., 2020; Tian and Andrade-Cetto, 2022). However, considering the high temporal resolution characteristics inherent in event data, models founded on general convolution structures struggle to comprehensively use the temporal features of the event data, thereby limiting the effectiveness of event-based optical flow estimation.

2.2. SNN Models for Optical Flow Estimation

Considering the inherent advantages of SNNs in spatio-temporal feature extraction (Tavanaei et al., 2019), our work focuses on models using spiking neurons which are naturally adept at capturing spatio-temporal visual cues embedded in event data for precise optical flow estimation. Recently, numerous models based on spiking neurons have been proposed for event-based optical flow estimation (Paredes-Vallés et al., 2019; Hagenaars et al., 2021; Zhang et al., 2023). Similar to traditional ANN models, shallow convolutional SNN models trained using spike-time-dependent-plasticity learning methods (Diehl and Cook, 2015) have been introduced, demonstrating promising performance for relatively simple tasks (Paredes-Vallés et al., 2019). To enhance SNN model performance in complex real-world scenarios, a logical approach involves constructing hybrid models that leverage the strengths of both ANNs and SNNs (Diehl and Cook, 2015). Additionally, drawing inspiration from the BP-style direct training method for SNN models (Wu et al., 2018), a series of fully spiking models have also attained state-of-the-art performance on event-based data when compared with advanced ANN models (Hagenaars et al., 2021; Zhang et al., 2023). However, constrained by training approaches and network architectures, we posit that event-based optical flow estimation by SNN models can be further optimized by ANN conversion (Deng and Gu, 2021).

3. Preliminary

To facilitate a clearer understanding of our work, this section provides a brief overview of SNNs. To date, a variety of spiking neuron models have been developed to emulate the spatio-temporal dynamics of biological neurons. Among these, the leaky integrate-and-fire (LIF) model has become the foundational computational unit in constructing SNNs, offering a balance between biological plausibility and computational efficiency (Abbott, 1999). The membrane potential dynamics of an LIF neuron are typically governed by the following differential equation:

$$\tau \frac{dv(t)}{dt} = V_{\text{rest}} - v(t) + I(t), \tag{1}$$

accompanied by the spike generation mechanism defined as:

$$s(t) = \begin{cases} 1, & \text{if } v(t) \geq \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $v(t)$, $I(t)$, and $s(t)$ denote the membrane potential, input current, and output spike at time t, respectively. The parameters τ, V_{rest}, and θ represent the membrane time constant, resting potential, and firing threshold. A spike is emitted when the membrane potential $v(t)$ reaches or exceeds the threshold θ, after which the membrane potential is reset.

For implementation convenience, the continuous-time differential Eqs (1) and (2) are typically approximated in discrete-time form. The membrane potential update at each time step can be expressed as:

$$v(t) = \alpha(t) \cdot \tilde{v}(t-1) + \beta \cdot I(t), \tag{3}$$

where V_{rest} is assumed to be zero for simplicity. The parameters $\alpha(t)$ and β control the decay of the membrane potential and the scaling of the input current, respectively. Here, $\tilde{v}(t-1)$ represents the reset membrane potential at the previous time step, and is defined as:

$$\tilde{v}(t-1) = \begin{cases} v(t-1) \cdot (1 - s(t-1)), & \text{for hard reset}, \\ v(t-1) - s(t-1) \cdot \theta, & \text{for soft reset}, \end{cases} \tag{4}$$

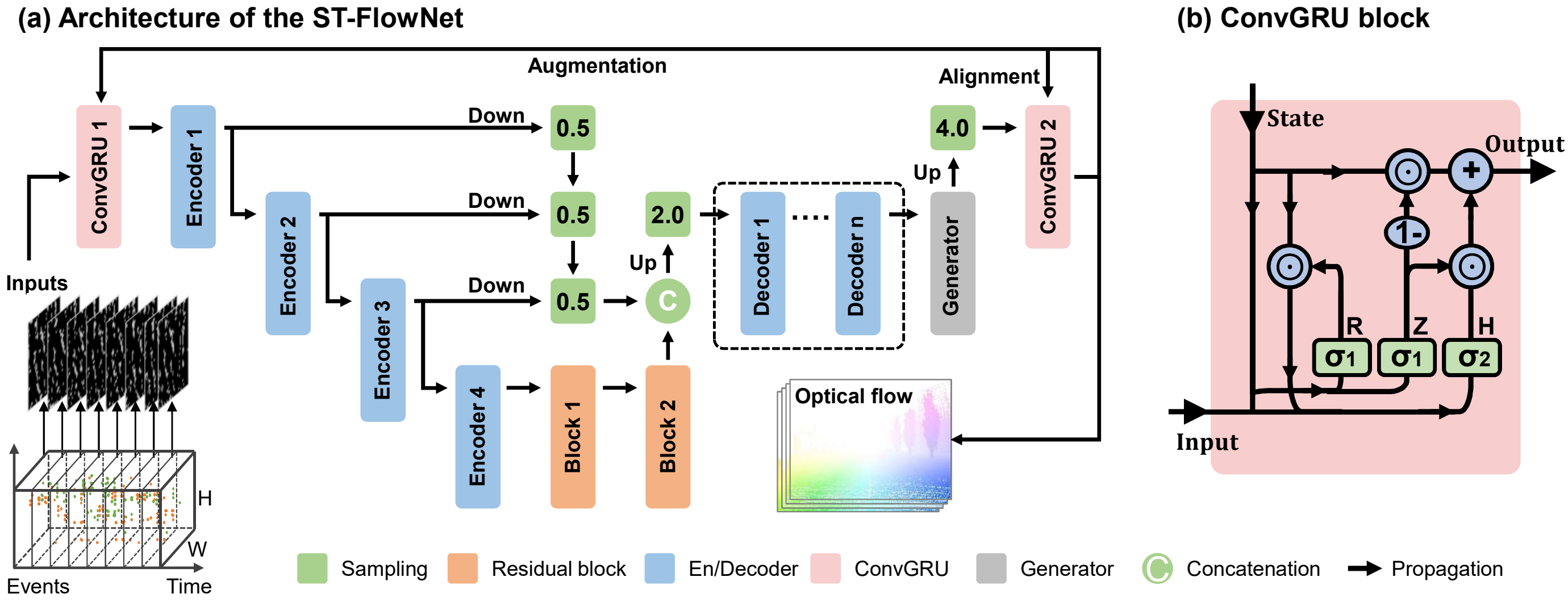

Figure 2: (a) The ST-FlowNet architecture is illustrated. Following pre-processing by a ConvGRU layer, the enhanced event-based input undergoes downsampling via four encoder layers. The resulting minimal feature maps produced by encoder4 traverse two residual block layers, ensuring robust feature extraction. Through the concatenation of feature maps at various levels, decoder layers and a generator are deployed for basic optical flow prediction. Furthermore, the basic predicted optical flow is fed through a ConvGRU layer to fuse historical sequential temporal feature and generate the final predicted optical flow. (b) Schematic illustration depicting the architecture of ConvGRU. A ConvGRU unit integrates both current input and state information to produce a corresponding output. The symbol ⊙ denotes the Hadamard product, and σ is the activation function.

where $s(t-1)$ indicates whether a spike was emitted at the previous time step. In the hard reset mechanism, the membrane potential is set to zero following a spike, while in the soft reset case, it is reduced by the threshold value θ.

4. Method

We present an overview of the proposed ST-FlowNet model in Fig. 2. Specifically, ST-FlowNet builds on the foundation of the end-to-end FlowNet model (Ilg et al., 2017), characterized by encoder and decoder layers. To effectively leverage the event modality, which inherently possesses rich spatio-temporal information, we incorporate ConvGRU layers (Cho et al., 2014) for sequential feature argumentation and alignment. Additionally, we use SNN models derived from pre-trained ANN models to enhance the ST-FlowNet's proficiency in spatio-temporal feature extraction. In this section, we elaborate on the event representation, network architecture, and training methodology used for ST-FlowNet.

4.1. Event Representation

In contrast to conventional frame-based images, which capture light intensity at discrete exposure times, event-based images ($\{[x_k, y_k, t_k, p_k]\}_{k=1}^{K}$) asynchronously record changes in light intensity at location $[x_k, y_k]$ along with their respective polarity ($p_k \in \{p^+, p^-\}$). Here, K represents the total number of events, and t_k denotes the timestamp of the k-th event. For the sake of convenience, the event inputs are aggregated into N group of frame-based representations $f_n^+(\tilde{x}, \tilde{y})$ and $f_n^-(\tilde{x}, \tilde{y})$ ($n = 1, ..., N$), which are defined by the following formulas:

$$f_n^+(\tilde{x}, \tilde{y}) = \textstyle\sum_i p_i^+, \tag{5}$$

$$f_n^-(\tilde{x}, \tilde{y}) = \textstyle\sum_i p_i^-, \tag{6}$$

subject to $i = \frac{K}{N}(n-1)+1, ..., \frac{K}{N}n$. The inputs to the models are presented as:

$$\text{ANN:}[f_1^+, f_1^-, ..., f_N^+, f_N^-], \tag{7}$$

$$\text{SNN:}\{[f_1^+, f_1^-], ..., [f_N^+, f_N^-]\}. \tag{8}$$

Concretely, we concatenate N groups of event-based images into a $2N$-channel data representation (Eq. 7), which serves as input to the ANN model, ensuring the complete information of the inputs is captured. For the SNN model, the value of N determines the temporal resolution of the raw data, and each pair of images is sequentially processed by the model (Eq. 8).

4.2. Network Architecture

As illustrated in Fig. 2, the ST-FlowNet model is structured in a semi-pyramidal encoder-decoder architecture (Ilg et al., 2017). Diverging from prior approaches, we incorporate ConvGRU layers to enhance the model's sequential feature representation. The ConvGRU layer retains its original architecture (shown in Fig. 2(b)). However, the hidden state, which is used to modify information across different time steps, has been redesigned.

Specifically, to process the input data, a **ConvGRU** layer is employed to combined the present input I_t with the inherent information O_{t-1}. The augmented input data $\tilde{I}_t$ is characterized as follows:

$$\tilde{I}_t = (1 - Z_t) \odot O_{t-1} + Z_t \odot H_t. \tag{9}$$

R_t, Z_t and H_t represent the reset gate, update gate and candidate hidden state respectively, and are calculated as follows:

$$R_t = \sigma_1(\xi_{\mathrm{R}}(O_{t-1}, I_t)), \tag{10}$$

$$Z_t = \sigma_1(\xi_{\mathrm{Z}}(O_{t-1}, I_t)), \tag{11}$$

$$H_t = \sigma_2(\xi_{\mathrm{I}}(R_t \odot O_{t-1}, I_t)). \tag{12}$$

Here, ξ_R, ξ_Z and ξ_I represent convolutional operations, σ_1 and σ_2 are the sigmoid and tanh activation functions. Intuitively, the candidate hidden state H_t balances and retains pertinent information from the previous optical flow and current input (Eq.12). The reset gate R_t governs the degree to which historical information is removed in H_t. Ultimately, the augmented input $\tilde{I}_t$ undergoes an update via a weighted summation of O_{t-1} and H_t (Eq.9). O_{-1} is established as 0 at the initial time step $t = 0$.

For every augmented input $\tilde{I}_t$, ST-FlowNet uses a convolutional architecture as **Encoders** to derive feature maps across multiple hierarchical levels, spanning from low-level to high-level representations (denoted as $\{F_t^l\}_{l=1}^4$). The dimensions of the feature maps undergo a reduction by half in each subsequent layer, while the channel count experiences a twofold increment. Furthermore, the resultant highest-level representation F_t^4 undergoes processing via two residual **Block**, aimed at conducting a more in-depth exploration of the underlying deep features (denoted as $\tilde{F_t^4}$).

To maximize the exploitation of optical flow information across various spatial scales, our architecture concatenates all feature maps that have undergone downsampling to attain a uniform size. Subsequently, this concatenated representation is upsampled to yield an input for subsequent decoder stages. This distinctive procedure is formally written as follows:

$$F_t = \mathcal{U}_2(\text{concate}(\mathcal{D}_8(F_t^1), \mathcal{D}_4(F_t^2), \mathcal{D}_2(F_t^3), \tilde{F_t^4})), \qquad (13)$$

where $\mathcal{D}$ and $\mathcal{U}$ represent the down and upsampling operations, respectively, with a subscript denoting the specific sampling factor. In our approach, we implement the downsampling process using convolutional layers with corresponding strides, while the upsampling process is achieved through bilinear interpolation.

After the feature extraction stage, tandem **Decoder** modules, each consisting of a convolutional layer with uniform input and output dimensions, decode the cross-scale feature representation into a predicted optical flow. The decoders are tasked with interpreting the flow information embedded in the feature space and generating a corresponding flow map. Finally, a **Generator** module, comprising a single convolutional layer with two filters, produces the basic predicted flow. In our method, the generator acts as a refinement stage, further enhancing the optical flow prediction generated by the decoders.

To further improve the optical flow prediction, we incorporate a **ConvGRU2** layer (Cho et al., 2014) following the **Generator** module to output the final optical flow. **ConvGRU2** uses the predicted flow from the previous time step as its state and the basic predicted flow from the current time step as its input. This temporal integration of optical flow information enables ST-FlowNet to capture long-range dependencies in the flow sequence. In addition, the output O_t of **ConvGRU2** is then used to augment the state information for **ConvGRU1** at the next time step, providing a more comprehensive representation of the optical flow dynamics.

4.3. Model Training

In this work, we use two types of cross-model transformation methodology, A2S conversion (Bu et al., 2022; Deng and Gu, 2021; Rathi and Roy, 2020) and hybrid bio-information fusion training, to generate an optimal SNN for optical estimation while preserving the original ANN model's accuracy.

4.3.1. ANN-to-SNN conversion

The fundamental process of the A2S conversion encompasses three key phases: first, constructing ANN and SNN models with identical architectures but employing distinct basic neurons; second, training ANN models using the standard BP method; and third, converting the trained ANN model into an SNN model, ensuring extensive retention of model accuracy.

In the ANN model, where information is represented with continuous values, the neuron is defined as follows:

$$a^l = h(w^l \cdot a^{l-1}), \qquad (14)$$

where a^l is the output in the l-th layer, w^l signifies the weight between the $(l-1)$-th and l-th layers, and $h(\cdot)$ denotes the activation function. In the SNN model, we use the LIF model (Abbott, 1999) as the spiking neuron for network construction. Without loss of generality, the iterative form of the membrane potential $v^l(t)$ of the LIF neuron is described as follows (Rathi and Roy, 2020):

$$v^l(t) = e^{-\tau} \cdot v^l(t-1) - s^l(t-1) \cdot \theta^l + w^l \cdot s^{l-1}(t) \cdot \theta^{l-1}. \qquad (15)$$

Here, θ^l is the spiking firing threshold, $\tau \in [0, +\infty)$ represents the membrane potential decay factor, and s^l denotes the spike output generated by:

$$s^l(t) = H(v^l(t) - \theta^l). \qquad (16)$$

Using the Heaviside step function $H(\cdot)$, the LIF spiking neuron emits a spike once the membrane potential surpasses the predetermined firing threshold. To circumvent notable accuracy diminution, we use the quantization clip-floor activation function (Bu et al., 2022) in the ANN model instead of the ReLU function:

$$h(x) = \lambda^l \cdot clip(\frac{1}{L}\left\lfloor \frac{x \cdot L}{\lambda^l} \right\rfloor, 0, 1). \qquad (17)$$

The clip function sets the upper bound to 1 and the lower bound 0. $\lfloor \cdot \rfloor$ denotes the floor function. Prior research has substantiated that the conversion process is theoretically lossless when the hyper-parameter L aligns with the desired time windows T of the SNN, and the trained parameter λ^l corresponds to the spike firing threshold (Bu et al., 2022; Deng and Gu, 2021).

The training process for the optical flow estimation network is conducted using the ANN model. Motivated by the limited availability of ground truth optical flow data, we train our optical flow estimation network through a self-supervised approach (Hagenaars et al., 2021). The comprehensive loss function encompasses two fundamental components: a contrast loss $\mathcal{L}_{\text{contrast}}$ and a smoothness loss $\mathcal{L}_{\text{smooth}}$. The contrast loss $\mathcal{L}_{\text{contrast}}$ uses a reformulated contrast maximization proxy loss to gauge the accuracy of optical flow estimation by assessing the motion compensation performance of an image reconstructed from the predicted optical flow (Hagenaars et al., 2021; Mitrokhin et al., 2018; Zhu et al., 2019). The smoothness

loss $\mathcal{L}_{\text{smooth}}$ uses Charbonneir smoothness function (Charbonnier et al., 1994; Zhu et al., 2018a, 2019) to regulate the optical flow variation between neighboring pixels. Consequently, the total loss is defined as:

$$\mathcal{L}_{\text{flow}} = \mathcal{L}_{\text{contrast}} + \lambda \mathcal{L}_{\text{smooth}}, \tag{18}$$

where the scalar λ balances the respective weights of the contrast loss and the smoothness loss.

Once the ANN model is fully trained, the network weights $\{w^l\}_{l=0}^{L}$ and parameters $\{\lambda^l\}_{l=0}^{L}$ of the ANN are transformed into the weights and spiking fire thresholds $\{\theta^l\}_{l=0}^{L}$ in the SNN (Deng and Gu, 2021; Deng et al., 2020).

4.3.2. Bio-information-fused training

Although the A2S conversion demonstrates superior performance in previous work, there are still some inherent problems that are hard to avoid. The biology parameters, spiking firing threshold and membrane potential decay factor in LIF neuron, significantly influence the spatio-temporal information processing capability of SNN models (Yu et al., 2023b; Sun et al., 2024; Fang et al., 2021). In A2S conversion method, these parameters are usually determined empirically or by threshold balancing strategies, limiting the performance of converted models.

To address these challenges, we propose a hybrid bio-information-fused training strategy (BISNN). This approach incorporates two key operations: (1) Cross-model initialization: The SNN models, which share an identical architecture, are initialized with the pre-trained weights $\{w^l\}_{l=0}^{L}$ of the ANN models. This facilitates efficient knowledge transfer between models; and (2) Parameter-free optimization: A supervised retraining procedure utilizing the STBP (Wu et al., 2018) method is employed to optimize the SNN models, thereby circumventing the need for complex biological parameter filtering processes. According to the chain rule, the mathematical formulation of the loss function's derivatives with respect to the learnable parameters $\{w^l\}_{l=0}^{L}$ can be expressed as follows:

$$\frac{\partial \mathcal{L}_{\text{flow}}}{\partial v^l(t)} = \frac{\partial \mathcal{L}_{\text{flow}}}{\partial s^l(t)} \frac{\partial s^l(t)}{\partial v^l(t)} + \frac{\partial \mathcal{L}_{\text{flow}}}{\partial s^l(t+1)} \frac{\partial s^l(t+1)}{\partial v^l(t)}, \tag{19}$$

$$\frac{\partial \mathcal{L}_{\text{flow}}}{\partial w^l} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_{\text{flow}}}{\partial v^l(t)} \frac{\partial v^l(t)}{\partial w^l}. \tag{20}$$

In this study, the approximate gradient function employed in the backpropagation process is formulated as follows (Fang et al., 2021):

$$H(\cdot) = \frac{\arctan[\pi(\cdot)]}{\pi} + \frac{1}{2}, \tag{21}$$

where arctan represents the inverse tangent function.

5. Experiments

In this section, we first provide a comprehensive overview of our experimental settings, encompassing datasets, evaluation metrics, and implementation details. Next, we present an in-depth performance comparison between ST-FlowNet and other state-of-the-art models on diverse benchmark datasets. For the purpose of visualization, we display representative examples for qualitative illustration. Finally, we conduct a series of ablation studies to demonstrate the significance of the proposed components.

Table 1: Comparison of event-based datasets for optical flow estimation. GT denotes the ground truth datasets.

Datasets	Year	Resolution	Train	Test	GT
UZH-FPV (Zhu et al., 2018b)	2018	346×260	✓	×	×
ECD (Mueggler et al., 2017)	2017	240×180	×	✓	×
MVSEC (Zhu et al., 2018b,a)	2019	346×260	×	✓	✓
HQF (Stoffregen et al., 2020)	2020	240×180	×	✓	×

5.1. Experimental Settings

5.1.1. Optical Flow Datasets

We train the models using the UZH-FPV drone racing dataset (Zhu et al., 2018b), which is distinguished by a diverse distribution of optical flow vectors. We evaluate the models' performance using the Event-Camera Dataset (ECD) (Mueggler et al., 2017), Multi-Vehicle Stereo Event Camera (MVSEC) (Zhu et al., 2018b,a), and High-Quality Frames (HQF) (Stoffregen et al., 2020) dataset, all captured in real-world scenarios using various DAVIS neuromorphic cameras. Tab. 1 shows a detailed overview of the dataset.

5.1.2. Evaluation Metrics

For the MVSEC (Zhu et al., 2018b) optical flow dataset, the ground truth is generated at each APS frame timestamp and scaled to represent the displacement for the duration of one (dt=1) and four (dt=4) APS frames (Zhu et al., 2018a). Consequently, optical flow is also computed at each APS frame timestamp, using all events within the time window as input for dt=1, or 25% of the window events at a time for dt=4. Both predicted optical flows are evaluated using the average endpoint error (AEE_1 for dt=1 and AEE_4 for dt=4).

The ECD and HQF datasets do not include ground truth, and thus we employ two image compensation quality metrics to estimate predicted flows. Specifically, the flow warp loss (FWL) assesses the sharpness of the image of warped events compared with the original event partition, and the variance of the contrast of the event images is reported (T. Stoffregen, 2020). Additionally, we also report the ratio of the squared average timestamps (RSAT), indicating the contrast of $\mathcal{L}_{\text{contrast}}$ between predicted optical flow and baseline null vectors (Hagenaars et al., 2021).

5.1.3. Implementation Details

We implement ST-FlowNet using the PyTorch framework and execute on an NVIDIA A100 GPU. The training process consists of 100 epochs for the ANN model, followed by 10 epochs of retraining in the BISNN method, with a batch size of 8. We use the adaptive moment estimation optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0002, subject

Table 2: Comparison of state-of-the-art models across different datasets. The ST-FlowNet model trained using the ANN, A2S, and BISNN methods are denoted as ST-FlowNet_1, ST-FlowNet_2, and ST-FlowNet_3, respectively. For the MVSEC dataset, the AEE results for each scenario are presented. For the ECD and HQF datasets, the average FWL and RSAT results across all scenarios are reported. The optimal and suboptimal results are highlighted in **bold**. The symbol ↓ (↑) indicates that a smaller (larger) value is preferred.

	Model	OD1		IF1		IF2		IF3		ECD		HQF	
		AEE_1 ↓	AEE_4 ↓	AEE_1 ↓	AEE_4 ↓	AEE_1 ↓	AEE_4 ↓	AEE_1 ↓	AEE_4 ↓	FWL↑	RSAT↓	FWL↑	RSAT↓
ANN	EV-FlowNet (Hagenaars et al., 2021)	**0.32**	**1.30**	0.58	2.18	1.02	3.85	0.87	3.18	1.31	**0.94**	1.37	**0.92**
	RNN-EV-FlowNet (Xu et al., 2025a)	–	1.69	–	**2.02**	–	**3.84**	–	**2.97**	**1.36**	0.95	**1.45**	0.93
	GRU-EV-FlowNet (Hagenaars et al., 2021)	0.47	1.69	0.60	2.16	1.17	3.90	0.93	3.00	–	–	–	–
	GRU-FireNet (Scheerlinck et al., 2020)	0.55	2.04	0.89	3.35	1.62	5.71	1.35	4.68	–	–	–	–
	ET-FlowNet (Tian and Andrade-Cetto, 2022)	**0.39**	1.47	**0.57**	2.08	1.20	3.99	0.95	3.13	–	–	–	–
	STT-FlowNet (Tian and Andrade-Cetto, 2025)	0.66	–	0.57	–	**0.88**	–	**0.73**	–	–	–	–	–
	ST-FlowNet_1 (ours)	0.40	**1.24**	**0.48**	**1.86**	**0.89**	**2.98**	**0.70**	**2.34**	**1.37**	**0.92**	**1.48**	**0.90**
SNN	LIF-EV-FlowNet (Hagenaars et al., 2021)	0.53	2.02	0.71	2.63	1.44	4.93	1.16	3.88	1.21	0.95	1.24	0.94
	XLIF-EV-FlowNet (Hagenaars et al., 2021)	0.45	1.67	0.73	2.72	1.45	4.93	1.17	3.91	1.23	0.95	1.25	0.93
	LIF-FireNet (Hagenaars et al., 2021)	0.57	2.12	0.98	3.72	1.77	6.27	1.50	5.23	1.28	0.99	1.34	1.00
	XLIF-FireNet (Hagenaars et al., 2021)	0.54	2.07	0.98	3.73	1.82	6.51	1.54	5.43	1.29	0.99	1.39	0.99
	Adaptive-SpikeNet (Kosta and Roy, 2023)	0.44	–	0.79	–	1.37	–	1.11	–	–	–	–	–
	SDformerFlow (Tian and Andrade-Cetto, 2025)	0.69	–	0.61	–	0.83	–	**0.76**	–	–	–	–	–
	FSFN (Apolinario and Roy, 2024)	0.50	–	0.76	–	1.19	–	1.00	–	–	–	–	–
	ST-FlowNet_2 (ours)	**0.37**	**1.24**	**0.50**	**1.86**	**0.84**	**2.78**	**0.70**	**2.34**	**1.37**	**0.91**	**1.47**	**0.90**
	ST-FlowNet_3 (ours)	**0.39**	**1.47**	**0.51**	**1.92**	**0.99**	**3.33**	0.77	**2.56**	**1.34**	**0.93**	**1.45**	**0.91**

Table 3: Comparison of models trained or fine-tuned on the MVSEC dataset.

Model	OD1	IF1	IF2	IF3
RNN-FireNet-S-FT (Xu et al., 2025b)	1.97 (AEE_4)	3.24 (AEE_4)	5.48 (AEE_4)	4.45 (AEE_4)
STE-FlowNet (Ding et al., 2022)	0.42 (AEE_1)	0.57 (AEE_1)	0.79 (AEE_1)	0.72 (AEE_1)
U-Net-like SNN (Cuadrado et al., 2023)	–	0.58 (AEE_1)	0.72 (AEE_1)	0.67 (AEE_1)
ST-FlowNet_2 (ours)	0.37 (AEE_1) 1.24 (AEE_4)	0.50 (AEE_1) 1.86 (AEE_4)	0.84 (AEE_1) 2.78 (AEE_4)	0.70 (AEE_1) 2.34 (AEE_4)

to exponential decay. We empirically set the scaling weight for $\mathcal{L}_{\text{smooth}}$ to $\lambda = 0.001$, as used in previous work (Zhu et al., 2018a).

In the SNN models converted from ANN models, the firing thresholds are determined using the threshold balance strategy. To simplify the parameter selection process, we focus exclusively on the membrane potential decay factors in the **Generator** (τ_0) and **ConvGRU2** (τ_1) modules, while setting those in other modules to 0. For consistency and a fair comparison, during the retraining procedure of the BISNN method, both the firing threshold and membrane potential decay factors are initialized with the same values as those used in the A2S method.

5.2. Comparison with State-of-the-art Methods

To validate the effectiveness of our proposed ST-FlowNet, we conduct a comprehensive comparative analysis, assessing its performance compared with other state-of-the-art models from both quantitative and qualitative perspectives (Scheerlinck et al., 2020; Tian and Andrade-Cetto, 2022). The ST-FlowNet model trained using the ANN, A2S, and BISNN methods are denoted as ST-FlowNet_1, ST-FlowNet_2, and ST-FlowNet_3, respectively.

5.2.1. Quantitative Evaluation

We conduct a comparative analysis between ST-FlowNet and other state-of-the-art models on the MVSEC dataset across four representative scenarios: outdoor_day1 (denoted as OD1), indoor_flying1 (IF1), indoor_flying2 (IF2) and indoor_flying3 (IF3), considering both the dt=1 and dt=4 conditions. As shown in Tab. 2, ST-FlowNet demonstrates superior AEE performance compared with other ANN and SNN models across most of scenarios. Notably, the optical flow estimation performance can be improved slightly by leveraging the converted SNN model, as particularly evident in the OD1 and IF2 scenarios. We also provide a comparison between our ST-FlowNet_2 SNN model and other models that are directly trained or fine-tuned on the MVSEC dataset. As shown in Tab. 3, ST-FlowNet_2 still demonstrates competitive performance compared to other models. We believe that training on a larger dataset may further enhance the optical flow estimation performance of our model. For the sake of comparison, on the ECD and HQF dataset, we present a comparative analysis of state-of-the-art models without distinguishing scenarios. As shown in Tab. 2, ST-FlowNet attains the optimal performance, excelling with respect to both the FWL and RSAT metrics.

5.2.2. Qualitative Evaluation

In Fig. 3, we present the visualization results of the predicted optical flows generated by various models. We use EV-FlowNet (ANN model) and LIF-EV-FlowNet (SNN model), which exhibit outstanding performance, for comparison with our ANN and SNN models, respectively. The AEE_1 (black) or FWL (red) values are provided at the upper-left of each predicted optical flow image. Overall, ST-FlowNet demonstrates evident superiority in the visual quality of its optical flows compared with competing models. Specifically, in challenging scenarios such as boundary regions with motion blur or sparse features (i.e. D6 and HTP), both EV-FlowNet and LIF-EV-FlowNet yield prediction failures or errors. ST-FlowNet excels in accurately capturing detailed scene information, thereby achieving reliable optical flow estimation. Additionally, qualitative comparisons highlight that the converted SNN model retains the original ANN model's proficiency in optical flow estimation.

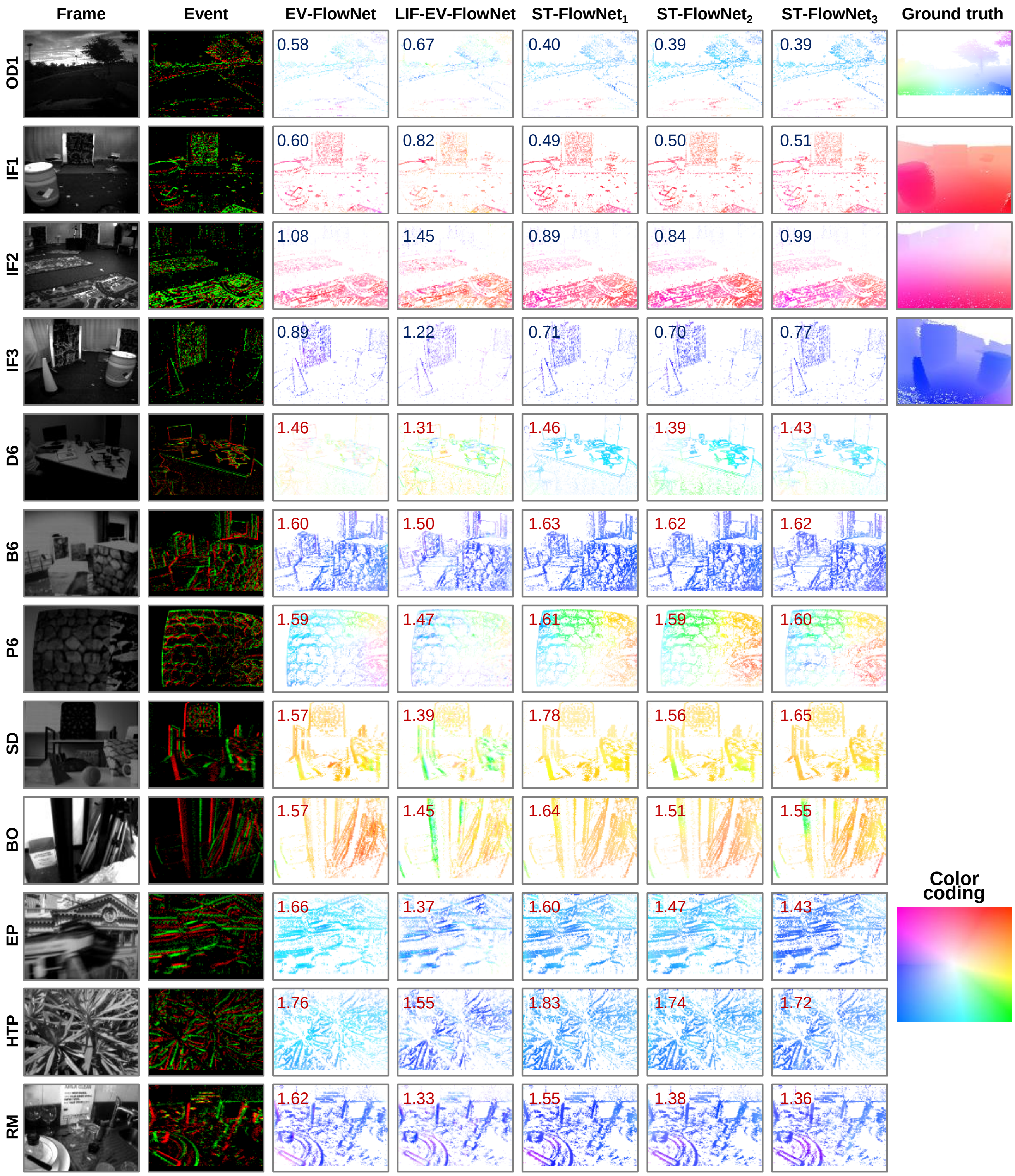

Figure 3: Visual comparison of ST-FlowNet with other models. Original frame-based images, ground truth of the MVSEC dataset, and the color coding of the optical flow are provided for reference. The AEE_1 (black) or FWL (red) results of each predicted optical flow are provided at the upper-left.

5.3. Ablation Analysis

Our novel ST-FlowNet architecture comprises the following principal components: decoders with a non-pyramid architecture (**Decoder**), spatio-temporal optical flow augmentation (**ConvGRU1**), and optical flow alignment (**ConvGRU2**). To demonstrate the effectiveness of these components, we establish a baseline architecture by removing two ConvGRU layers and retaining only one decoder in ST-FlowNet. We conduct relative ablation analyses on the baseline architecture by progressively introducing the relevant components.

Table 4: Performance analysis of the key modules in the ST-FlowNet architecture on the MVSEC dataset. The number of **Decoder** modules is denoted as '#Ds', while the **ConvGRU1** and **ConvGRU2** modules are referred to as 'Aug' and 'Align', respectively. The invalid results is highlighted in *underline*.

Method	#Ds	Aug	Align	OD1 $AEE_1\downarrow$	IF1 $AEE_1\downarrow$	IF2 $AEE_1\downarrow$	IF3 $AEE_1\downarrow$
ANN	4	×	×	0.58	0.60	1.07	0.99
	1	×	×	0.45	0.58	1.02	0.84
	1	✓	×	0.44	0.53	0.99	0.81
	1	✓	✓	**0.40**	**0.48**	**0.89**	**0.70**
A2S	4	×	×	0.56	0.63	1.12	0.94
	1	×	×	0.48	0.62	1.03	0.81
	1	✓	×	0.53	0.54	0.98	0.80
	1	✓	✓	**0.37**	**0.50**	**0.84**	**0.70**
BISNN	4	×	×	0.67	0.82	1.45	1.22
	1	×	×	0.47	0.64	1.13	0.91
	1	✓	×	0.49	0.58	1.05	0.83
	1	✓	✓	**0.39**	**0.51**	**0.99**	**0.77**

Table 5: Performance analysis of the key modules in the ST-FlowNet architecture on the ECD and HQF datasets. The number of **Decoder** modules is denoted as 'Ds', while the **ConvGRU1** and **ConvGRU2** modules are referred to as 'Aug' and 'Align', respectively.

	Method	#Ds	Aug	Align	D6		B6		P6		SD	
					FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓
ECD	ANN	4	×	×	1.29	0.90	1.51	0.92	1.45	0.92	1.40	0.92
		1	×	×	1.45	**0.87**	1.62	0.92	**1.61**	0.92	1.61	0.90
		1	✓	×	1.42	0.88	1.62	**0.91**	1.59	**0.91**	1.73	**0.89**
		1	✓	✓	**1.46**	**0.87**	**1.63**	0.92	**1.61**	**0.91**	**1.77**	**0.89**
	A2S	4	×	×	1.24	0.93	1.44	0.94	1.39	0.94	1.30	0.93
		1	×	×	1.40	**0.88**	1.60	**0.92**	1.57	0.92	1.51	0.91
		1	✓	×	1.40	0.89	1.61	**0.92**	1.59	**0.91**	1.56	0.91
		1	✓	✓	**1.44**	**0.88**	**1.63**	**0.92**	**1.62**	**0.91**	**1.62**	**0.90**
	BISNN	4	×	×	1.31	0.90	1.50	0.93	1.47	0.93	1.39	**0.89**
		1	×	×	1.44	**0.88**	1.59	0.93	1.57	0.93	1.74	0.91
		1	✓	×	1.44	**0.88**	1.61	**0.91**	1.59	**0.91**	**1.83**	**0.89**
		1	✓	✓	**1.45**	**0.88**	**1.62**	0.92	**1.61**	**0.91**	1.66	**0.89**
	Method	#Ds	Aug	Align	BO		EP		HTP		RM	
					FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓
HQF	ANN	4	×	×	1.37	0.91	1.29	0.93	1.49	0.90	1.24	0.94
		1	×	×	1.57	**0.90**	1.57	**0.90**	1.79	0.89	1.46	0.91
		1	✓	×	**1.65**	**0.90**	1.58	0.91	1.81	**0.88**	1.52	0.91
		1	✓	✓	**1.65**	**0.90**	**1.60**	0.91	**1.82**	**0.88**	**1.54**	**0.90**
	A2S	4	×	×	1.30	0.93	1.24	0.95	1.37	0.92	1.20	0.95
		1	×	×	1.46	0.91	1.46	**0.91**	1.64	**0.89**	1.35	0.92
		1	✓	×	1.58	0.91	1.58	**0.91**	1.79	**0.89**	1.48	0.91
		1	✓	✓	**1.60**	**0.90**	**1.59**	**0.91**	**1.79**	0.90	**1.49**	**0.90**
	BISNN	4	×	×	1.45	**0.89**	1.37	**0.91**	1.55	**0.88**	1.33	0.92
		1	×	×	1.56	0.91	1.57	0.92	1.67	0.91	1.52	0.92
		1	✓	×	1.56	0.91	**1.59**	**0.91**	1.71	0.89	**1.53**	**0.91**
		1	✓	✓	**1.57**	0.90	1.56	**0.91**	**1.77**	**0.88**	1.49	**0.91**

5.3.1. Analysis of the Number of Decoders

We compare the baseline models with varying numbers of decoders to examine the effectiveness of the decoders. Specifically, we increase the number of decoders from one in the baseline model to four in the comparison models. To conduct a thorough comparison, we select four representative scenarios from the ECD and HQF datasets, each denoted as follows: dynamic_6dof (D6), boxes_6dof (B6), poster_6dof (P6), slider_depth (SD), along with boxes (BO), engineering_posters (EP), high_texture_plants (HTP), and reflective_materials (RM). As shown in Tabs. 4 and 5, the performance decreases across all scenarios when more decoders are employed. This suggests that the use of more decoders is not necessarily associated with improved performance. This observation not only alleviates the pressure to increase the model size but also reinforces the advantage of the SNN model in terms of energy consumption.

5.3.2. Analysis of the Spatio-temporal Augmentation

Building on the baseline model with a single decoder, we integrate **ConvGRU1** as the spatio-temporal optical flow augmentation component. Unlike the complete ST-FlowNet, the **ConvGRU1** layer receives the upsampled basic predicted optical flow as state information. The detailed results are shown in Tabs. 4 and 5. The AEE results for the MVSEC dataset demonstrate that the ConvGRU augmentation layer enhances optical flow estimation performance in most scenarios, except for OD1. The FWL and RSAT results computed for the ECD and HQF datasets also generally corroborate the improvement, which is more pronounced for the HQF dataset. Furthermore, the ConvGRU layers assist in mitigating the performance loss of converted SNN models. In certain scenarios, SNN models exhibit superior performance compared with ANN models, as observed for the IF2 and IF3 scenarios.

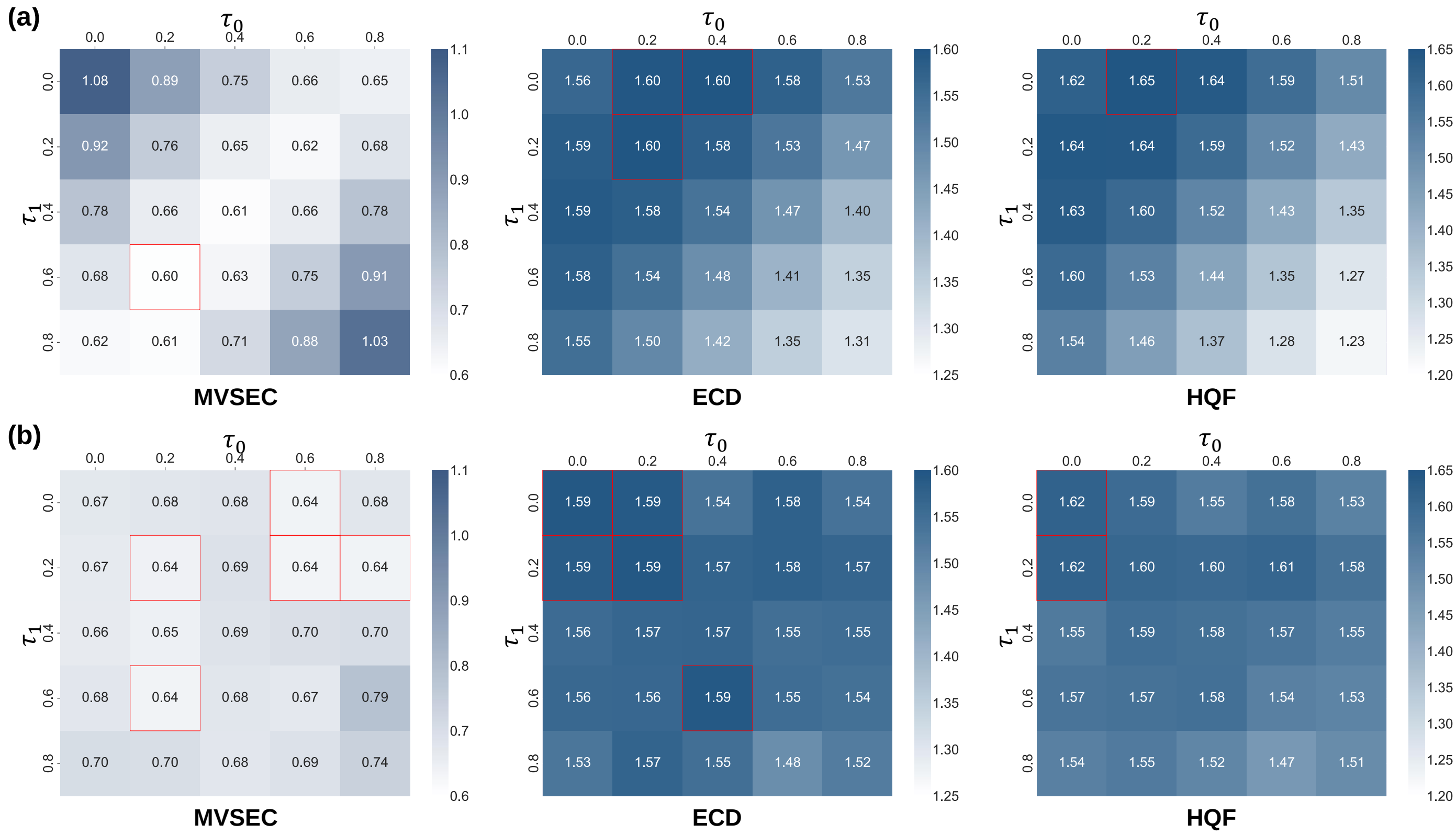

Figure 4: The performance comparison of SNN models initialized with different combinations of membrane potential decay factors. (a) SNN models trained using the A2S method. (b) SNN models trained using the BISNN method. The optimal results are highlighted within red boxes.

Table 6: Performance analysis of SNNs trained with different method.

	Method	OD1 AEE$_1$↓		IF1 AEE$_1$↓		IF2 AEE$_1$↓		IF3 AEE$_1$↓	
MVSEC	STBP	0.78		1.00		1.78		1.57	
	A2S	0.37		0.50		0.84		0.70	
	BISNN	**0.39**		**0.51**		**0.99**		**0.77**	
		D6		B6		P6		SD	
ECD		FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓
	STBP	1.18	0.95	1.31	0.96	1.30	0.96	1.46	0.92
	A2S	1.44	**0.88**	**1.63**	**0.92**	**1.62**	**0.91**	1.62	0.90
	BISNN	**1.45**	**0.88**	1.62	**0.92**	1.61	**0.91**	**1.66**	**0.89**
		BO		EP		HTP		RM	
HQF		FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓	FWL↑	RSAT↓
	STBP	1.34	0.93	1.16	0.97	1.44	0.93	1.17	0.96
	A2S	**1.60**	**0.90**	**1.59**	**0.91**	**1.79**	0.90	**1.49**	**0.90**
	BISNN	1.57	**0.90**	1.56	**0.91**	1.77	**0.88**	**1.49**	0.91

5.3.3. Analysis of the Spatio-temporal Alignment

Expanding on the previous model, we continue to introduce **ConvGRU2** as a spatio-temporal alignment module. This module is designed to project the basic and historical predicted optical flow into a standardized state space. By doing so, it not only enhances the precision of optical flow prediction but also establishes an effective reference state for input data. As illustrated in Tabs. 4 and 5, the optical flow estimation performance further improves in most scenarios. Moreover, in comparison to the spatio-temporal augmentation module, the alignment module demonstrates superior effectiveness.

5.4. Analysis of Training Methods

To thoroughly assess the efficacy of various training approaches, we also conduct experiments involving the direct training of SNNs using the STBP method. The initialization operations of biological parameters are consistent with those in the BISNN method. As presented in Tab. 6, the performance of models trained using the A2S and BISNN methods significantly outperforms those trained with the STBP method. These findings suggest that direct training of SNNs for optical flow estimation remains a challenging task, and further highlight the competitive performance improvements achieved by both the A2S and BISNN methods.

The performance of SNN models converted using the A2S method is influenced by the selection of biological parameters. In our experiments, the spike firing thresholds are determined using the threshold balance strategy, while the membrane potential decay factors are chosen empirically. To further verify the impact of parameter settings on model performance, we perform a search over the parameters τ_0 and τ_1 within the $[0, 0.8]$ range. Experiments are conducted on models trained using both the A2S and BISNN methods under various parameter combinations, and the average results across all scenarios are presented in Fig. 4. The optimal results for each set of experiments are highlighted within red boxes.

As shown in Fig. 4(a), the performance of the models fluctuates significantly with changes in the membrane potential decay factors. On the MVSEC dataset, parameter combinations along

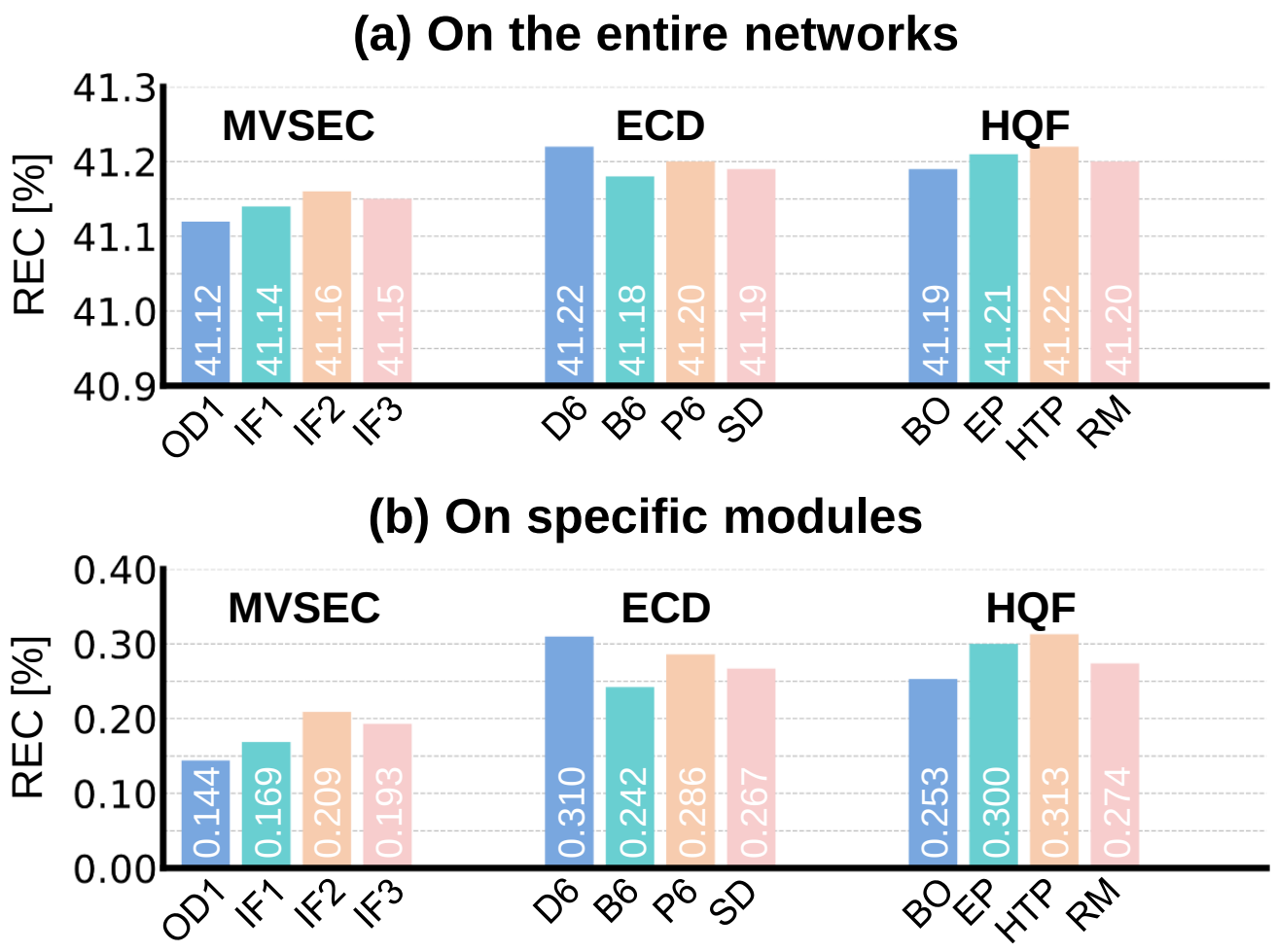

Figure 5: The energy consumption of SNN ST-FlowNet models relative to ANN models.

the secondary diagonal tend to yield the best results. However, for the ECD and HQF datasets, smaller membrane potential decay factor combinations appear to be more effective. Overall, we are still unable to summarize a general principle for parameter selection from this sufficiently large search space. These findings underscore the challenges associated with selecting optimal biological parameters. Nevertheless, the performance of models trained using the BISNN method demonstrates notable robustness across varying parameter combinations. These results substantiate our claim that BISNN operates as a parameter-free training method, effectively alleviating the complexities of parameter tuning while consistently delivering robust optical flow estimation performance.

5.5. Analysis of Energy Consumption

Finally, we assess the energy consumption for different ST-FlowNet models. In contrast to the ANN ST-FlowNet model that mainly employs multiply-accumulate operations, the SNN ST-FlowNet model predominantly leverages sparse accumulate (AC) operations as computational units, but excluding the **ConvGRU1/2**, **Encoder1** and **Decoder1** modules. The relative energy consumption (REC), denoted as $\eta = \frac{\Phi_{\text{SNN}}}{\Phi_{\text{ANN}}}$, is employed to evaluate the energy-saving advantages of SNN ST-FlowNet models. Here, Φ represents the theoretical energy consumption as calculated in previous work (Yao et al., 2023).

We conduct experiments on ST-FlowNet models trained using the ANN and A2S methods, respectively, and the results are summarized in Fig. 5. As expected, the energy consumption reduction achieved by the SNN ST-FlowNet model is approximately 60% across the entire networks (Fig. 5 (a)). If we focus only on specific modules that exclusively use AC operations in the SNN ST-FlowNet model, a remarkable reduction (more than 300-fold) in energy consumption can be observed (Fig. 5 (b)). These observations clearly demonstrate the advantage of energy consumption in SNN ST-FlowNet models.

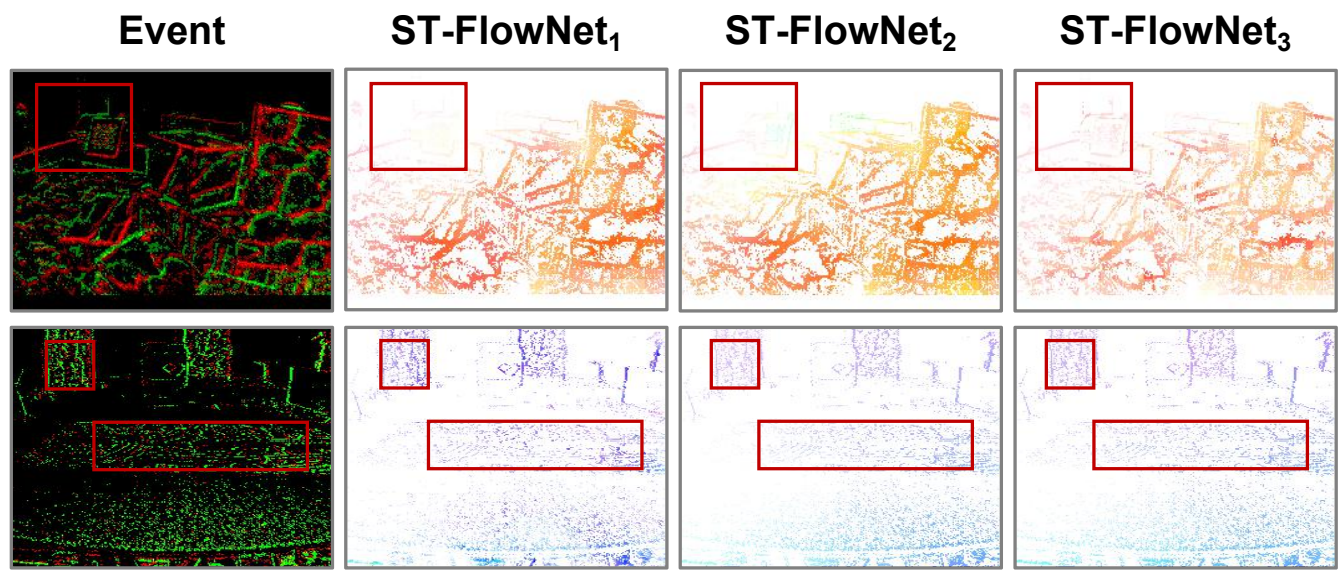

Figure 6: Representative examples of failure cases from our ST-FlowNet models. The regions with failed optical flow estimations are highlighted with red boxes.

5.6. Analysis of Failure Cases

Although our proposed method demonstrates promising performance in optical flow estimation, several challenges remain unresolved. Two representative failure cases are illustrated in Fig. 6. In the first row, the ST-FlowNet models fail to accurately capture a small object, as highlighted by the red box. This observation is consistent with the visualization results in Fig. 3, which indicate a relatively degraded performance in scenarios containing numerous small-scale objects (i.e. D6 and RM). In addition, distinguishing between background noise and high-density event streams presents another significant challenge, as shown in the second row of Fig. 6. This issue, common in event-based vision tasks, may potentially be addressed through the development of more suitable event representation strategies.

6. Discussion and Conclusion

In this study, we present a novel neural network architecture, termed ST-FlowNet, which incorporates enhanced ConvGRU layers, designed to align and augment spatio-temporal information, thereby improving optical flow estimation based on event-driven inputs. To address the challenges in training SNN models for optical flow estimation, we propose two methods: (1) the A2S method, which generates an SNN model from a pre-trained ANN model, and (2) a novel BISNN strategy, aimed at mitigating the complexities associated with the selection of biological parameters.

Overall, our work demonstrates a notable level of superiority. (1) Our experimental results across a variety of representative scenarios validate the effectiveness of the proposed ST-FlowNet, which outperforms current state-of-the-art optical flow estimation models. In particular, for the challenging boundary regions characterized by motion blur or sparse features (commonly introduced during event representation), ST-FlowNet exhibits a robust capability for capturing scene information. (2) Ablation studies further highlight the critical roles of the integrated **ConvGRU** modules in spatio-temporal augmentation and alignment, establishing a promising tool for future model design. Meanwhile, increasing the number of decoder layers is shown to be non-essential in semi-pyramidal architectural models, thereby contributing to a lightweight and

energy-efficient design. (3) A comparative analysis of three training paradigms demonstrates that indirect A2S conversion and hybrid BISNN methods can generate superior SNN models compared to direct STBP training. Notably, the BISNN method exhibits greater robustness to variations in initialized biological parameters, thereby alleviating the complexity of biological parameter selection.

The limitations and shortcomings of the ST-FlowNet model are discussed below. First, we observe performance degradation in the converted SNN models under specific conditions (e.g., IF1 in Tab. 2 and HTP in Tab. 5), which may be attributed to theoretical conversion errors inherent in the A2S process (Deng and Gu, 2021). This issue could be addressed through the development of more efficient conversion methods. Second, the spatio-temporal augmentation module exhibits a less robust ability to improve performance, potentially due to modality bias between the input event data and the optical flow state. A well-designed **ConvGRU1** module may further improve the performance of the ST-FlowNet model by reducing modality bias, deserving to be examined in our further studies. Third, analysis of the failure cases reveals that the ST-FlowNet models exhibit limited sensitivity to small objects, background noise, and high-intensity event streams. Addressing these challenges represents a potential avenue for further enhancing model performance.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under grant 2023YFF1204200, in part by the STI 2030–Major Projects under grant 2022ZD0208500, in part by the Sichuan Science and Technology Program under grant 2024NSFJQ0004 and grant DQ202410, and in part by the Natural Science Foundation of Chongqing, China under grant CSTB2024NSCQ-MSX0627, in part by the Science and Technology Research Program of Chongqing Education Commission of China under Grant KJZD-K202401603, and in part by the China Postdoctoral Science Foundation under grant 2024M763876.

References

Abbott, L. F., 1999. Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain Research Bulletin 50 (5-6), 303–304.

Apolinario, M. P. E., Roy, K., 2024. S-TLLR: STDP-inspired temporal local learning rule for spiking neural networks. Transactions on Machine Learning Research.

Ballas, N., Yao, L., Pal, C., Courville, A., 2016. Delving deeper into convolutional networks for learning video representations. In: 4th International Conference on Learning Representations.

Benosman, R., Clercq, C., Lagorce, X., Ieng, S.-H., Bartolozzi, C., 2013. Event-based visual flow. IEEE Transactions on Neural Networks and Learning Systems 25 (2), 407–417.

Bu, T., Fang, W., Ding, J., DAI, P., Yu, Z., Huang, T., 2022. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. In: International Conference on Learning Representations.

Cai, W., Sun, H., Liu, R., Cui, Y., Wang, J., Xia, Y., Yao, D., Guo, D., 2024. A spatial–channel–temporal-fused attention for spiking neural networks. IEEE Transactions on Neural Networks and Learning Systems 35 (10), 14315–14329.

Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M., 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st International Conference on Image Processing. Vol. 2. pp. 168–172.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Cuadrado, J., Rançon, U., Cottereau, B. R., Barranco, F., Masquelier, T., 2023. Optical flow estimation from event-based cameras and spiking neural networks. Frontiers in Neuroscience 17, 1160034.

Deng, L., Wu, Y., Hu, X., Liang, L., Ding, Y., Li, G., Zhao, G., Li, P., Xie, Y., 2020. Rethinking the performance comparison between SNNs and ANNs. Neural Networks 121, 294–307.

Deng, S., Gu, S., 2021. Optimal conversion of conventional artificial neural networks to spiking neural networks. In: International Conference on Learning Representations.

Diehl, P. U., Cook, M., 2015. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Frontiers in Computational Neuroscience 9, 99.

Ding, Z., Zhao, R., Zhang, J., Gao, T., Xiong, R., Yu, Z., Huang, T., 2022. Spatio-temporal recurrent networks for event-based optical flow estimation. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 36. pp. 525–533.

Fan, L., Zhang, T., Du, W., 2021. Optical-flow-based framework to boost video object detection performance with object enhancement. Expert Systems with Applications 170, 114544.

Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., Tian, Y., 2021. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2661–2671.

Gaur, V., 2022. Lucas-kanade optical flow machine learning implementations. Journal of Student Research 11 (3).

Hagenaars, J., Paredes-Vallés, F., De Croon, G., 2021. Self-supervised learning of event-based optical flow with spiking neural networks. Advances in Neural Information Processing Systems 34, 7167–7179.

Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xu, C., Huang, T., Shi, B., 2020. Neuromorphic camera guided high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1730–1739.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2462–2470.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations.

Kosta, A. K., Roy, K., 2023. Adaptive-SpikeNet: Event-based optical flow estimation using spiking neural networks with learnable neuronal dynamics. In: 2023 IEEE International Conference on Robotics and Automation. pp. 6021–6027.

Maass, W., 1997. Networks of spiking neurons: The third generation of neural network models. Neural Networks 10 (9), 1659–1671.

Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y., 2018. Event-based moving object detection and tracking. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1–9.

Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D., 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. The International Journal of Robotics Research 36 (2), 142–149.

Paredes-Vallés, F., Scheper, K. Y., De Croon, G. C., 2019. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. IEEE transactions on Pattern Analysis and Machine Intelligence 42 (8), 2051–2064.

Rathi, N., Roy, K., 2020. Diet-SNN: Direct input encoding with leakage and threshold optimization in deep spiking neural networks. International Conference on Learning Representations.

Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D., 2020. Fast image reconstruction with an event camera. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 156–163.

Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N.,

Kleeman, L., Mahony, R., 2020. Reducing the sim-to-real gap for event cameras. In: 16th European Conference on Computer Vision. pp. 534–549.
Sun, H., Cai, W., Yang, B., Cui, Y., Xia, Y., Yao, D., Guo, D., 2024. A synapse-threshold synergistic learning approach for spiking neural networks. IEEE Transactions on Cognitive and Developmental Systems 16 (2), 544–558.
T. Stoffregen, C. Scheerlinck, D. S. T. D. N. B. L. K. R. M., 2020. Reducing the sim-to-real gap for event cameras. In: 16th European Conference on Computer Vision.
Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., Maida, A., 2019. Deep learning in spiking neural networks. Neural Network 111, 47–63.
Tian, Y., Andrade-Cetto, J., 2022. Event transformer FlowNet for optical flow estimation. In: British Machine Vision Conference.
Tian, Y., Andrade-Cetto, J., 2025. SDformerFlow: Spiking neural network transformer for event-based optical flow. In: Pattern Recognition. pp. 475–491.
Ussa, A., Rajen, C. S., Pulluri, T., Singla, D., Acharya, J., Chuanrong, G. F., Basu, A., Ramesh, B., 2024. A hybrid neuromorphic object tracking and classification framework for real-time systems. IEEE Transactions on Neural Networks and Learning Systems 35 (8), 10726–10735.
Vandal, T. J., Nemani, R. R., 2023. Temporal interpolation of geostationary satellite imagery with optical flow. IEEE Transactions on Neural Networks and Learning Systems 34 (7), 3245–3254.
Wang, L., Guo, Y., Liu, L., Lin, Z., Deng, X., An, W., 2020. Deep video super-resolution using HR optical flow estimation. IEEE Transactions on Image Processing 29, 4323–4336.
Wu, Y., Deng, L., Li, G., Zhu, J., Shi, L., 2018. Spatio-temporal backpropagation for training high-performance spiking neural networks. Frontiers in Neuroscience 12, 331.
Xu, Y., Tang, G., Yousefzadeh, A., de Croon, G. C., Sifalakis, M., 2025a. Event-based optical flow on neuromorphic processor: Ann vs. snn comparison based on activation sparsification. Neural Networks 188, 107447.
Xu, Y., Tang, G., Yousefzadeh, A., de Croon, G. C., Sifalakis, M., 2025b. Event-based optical flow on neuromorphic processor: ANN vs. SNN comparison based on activation sparsification. Neural Networks, 107447.
Yao, M., Zhang, H., Zhao, G., Zhang, X., Wang, D., Cao, G., Li, G., 2023. Sparser spiking activity can be better: Feature refine-and-mask spiking neural network for event-based visual recognition. Neural Networks 166, 410–423.
Yu, F., Wu, Y., Ma, S., Xu, M., Li, H., Qu, H., Song, C., Wang, T., Zhao, R., Shi, L., 2023a. Brain-inspired multimodal hybrid neural network for robot place recognition. Science Robotics 8 (78), eabm6996.
Yu, Q., Gao, J., Wei, J., Li, J., Tan, K. C., Huang, T., 2023b. Improving multispike learning with plastic synaptic delays. IEEE Transactions on Neural Networks and Learning Systems 34 (12), 10254–10265.
Zhai, M., Xiang, X., Lv, N., Kong, X., 2021. Optical flow and scene flow estimation: A survey. Pattern Recognition 114, 107861.
Zhang, Y., Lv, H., Zhao, Y., Feng, Y., Liu, H., Bi, G., 2023. Event-based optical flow estimation with spatio-temporal backpropagation trained spiking neural network. Micromachines 14 (1), 203.
Zheng, H., Wu, Y., Deng, L., Hu, Y., Li, G., 2021. Going deeper with directly-trained larger spiking neural networks. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 35. pp. 11062–11070.
Zheng, H., Zheng, Z., Hu, R., Xiao, B., Wu, Y., Yu, F., Liu, X., Li, G., Deng, L., 2024. Temporal dendritic heterogeneity incorporated with spiking neural networks for learning multi-timescale dynamics. Nature Communications 15 (1), 277.
Zhu, A., Yuan, L., Chaney, K., Daniilidis, K., June 2018a. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In: Proceedings of Robotics: Science and Systems.
Zhu, A. Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K., 2018b. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters 3 (3), 2032–2039.
Zhu, A. Z., Yuan, L., Chaney, K., Daniilidis, K., 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997.
Zitnick, C., Jojic, N., Kang, S. B., 2005. Consistent segmentation for optical flow estimation. In: 8th IEEE International Conference on Computer Vision. Vol. 2. pp. 1308–1315.