VOLEX-Fusion: Reliability-Aware Gated Fusion for Robust Event-Based Monocular Depth Estimation

Şebnem Sarıözkan, Alper Yegenoglu, and Erdal Kayacan

***Abstract*— We introduce VOLEX-Fusion, a monocular depth estimation framework designed to maximize the utility of sparse event data through an adaptive gated fusion with RGB frames. The core innovation lies in the network's ability to act as a reliability-aware arbitrator for depth inference; in scenarios where RGB quality degrades—due to severe motion blur or low-light—the fusion mechanism intelligently shifts its priority to the high-temporal event stream to maintain accurate depth consistency. By formulating depth estimation as a sensor-aware weighting problem, VOLEX-Fusion effectively bridges the gap between dense photometric textures and sparse event signals, ensuring robust topographic reconstruction even when the primary visual modality becomes unreliable.**

***Index Terms*— Event-Based Vision, Deep Learning for Visual Perception, Perception in Aerial Systems.**

I. Introduction

This work presents **VOLEX-Fusion**, a novel framework for reliability-aware gated fusion of RGB and event data. Unlike prior hybrid approaches that rely on simple concatenation or fixed-weight fusion, VOLEX-Fusion acts as an **intelligent arbitrator**, dynamically weighting features based on instantaneous reliability. During motion blur or illumination dropouts, the system prioritizes the high-temporal event stream to preserve topographic consistency. IMU-stabilized motion compensation and backward-compensated rectification ensure spatially aligned, artifact-free event voxels.

The method is validated on public benchmarks and real-world scenarios. Results show that VOLEX-Fusion significantly outperforms conventional baselines, maintaining robust depth estimation even when RGB input is severely degraded.

The contributions of this paper are as follows:

- A novel **reliability-aware gated fusion** architecture that dynamically weights RGB and event features for depth inference under varying visual qualities.
- An **IMU-guided motion compensation** scheme integrated with a backward-compensated VOLEX representation to generate sharp and spatially consistent multimodal inputs.
- An adaptive sensor arbitration mechanism enabling reliable depth estimation during motion blur and low-light degradation.
- Superior depth accuracy and consistency over state-of-the-art monocular baselines in challenging high-speed and low-illumination aerial scenarios.

*This work was supported by the Horizon Europe Grant Agreement No. 101136056.

Ş. Sarıözkan, A. Yegenoglu, and E. Kayacan are with the Automatic Control Group (RAT), Paderborn University, 33098 Paderborn, Germany (e-mail: sariozka@mail.uni-paderborn.de, alper.yegenoglu@uni-paderborn.de, and erdal.kayacan@uni-paderborn.de).

II. Methodology

The architecture of **Volex-Fusion** is designed to perform monocular depth estimation by synergizing asynchronous event streams with conventional RGB frames. As shown in Fig. 1, our approach moves away from rigid fusion by employing a dynamic gating strategy that prioritizes the most reliable modality under challenging conditions.

A. Event presentation and motion correction

To effectively process asynchronous event streams and standard RGB frames, our method integrates geometric motion correction with an adaptive fusion strategy. We incorporate IMU-based synchronization to mitigate motion blur and temporal smearing within the event data. By utilizing gyroscope measurements, the raw events are spatially warped to a common reference timestamp based on the camera's rotational movement. These motion-compensated events are then aggregated into a spatially consistent grid, ensuring that structural and geometric information remains sharp even during rapid sensor maneuvers.

B. Fusion gate

Following this pre-processing, the extracted features from both the event stream and the RGB images are fed into the **Dynamic Fusion Module (DFM)**. Rather than treating both modalities equally through standard concatenation, the DFM evaluates them as competing signals. A learnable gating mechanism computes a pixel-wise reliability mask based on the current context. This spatial mask dynamically adjusts the weight of each modality during fusion. Consequently, the network autonomously suppresses degraded RGB features—such as those suffering from high-speed motion blur or low illumination—and prioritizes the robust, high-temporal-resolution event stream to maintain accurate depth inference.

C. System architecture

Our architecture employs a recurrent encoder-decoder topology to generate dense depth maps from the dynamically fused features. The encoder's hierarchical spatial representations are progressively upsampled by the decoder via skip connections, preserving fine-grained geometric details. To ensure temporal consistency and mitigate flickering artifacts, a **ConvGRU** unit is embedded at the bottleneck.

The decoder regresses the final depth map in the logarithmic space to effectively handle the wide dynamic range of

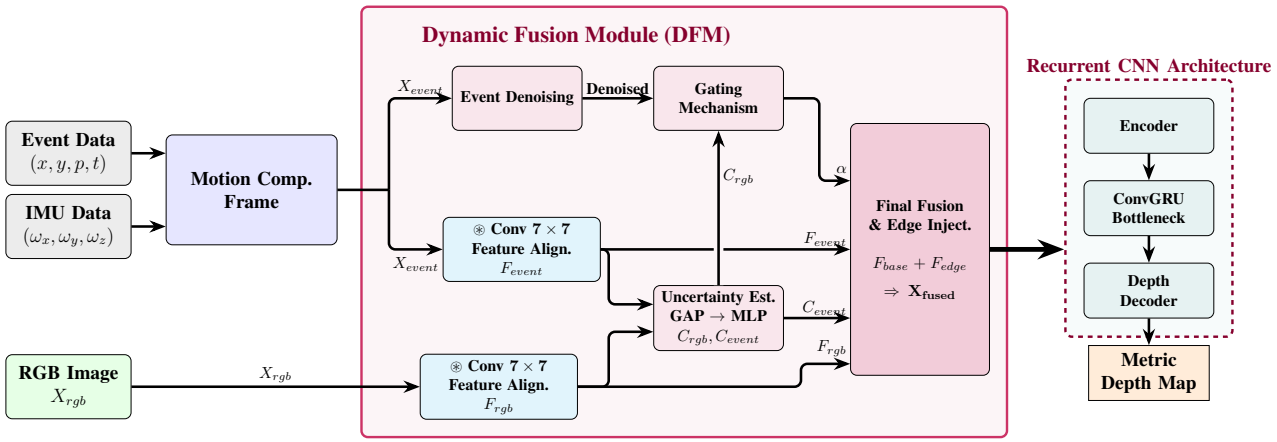

Fig. 1. Detailed architecture of the VOLEX-Fusion pipeline. Raw inputs are preprocessed into motion-compensated frames and aligned using 7×7 convolutions. The Dynamic Fusion Module calculates sensor uncertainty (C_{rgb}, C_{event}) and extracts geometric boundaries to generate an adaptive weighting map (α). The fully fused features are then processed by a vertically aligned recurrent CNN framework comprising an encoder, a ConvGRU bottleneck for temporal consistency, and a decoder to regress the final metric depth map.

complex environments. The network is trained using a scale-invariant logarithmic (SiLog) loss function. By penalizing relative errors rather than absolute metric differences, this formulation encourages the model to accurately capture the global structural layout of the scene, yielding high-quality and temporally smooth depth estimations.

III. Experiments

We trained our proposed fusion framework on the TartanAir [1] dataset (excluding the Abandoned Factory Night sequence). Evaluation was performed on two challenging edge-case test sets: the MVSEC [2] indoor flying sequences for real-world high-speed dynamics, and the unseen TartanAirV2 Abandoned Factory (Night/Easy) for synthetic, extreme low-light conditions. These benchmarks effectively assess the framework's structural integrity under severe motion blur and near-zero illumination.

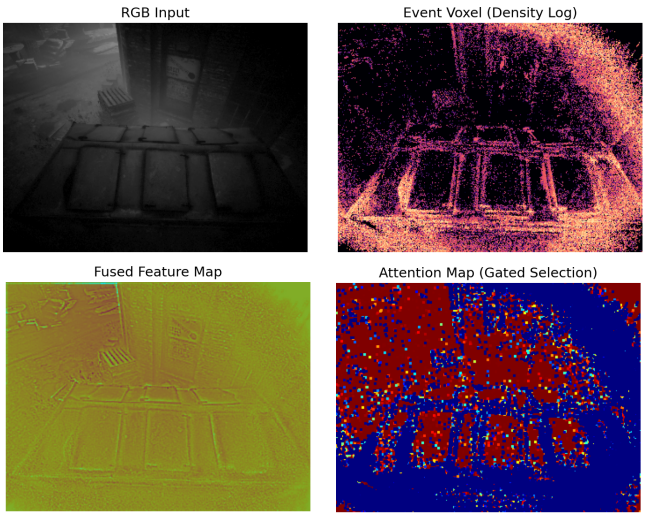

Fig. 2. The TartanAir night views. The Fused Feature Map shows the integration of RGB and event data, while the Attention Map visualizes the dynamic gating mechanism.

The quantitative performance of our framework is summarized in Table I. In real-world indoor scenarios, the model achieves high precision, notably reaching an RMSE of **0.7481m** in the MVSEC Indoor Flying 2 sequence. The consistency across both Indoor 2 and 3 sequences suggests that our gated attention mechanism effectively prioritizes the event stream when rapid camera maneuvers induce blur in the RGB frames.

Furthermore, our evaluation on the TartanAirV2 nighttime sequences demonstrates the robustness of the fusion strategy in degraded environments. Despite the challenging 20m depth range and lack of active lighting, the model achieves an absolute relative of **0.1763**. This suggests that the adaptive gating successfully leverages event voxels to "fill in" the spatial gaps where the RGB signal is unreliable. As shown in Fig. 2, the attention map dynamically controls the sensor fusion process. Specifically, blue regions indicate a higher reliance on event data, whereas red regions correspond to the RGB features.

TABLE I

Depth evaluation results of our proposed method on MVSEC datasets (Median Scaling applied, Range: 0.1m - 5.0m for Indoor, 0.1m - 20.0m for Night).

Metric	Indoor Flying 2	Indoor Flying 3	Night (Easy)
AbsRel ↓	0.2124	0.2062	0.1763
SqRel ↓	0.2179	0.2552	1.8269
RMSE (m) ↓	0.7481	0.8662	2.6711
RMSElog ↓	0.2654	0.2559	0.2263
SiLog ×100 ↓	25.517	24.435	21.724
$\delta < 1.25$ (↑)	0.6490	0.6702	0.7885
$\delta < 1.25^2$ (↑)	0.9081	0.9153	0.9483
$\delta < 1.25^3$ (↑)	0.9759	0.9793	0.9791

A. Bridging the gap to outdoor scenarios

Currently, our algorithm is optimized for short-to-medium range depth estimation and has not yet been fully adapted

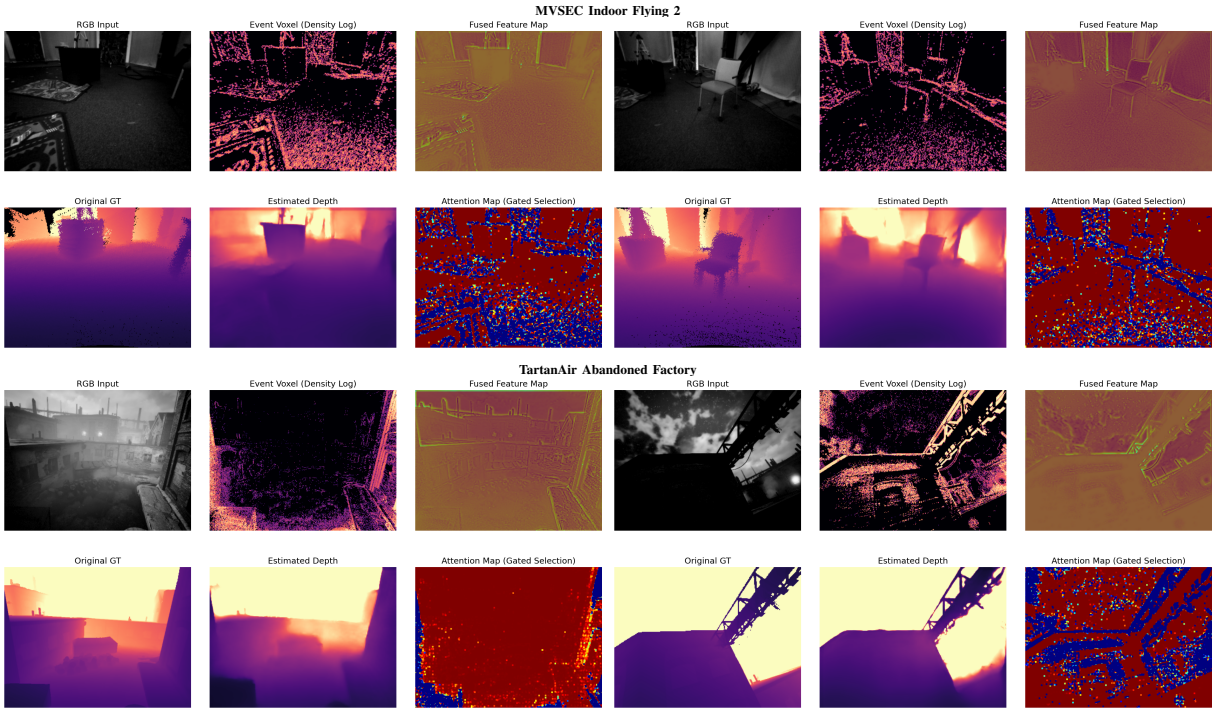

Fig. 3. Qualitative results for MVSEC Indoor and TartanAir abandoned factory night sequences.

for the large-scale spatial structures of real-world outdoor sequences (e.g., MVSEC Outdoor Day). However, our evaluations on the TartanAirV2 Abandoned Factory (Night/Easy) dataset provide a promising proxy for outdoor performance. This sequence presents challenging low-light conditions and longer depth ranges (up to 20m), serving as a precursor to real-world night driving scenarios.

Our model achieved a relative error of 0.1763 and a $\delta < 1.25$ accuracy of 78.8% on this dataset. These results are promising, especially when considering the inherent difficulty of nocturnal depth estimation where RGB signals are near-zero.

B. Algorithmic and causal comparison with SOTA

When compared with state-of-the-art (SOTA) methods such as FUSE [3], SRFNet [4], and UniCTDepth [5], certain causal advantages of our architecture emerge:

Unlike methods relying on rigid cross-modal consistency or complex kernels, our adaptive gated selection mechanism effectively suppresses noise propagation from unreliable sensor streams while maintaining sharp structural boundaries with minimal computational latency.

IV. Conclusion and Future Work

In this study, we present an adaptive gated fusion framework that effectively combines event-based data with traditional RGB frames for robust depth estimation. Our approach successfully mitigates common vision challenges such as motion blur in high-speed indoor maneuvers and visual degradation in low-light environments. Experimental results demonstrate that the proposed gated mechanism provides high-fidelity depth maps that maintain structural integrity across diverse scenarios.

While our current results are promising, model refinements are still ongoing to further enhance the spatial resolution and long-range accuracy of the estimation. Our goal is to deploy this framework as a core perception layer in autonomous systems. Future work will focus on integrating our depth estimation model into depth-based navigation pipelines, including both classical path-planning and Reinforcement Learning (RL) based control strategies. Additionally, we plan to investigate the synergy between our fusion framework and SLAM (Simultaneous Localization and Mapping) algorithms to improve tracking robustness in visually degraded conditions.

References

[1] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "TartanAir: A dataset to push the limits of visual SLAM," 2020.

[2] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, July 2018.

[3] P. Sun, J. Jiang, Y. Yao, Y. Chen, W. Zhao, K. Jiang, and X. Liu, "FUSE: Label-free image-event joint monocular depth estimation via frequency-decoupled alignment and degradation-robust fusion," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[4] T. Pan, Z. Cao, and L. Wang, "SRFNet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024.

[5] L. Jing, D. Shi, Z. Liu, S. Jin, C. Qiu, Z. Qiao, Y. Li, and J. Xia, "UniCT Depth: Event-image fusion based monocular depth estimation with convolution-compensated vit dual sa block," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 8 2025, main Track.