photonics

MDPI

Article

Dark-Scene Neuromorphic Imaging for Human Pose Estimation

Hezu Bai [1], Xulei Qin [1,2,*], Yayu Dai [3], Zhou Ge [4,5,*], Chengze Li [4,5], Chaofan Wu [4,5], Feng Shi [2], Hongchang Cheng [2], Haibo Fan [2] and Shilong Li [2]

[1] School of Physics, Changchun University of Science and Technology, Changchun 130015, China
[2] Science and Technology on Low-Light-Level Night Vision Laboratory, Xi'an 710065, China
[3] College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou 225127, China
[4] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China
[5] SHU General Intelligent Robotics Research Institute, Baoshan District, Shanghai 200444, China
* Correspondence: qxl@cust.edu.cn (X.Q.); gezhou@shu.edu.cn (Z.G.)

Abstract

Neuromorphic imaging sensors (event cameras) offer a promising paradigm for computational imaging and human pose estimation (HPE) under extreme illumination conditions. Nevertheless, dark-scene background activity originating from photodiode dark current and circuit thermal noise, together with hot-pixel noise, severely corrupts event streams and impedes reliable HPE in low-light scenarios. To address this issue, we propose an adaptive event denoising framework built upon a spatiotemporal Gaussian-weighted neighborhood model with a dynamic thresholding mechanism. It can effectively suppress background activity and hot-pixel noise while preserving edge and motion details critical for pose estimation. Leveraging this denoising front-end, we construct a complete dark-scene neuromorphic HPE pipeline by transferring the pre-trained MediaPipe model onto event-based time-surfaces. Quantitative and qualitative evaluations on public and self-collected datasets demonstrate that our approach outperforms state-of-the-art denoising methods with an improvement of over 20% on public benchmarks and over 30% on self-collected dark-scene data. We expect our work to pave the way toward reliable dark-scene human–robot interaction through robust neuromorphic pose estimation.

Keywords: neuromorphic imaging; dark scene; human pose estimation; adaptive denoising

1. Introduction

Dark-scene computational imaging has emerged as a pivotal technique in modern visual perception, underpinning a wide range of critical low-light applications such as night sensing, low-illumination motion detection, and dim-environment monitoring [1,2]. High-fidelity, low-latency imaging in unstructured dark scenes imposes rigorous demands on visual sensing paradigms. Specifically, perception systems must retain fine structural features and dynamic motion information while maintaining robust resistance against extreme low-light interference. As a cornerstone of human-centric vision tasks, human pose estimation (HPE) plays an instrumental role in smart video surveillance, sports motion analysis, and human–computer interaction [3]. For scenarios requiring assessment of human agents, the ability to perform accurate pose estimation constitutes a fundamental prerequisite for downstream pipelines, including action recognition, posture analysis, emotion tracking, and intent estimation. Consequently, HPE has been intensively investigated in recent years across various sensory modalities, including conventional digital cameras, marker-based motion capture systems, and RGB-D sensors.

Check for updates

Received: 29 June 2026
Revised: 28 August 2026
Accepted: 28 August 2026
Published: 30 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Despite recent progress, performing reliable HPE under night scenes remains an unresolved frontier, as standard synchronous vision sensors are fundamentally constrained by their frame-integration paradigm. This mechanism introduces an intractable trade-off: long exposure times induce pronounced motion blur that destroys tracking accuracy, whereas short exposure times cause severe perceptual "blindness" in low-light conditions. Furthermore, the limited dynamic range (typically ∼60 dB) of standard CMOS sensors frequently results in under- or over-saturation in high-contrast or silhouette scenarios. In safety-critical contexts such as human–robot interaction, a temporary loss of perception due to these environmental factors immediately ruptures the control loop, potentially leading to erratic robot behaviors, collisions, or structural damage.

Neuromorphic event cameras have gained significant traction as an effective means to overcome these boundaries, offering a promising avenue for "seeing human pose at nighttime" [4,5]. By asynchronously capturing pixel-level intensity changes rather than static frames, these bio-inspired sensors deliver information with microsecond-level latency and high dynamic range. Such advantages benefit night-scene HPE substantially, particularly for tracking fast, varying-speed human gestures in unconstrained dark environments without markers. Nevertheless, operating in extreme low-light conditions triggers severe signal degradations. Due to the high temporal sensitivity of event sensors, dark environments significantly amplify non-physical disturbances, manifesting background noise that heavily corrupts the valid human motion signal. Specifically, this dark-scene degradation stems from background activity (BA) noise driven by photodiode dark currents and circuit thermal noise, alongside hot pixel noise caused by structurally compromised pixels that continuously fire in total darkness. These non-physical events lack structural relevance, drastically increase downstream computational complexity, and severely submerge sparse human motion features. Therefore, accurately distinguishing real, motion-induced events from stochastic noise under low-light conditions remains a critical challenge for event-based HPE pipelines.

To overcome these low-light degradation bottlenecks, we present an unsupervised event restoration framework integrated into a complete dark-scene neuromorphic HPE pipeline. The core of our method is a spatiotemporal Gaussian-weighted neighborhood model coupled with an dynamic thresholding strategy: it evaluates each event's causal local density using a Gaussian decay kernel over both spatial and temporal dimensions and dynamically adjusts the discrimination threshold according to the computed density to suppress background activity and hot-pixel noise while preserving edge and motion details. This design eliminates reliance on hand-crafted filters or labeled training data, ensuring robust generalization across varying conditions. Our work demonstrates that unsupervised spatiotemporal density modeling alone can effectively restore event streams in low-light environments. This framework paves the way for reliable human–robot interaction in dark scenes, enabling applications ranging from night-time human motion analysis to autonomous systems operating under adverse illumination.

2. Related Works

Dark-scene visual perception has been extensively studied within the frame-based paradigm. Traditional low-light image enhancement approaches primarily relied on hand-crafted priors, such as histogram equalization [6,7], Retinex theory [8], and statistical sparsity [9]. Concurrently, alternative sensory modalities like infrared imaging [10] and hybrid systems [11] were developed to broaden spectral acquisition. In recent years, learning-based methods have achieved dominant success [1,12], with several video enhancement methods leveraging paired datasets to enforce cross-frame temporal consistency in dynamic scenes [13,14]. However, these frame-based paradigms face an insurmountable

physical bottleneck: enforcing short exposure times in fast-moving dark scenes inherently leads to severe photon starvation and irreversible loss of structural details, which downstream enhancement algorithms fail to recover. To circumvent these limitations, neuromorphic imaging has emerged as a promising frontier for night vision. Frameworks like DVS-Dark [15] transfer daylight textures to the nighttime domain by learning shared structural features, while various event-guided enhancement methods [16,17] utilize the high dynamic range (HDR) and microsecond latency of event cameras to guide intensity reconstruction. Despite their ingenuity, these methods often overlook the fact that the spatiotemporal distribution of event streams in real-world nighttime environments suffers from severe degradation, which can severely degrade downstream HPE quality.

To cope with dark-scene neuromorphic imaging degradation, various event signal restoration and filter mechanisms have been proposed, which can be broadly categorized into traditional spatiotemporal modeling and data-driven learning approaches. Traditional methods typically evaluate event consistency by exploiting localized neighborhood correlations or density to filter out dark-induced non-idealities. For instance, Feng et al. [18] designed a density-based filter to preserve valid structures based on local spatiotemporal event density. Mueggler et al. [19] utilized time surfaces for fast corner detection, leveraging temporal weighting to amplify edge features, a concept further advanced by Baldwin et al. [20] via the Inceptive Event Time Surface (IETS). Additionally, the Double Window Filter (DWF) [21] structures spatiotemporal correlations using sequential queues to isolate background disturbances from valid motion signals. More recently, Fang et al. [22] proposed a window-based denoiser to enhance temporal consistency, while Lin et al. [23] introduced a Gaussian self-adjusting threshold filter (GMCM) for motion denoising, but it relies on explicit motion models for two-stage denoising. For complex or ultra-low-light scenarios where hand-crafted filters struggle to model non-stationary noise, deep learning-based methods have attracted substantial interest. Frameworks like EDnCNN [24], AEDNet [25], and the Quantified Multi-Layer Perception Filter (QMLPF) [26] attempt to learn the boundary between valid signals and dark-induced noise by processing time surfaces or modeling spatiotemporal components separately. Despite these advancements, learning-based regression models often lead to poor generalization when transferred to unconstrained testing scenarios.

Unlocking these degraded event streams holds profound potential for high-speed, human-centric perception tasks, particularly HPE. Event cameras offer a paradigm shift for capturing high-frequency dynamics while inherently suppressing redundant static data [27,28]. This sensor capability is uniquely suited for "seeing motion at nighttime", maintaining tracking fidelity during fast, varying-speed human gestures that would otherwise suffer from severe inter-frame aliasing or temporal undersampling in standard 30 Hz or 60 Hz frame-based vision pipelines. The recent literature has begun to address HPE directly from event streams using representations like time surfaces, voxel grids, and spiking neural networks (SNNs) [29]. However, existing event-based HPE models are heavily constrained by their dependency on high-quality intensity signals, making them highly vulnerable to dark-scene degradation. For instance, EventCap [30] and EventHPE [31] strictly rely on hybrid event cameras that output both asynchronous events and low-frequency grayscale images to optimize or initialize 3D human poses. However, the concurrent grayscale frames provided by such hybrid sensors (e.g., DAVIS) are themselves severely degraded or entirely lost in extreme low-light conditions, making these frame-dependent pipelines ill suited for robust nighttime operation.

3. Methodology

The overall framework of the proposed method is illustrated in Figure 1, which comprises three key components: neuromorphic imaging, dark-scene restoration, and human pose estimation.

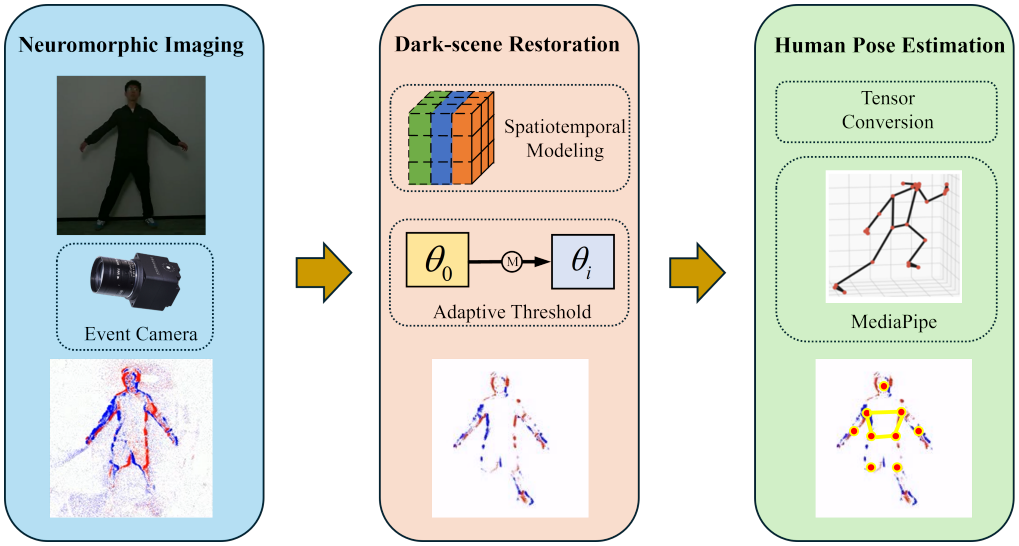

Figure 1. An overview of the proposed dark-scene neuromorphic human pose estimation framework. The pipeline consists of three primary stages: (**Left**) Neuromorphic Imaging: A bio-inspired event camera captures raw human-motion event streams under low-illumination conditions, which are contaminated by dark-scene background noise and hot-pixel artifacts. (**Middle**) Dark-scene Restoration: Joint spatiotemporal neighborhood modeling with adaptive threshold mechanism to filter noise and recover fine structural motion edges. (**Right**) Human Pose Estimation: The denoised event streams are transformed into event-based time surfaces, which are fed into the pre-trained MediaPipe model for human landmark detection.

3.1. Event Representation

Event cameras generate asynchronous pixel response signals triggered by luminance changes in the scene, which differ fundamentally from the frame-based output of traditional optical cameras. Each independent response signal can be characterized by a quadruple parameter set. For the k-th acquired event sample in a data stream with N samples, the mathematical definition is expressed as:

$$s_k = (x_k, y_k, t_k, p_k), \quad k = 1, 2, \ldots, N, \tag{1}$$

where (x_k, y_k) corresponds to the two-dimensional pixel position of the triggered event on the sensor imaging plane, t_k denotes the precise recording timestamp of the event, and the binary variable $p_k \in \{-1, 1\}$ indicates the luminance change polarity of the pixel. A pixel is assigned a polarity of $+1$ when its local logarithmic intensity increment surpasses the positive contrast threshold, indicating an ON event (dark-to-bright transition); otherwise, the polarity is -1, indicating an OFF event. Due to the inherent characteristics of event cameras, there exists a prominent mismatch between temporal and spatial resolution: the temporal sampling precision of event signals is far higher than the spatial pixel resolution, which causes scale inconsistency in subsequent feature extraction. To solve this issue, we analyze the temporal interval distribution characteristics of continuous event sequences and adaptively determine the effective time window range based on statistical percentile features. An adaptive time window is determined using percentile estimation:

$$\Delta t = \max(\text{Percentile}(\{\Delta t_k\}, \alpha), t_{low}), \tag{2}$$

where Δt_k represents the time interval between consecutive event samples, Percentile$(\cdot, \alpha)$ denotes the α% percentile function, and t_{low} is the preset minimum threshold of the time window. The coefficient α reflects the statistical confidence level for capturing valid local motion events. The upper $(1-\alpha)$ fraction of long-interval outliers, dominated by noise, are truncated, avoiding excessive window expansion caused by sparse noise. The lower bound t_{low} prevents window collapse during high-speed motion, enforcing a minimum integration depth compatible with sensor bandwidth. This robust formulation effectively eliminates outlier time interference and enables the temporal scale to dynamically adapt to varying event densities across different dynamic scenes.

3.2. Spatiotemporal Gaussian-Weighted Denoising Model

A spatiotemporal Gaussian-weighted neighborhood model is then constructed to distinguish effective signal events from noise events. Traditional event camera denoising methods mostly rely on single spatial neighborhood feature judgment, which use the spatial distribution consistency of adjacent pixels as the basis for noise discrimination. Such methods can suppress partial fixed-pixel noise but ignore the strong temporal correlation characteristics of real scene events, resulting in incomplete noise suppression and easy loss of weak effective signal events. Different from single spatial domain modeling, the proposed method integrates spatial adjacency and temporal continuity features of events and uses a Gaussian function to assign differential weights to neighboring events, so as to highlight effective events with continuous spatiotemporal distribution and suppress isolated abnormal noise points.

For any event $s_k = (x_k, y_k, t_k, p_k)$, its attribute (valid signal or isolated noise) is determined by analyzing the spatiotemporal distribution of surrounding neighboring events with the same polarity. To strictly obey temporal causality for online streaming processing and prevent future event buffering, the causal spatiotemporal neighborhood set of s_k is defined as:

$$\Omega(s_k) = \left\{ s_m \;\middle|\; (x_m - x_k)^2 + (y_m - y_k)^2 \le (W/2)^2,\ 0 < t_k - t_m \le \Delta t,\ p_m = p_k \right\}, \quad (3)$$

where W denotes the spatial neighborhood window size. The inequality $0 < t_k - t_m$ explicitly guarantees that only historical preceding events are referenced while excluding self-referencing. The spatio-temporal correlation weight $w_{m,k}$ between a past neighboring event $s_m \in \Omega(s_k)$ and the target event s_k is governed by a Gaussian decay kernel:

$$w_{m,k} = \exp\left(-\frac{(x_m - x_k)^2 + (y_m - y_k)^2}{2\sigma_s^2} - \frac{(t_k - t_m)^2}{2\sigma_t^2} \right), \quad (4)$$

where σ_s and σ_t represent the spatial and temporal Gaussian standard deviations, respectively. Physically, σ_s denotes the spatial decay length of valid motion edges on the pixel plane, reflecting the thickness of moving contours. σ_t controls the temporal decay of events from the same edge crossing a pixel, prioritizing temporally adjacent events and diminishing sporadic noise. The localized density ρ_k of the target event s_k is accumulated across all past neighbors in its causal set:

$$\rho_k = \sum_{s_m \in \Omega(s_k)} w_{m,k}. \quad (5)$$

In complex scenes, the event distribution presents spatial and temporal heterogeneity: high-speed moving targets produce dense continuous event streams, while static background areas only generate sparse discrete events. Traditional fixed threshold strategy is prone to misjudgment problems: a high threshold will filter out weak effective signals in sparse regions, while a low threshold cannot completely eliminate discrete noise in dense

regions. To solve this defect, the noise discrimination threshold is adjusted dynamically according to the local weighted density of each event, realizing adaptive decision making of different regional events. The adaptive threshold for determining whether an event is valid or noise is formulated as follows:

$$\vartheta_k = \max\left(\vartheta_{\min}, \min\left(\vartheta_{\min} + \left\lfloor \frac{\rho_k}{\beta} \right\rfloor, \vartheta_{\max}\right)\right) \tag{6}$$

where $\vartheta_{\min}$ and $\vartheta_{\max}$ represent the minimum and maximum discrimination threshold, respectively, and β denotes the dynamic scaling factor. Bounded by $\vartheta_{\min}$ and $\vartheta_{\max}$, the threshold ϑ_k operates as a non-linear piecewise step function of local event density ρ_k. Specifically, for sparse background areas ($\rho_k < \beta$), the threshold defaults to the base value $\vartheta_{\min}$ to eliminate isolated noise. For medium-density regions, ϑ_k increases discretely in step-wise fashion to tighten filtering criteria. For ultra-dense event streams, ϑ_k saturates at $\vartheta_{\max}$, avoiding excessive filtering and preserving fine motion edge structures.

3.3. Pose Estimation Pipeline

To achieve accurate pose estimation, we transfer the restored event streams into a continuous, lightweight spatiotemporal time-surface representation. For each pixel coordinate (x, y) within a temporal accumulation interval T, the continuous time-surface $\mathcal{TS}$ is mathematically modeled via an exponential decay function:

$$\mathcal{TS}(x, y, t) = \sum_{t_k \in T} p_k \cdot \exp\left(-\frac{t - t_k}{\gamma}\right). \tag{7}$$

where γ dictates the memory retention span of historical motion trail traces. The converted spatiotemporal tensors are subsequently fed into a high-performance, lightweight MediaPipe pose estimation framework [32]. Notably, the MediaPipe backbone is originally pre-trained on conventional frame images rather than event-based inputs. Nevertheless, event-based time-surfaces retain the human body's structural boundaries and motion contours, which constitute the visual cues for MediaPipe.

The pose estimation engine processes the time surfaces to regress the spatial coordinates of a standardized human skeletal topology, which encompasses 33 prominent body keypoints (including the shoulders, elbows, wrists, hips, knees, and ankles). To quantitatively evaluate pose estimation performance, we define a keypoint detection rate as the ratio of the total number of valid keypoints to the theoretical maximum number of keypoints across all time surfaces:

$$R_{\text{det}} = \frac{\sum_{f=1}^{N_f} \sum_{i=1}^{33} \mathbf{1}[\text{keypoint}_{i,f} \text{ is valid}]}{33 \cdot N_f} \times 100\% \tag{8}$$

where N_f denotes the total number of time surfaces.

4. Experimental Results

4.1. Implementation Details

All algorithms were implemented in Python and executed on a laptop with Intel Core i9 CPU and NVIDIA RTX 5060 GPU. For the denoising module, the adaptive temporal window parameters were set to $\alpha = 0.95$ and $t_{low} = 0.5$ ms. The spatiotemporal Gaussian kernel operated over a $W = 5 \times 5$ spatial neighborhood with spatial standard deviation $\sigma_s = W/2 = 2.5$ pixels with temporal standard deviation $\sigma_t = \Delta t/2$. The adaptive discrimination threshold range was set with $\vartheta_{\min} = 5$, $\vartheta_{\max} = 12$, and the scaling factor

was set to $\beta = 100$. For time-surface accumulation, the temporal interval T was set to 60 ms, and the exponential decay constant γ was set to 20 ms.

4.2. Public Dataset Validation

First, we evaluated the proposed method on the public E-MLB dataset [33], where neutral-density (ND) filters are attached to a DAVIS 346 event camera (346 × 260 resolution) to create controlled real-world low-light conditions. The proposed method is compared against three representative denoising approaches: DWF [21], YNoise [18], and EDnCNN [24]. The visual comparison results are shown in Figure 2, where red and blue colors represent different polarities of events. Columns correspond to results captured with different ND filters. Rows sequentially display the noisy raw events, our proposed method, DWF, YNoise, and EDnCNN, respectively.

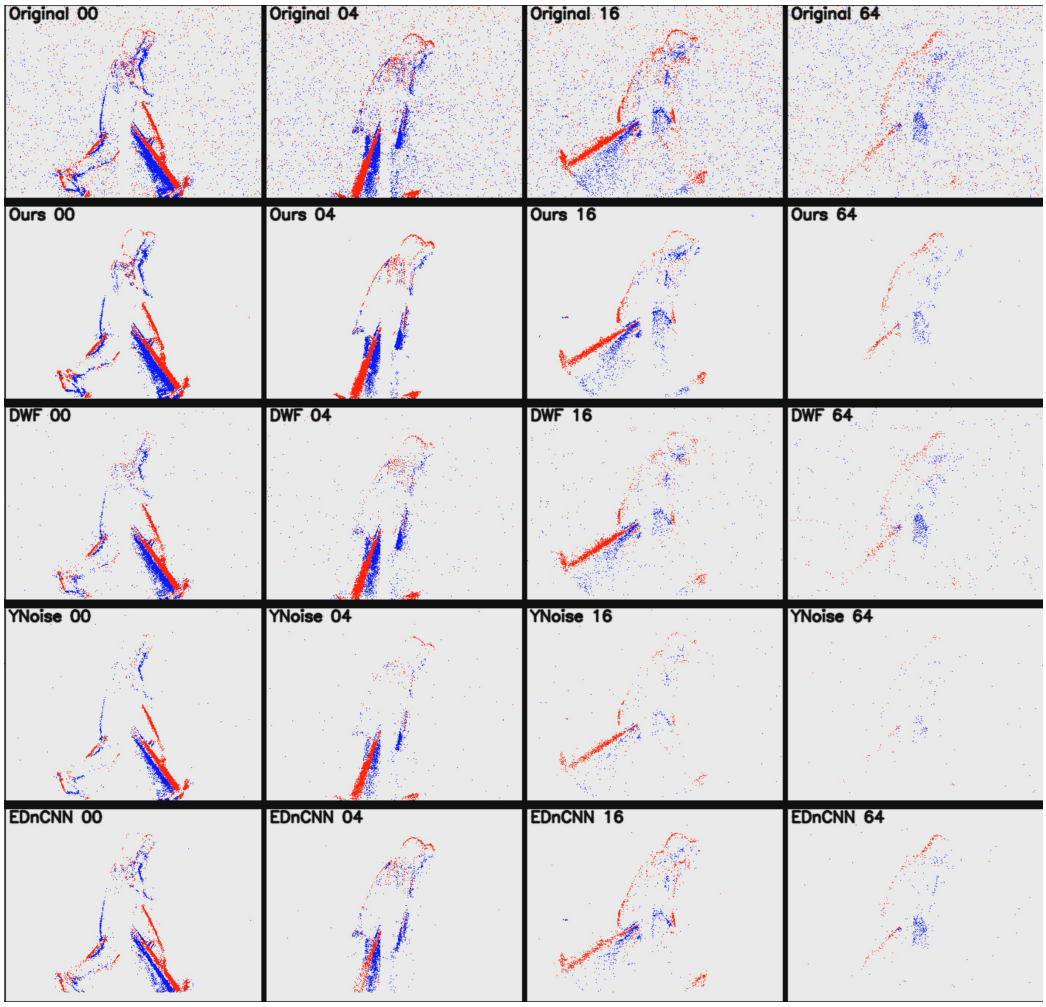

Figure 2. Visual comparisons of "Ski_Walker" sequence across different low-light degradations. Rows: raw noisy events, our method, DWF, YNoise, EDnCNN. Columns: degradation level induced by ND filters of 00, 04, 16, and 64.

Among the four compared methods, DWF and YNoise exhibit severe over-suppression, erasing substantial valid structural events and causing irreversible loss of textures. EDnCNN cannot fully suppress noise interference and produces noticeable blurring along contours. By contrast, our method achieves a desirable trade-off between noise removal and valid-event preservation across all degradation levels, showing better noise robustness and superior structural feature retention compared with competing baselines.

For quantitative evaluation, we adopt two label-free metrics proposed in [34]: the Continuous Contrast Curve (CCC) and the Area of Continuous Contrast Curve (AOCC). The CCC characterizes structural contrast variations of event streams across diverse time scales, where higher curve values indicate more distinct edges and contours. It should be noted that this metric is non-monotonic by design, meaning that a noisy raw stream does not necessarily produce the highest CCC value across all time windows. The AOCC, obtained via numerical integration of the CCC, provides a comprehensive measure of structural preservation over multiple time windows. A higher AOCC reflects that the method effectively retains informative event structures while suppressing noise. We present the CCC and AOCC results in Figure 3 and Table 1, respectively. Our proposed method achieves the highest AOCC score among all compared approaches, demonstrating its superior and stable capability in preserving structural event information under diverse noise levels.

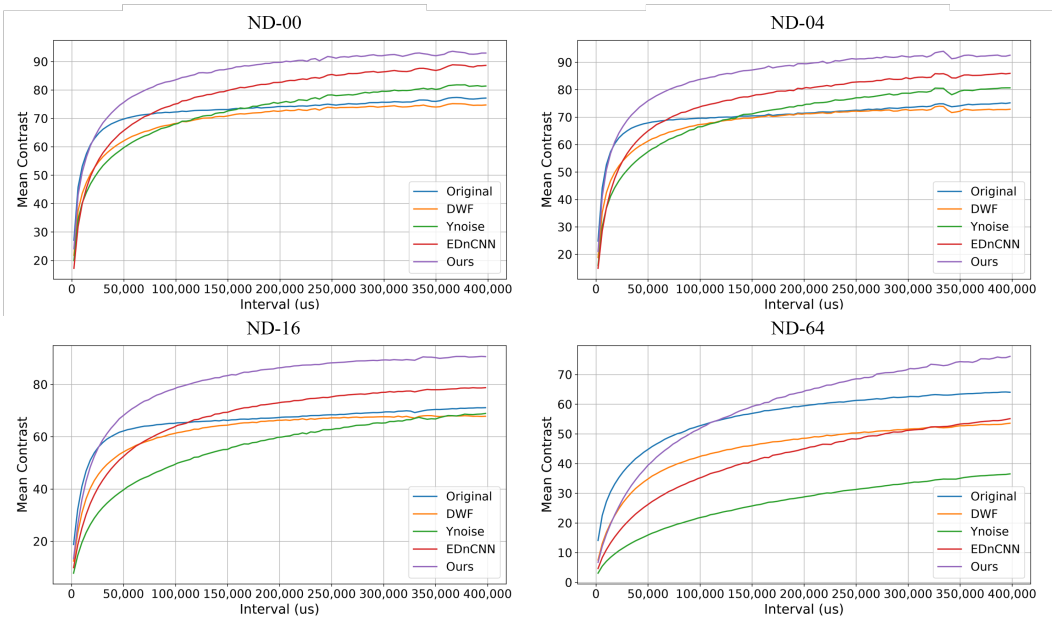

Figure 3. Quantitative evaluation of CCC performance on the E-MLB dataset. Higher curve values correspond to more distinct edges and contours.

Table 1. Quantitative evaluation of AOCC ($\times 10^7$) performance on the E-MLB dataset. Best results are highlighted in bold.

Degradation	ND-00	ND-04	ND-16	ND-64
Original	2.88	2.79	2.60	2.21
DWF	2.75	2.69	2.47	1.79
YNoise	2.84	2.79	2.20	1.06
EDnCNN	3.10	3.02	2.68	1.65
Ours	**3.39**	**3.38**	**3.21**	**2.35**

The results of keypoint detection rate are shown in Figure 4. With the intensification of light attenuation, all competing methods suffer from noticeable performance degradation, but our method consistently maintains the top detection accuracy, and at the ND-16 level, it still retains a detection rate of nearly 78%, which is significantly higher than that of other comparison methods. More importantly, under the most challenging extreme low-light condition (ND-64), our method still preserves a keypoint detection rate above 50%, which demonstrates that the proposed denoising strategy can effectively suppress noise and retain the key structural information of events in harsh low-light scenarios. On average,

our method achieves an over 20% higher detection rate than the competing methods, demonstrating its superiority in robustness and reliability for dark-scene HPE.

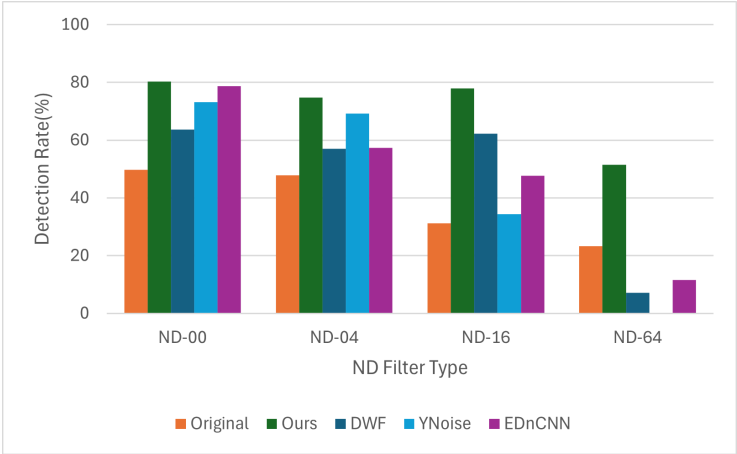

Figure 4. Keypoint detection rate of the original raw events, DWF, YNoise, EDnCNN and our method under degradation levels from ND-00 to ND-64. The vertical axis denotes keypoint detection rate (%), and the horizontal axis represents ND filter type.

4.3. Evaluation on Self-Collected Dataset

Then, we evaluated the proposed method on our self-collected dataset. We captured four typical human poses, including squat, shuffle, jump and kick at 30 lux using a Prophesee EVK4 event camera (1280 × 720 resolution). Visual comparisons are shown in Figure 5. Columns correspond to different human poses. Rows sequentially display the noisy raw events, our proposed method, DWF, YNoise, and EDnCNN, respectively. Similar to previous comparative results, our method achieves better balance between noise removal and structure preservation across all degradation levels, showing stronger robustness and feature retention than the baselines.

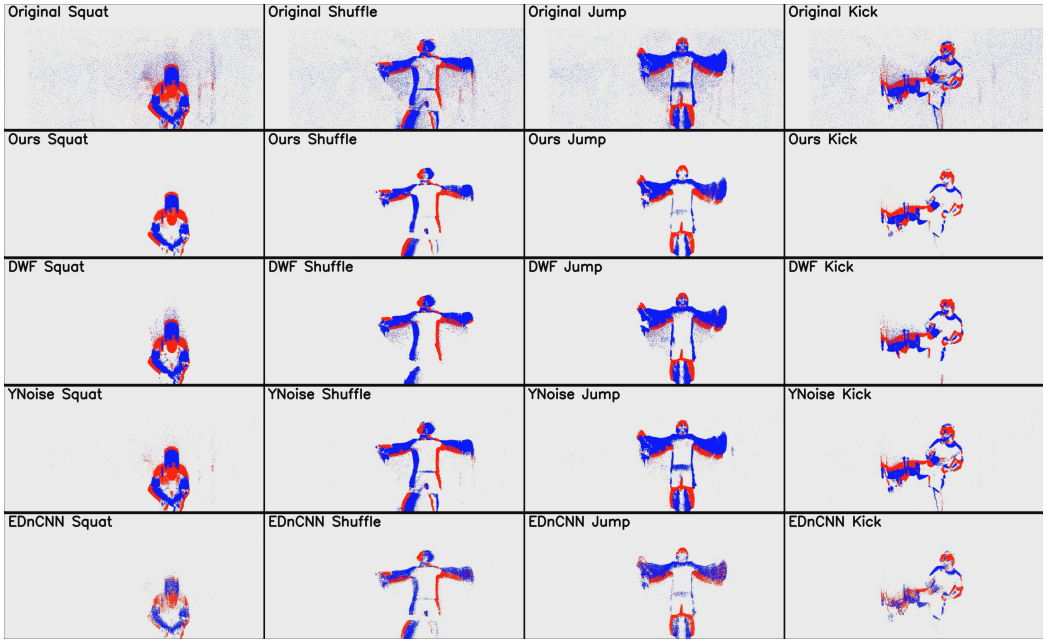

Figure 5. Visual comparisons of self-collected dark-scene human poses. Rows: raw noisy events, our method, DWF, YNoise, EDnCNN. Columns: Human poses of squat, shuffle, jump and kick.

Figure 6 and Table 2 illustrate the performance of different event denoising methods in terms of CCC and AOCC across four human motion sequences. On the Squat sequence,

the original raw events obtain a marginally higher AOCC score than our proposed method. This phenomenon arises due to false contrast from noise, so we take the original raw events as the reference baseline and mainly focus on comparing the AOCC and CCC performance among different denoising algorithms. When comparing among denoising methods, our approach achieves the highest AOCC among all competitors, which verifies that the proposed approach effectively preserves the structural edge information of event streams while suppressing noise, delivering higher-contrast event representations that benefit downstream human pose estimation tasks.

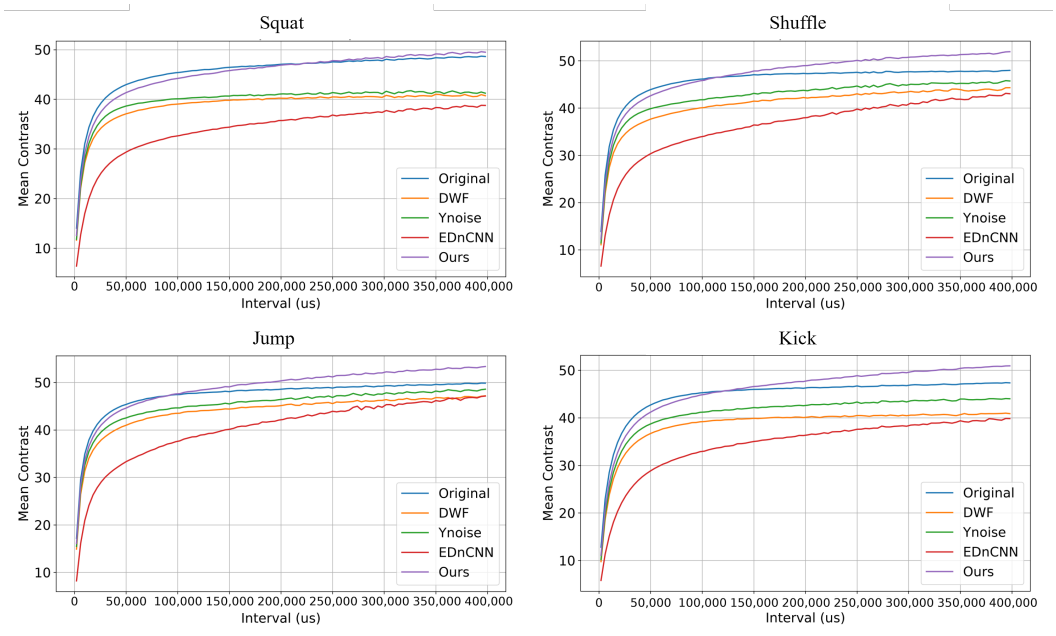

Figure 6. Quantitative evaluation of CCC performance on the self-collected dataset. Higher curve values correspond to more distinct edges and contours.

Table 2. Quantitative evaluation of AOCC ($\times 10^7$) performance on the self-collected dataset. Best results are highlighted in bold.

Pose	Squat	Shuffle	Jump	Kick
Original	**1.81**	1.82	1.88	1.78
DWF	1.54	1.62	1.74	1.54
YNoise	1.58	1.69	1.80	1.63
EDnCNN	1.35	1.44	1.60	1.37
Ours	1.80	**1.87**	**1.94**	**1.82**

The keypoint detection results are shown in Figure 7. The experimental results reveal that our method consistently achieves the highest keypoint detection rate across all four motion scenarios, demonstrating its stable superiority over baseline methods. Specifically, for the squat sequence with relatively small motion amplitude, all methods deliver limited detection performance, while our method still obtains a detection rate of approximately 23%, outperforming the original events and other denoising methods. For sequences with more vigorous movements including shuffle, jump and kick, the overall detection performance of all methods improves significantly, and our method maintains a remarkable leading position. Notably, EDnCNN suffers from severe performance degradation on the kick sequence due to over-suppression, whereas our method remains robust and maintains a clear advantage over all competing approaches. On average, our method achieves over 30% higher detection rate than the competing methods. These findings verify that

the proposed denoising method can effectively preserve the motion edge and structure information contained in event streams while suppressing noise, which greatly enhances the accuracy and generalization ability of event-based human pose keypoint detection under diverse motion patterns.

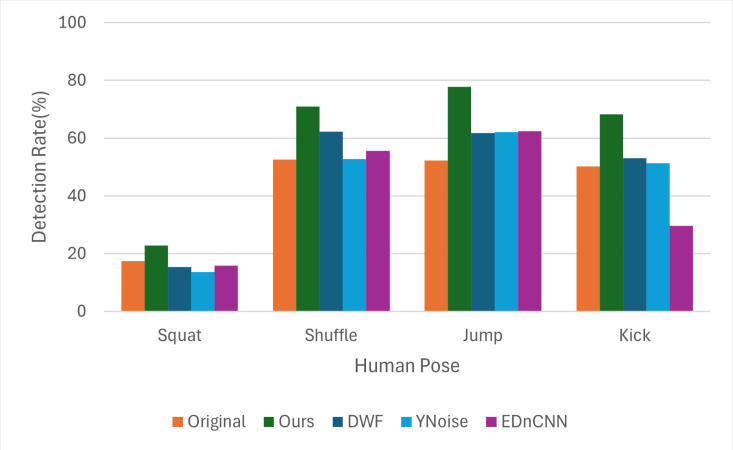

Figure 7. Keypoint detection rate of the original raw events, DWF, YNoise, EDnCNN and our method on dark-scene human poses. The vertical axis denotes keypoint detection rate (%), and the horizontal axis represents human pose.

5. Discussion

Despite these encouraging results, the current study still has notable limitations that constrain its broader applicability. First, when illuminance drops to the sub-lux level, valid motion events become extremely sparse, and the signal-to-noise ratio decreases sharply. Limited by laboratory conditions, extreme darkness is not achieved in the current experiments, and we will extend validation to lower illuminance levels in future work. Meanwhile, the adaptive mechanisms could be further optimized to handle high-dynamic-range (HDR) transitions, such as sudden flashes or moving light sources in the dark. Second, the current HPE pipeline adopts a transfer learning paradigm that feeds accumulated time-surfaces into a frame-pretrained MediaPipe model. Despite its practicality, this frame-converted representation cannot fully exploit the fine-grained asynchronous temporal information of native event streams, limiting the upper bound of pose estimation accuracy. Finally, the throughput of our current implementation is approximately 10,000 events per second on a consumer-level laptop. We aim to implement the entire front-end directly onto edge hardware, such as Field-Programmable Gate Arrays (FPGAs) or neuromorphic processors (e.g., Intel Loihi), to further minimize latency and power consumption, paving the way for fully autonomous, night-operational humanoid robots.

6. Conclusions

In this work, we target the challenging problem of event-based human pose estimation under dark-scene conditions, where event cameras suffer from severe background activity noise and hot-pixel artifacts. We presented an unsupervised adaptive event denoising framework built upon a spatiotemporal Gaussian-weighted neighborhood model with dynamic thresholding mechanism, which removes noise events while preserving critical motion-related edge structures without requiring labeled training data. By converting the restored event streams into time-surface representations, we construct a human pose estimation pipeline adapted for low-light environments based on MediaPipe pose detector. Extensive evaluations were conducted on both the public dataset and our self-collected dark-scene human motion dataset. Quantitative results in terms of CCC, AOCC, and

keypoint detection rate, together with qualitative visual comparisons, demonstrate that our method achieves a favorable trade-off between noise suppression and valid-event preservation. Compared with competing denoising approaches, our approach yields higher structural contrast and substantially improves keypoint detection robustness under various low-light degradation levels and human poses.

Author Contributions: Conceptualization, H.B. and X.Q.; methodology, Y.D. and C.L.; experiment, F.S. and C.L.; writing—original draft preparation, H.B., H.F. and S.L.; writing—review and editing, C.W. and C.L.; supervision, H.C. and Z.G.; funding acquisition, X.Q. and Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Science and Technology on Low-Light-Level Night Vision Laboratory, China (No.2025-JCJQ-LB-025-04), the National Natural Science Foundation of China (62505167), and the Science and Technology Commission of the Shanghai Municipality (25ZR1402149).

Data Availability Statement: The data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to See in the Dark. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 3291–3300.
2. Chen, X.; Yu, Y. Image Illumination Enhancement for Construction Worker Pose Estimation in Low-light Conditions. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 147–162.
3. Goyal, G.; Di Pietro, F.; Carissimi, N.; Glover, A.; Bartolozzi, C. MoveEnet: Online High-Frequency Human Pose Estimation with an Event Camera. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 4024–4033.
4. Lichtsteiner, P.; Posch, C.; Delbruck, T. A 128× 128 120 dB 15 µs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE J. Solid-State Circuits* **2008**, *43*, 566–576.
5. Liu, H.; Xu, J.; Peng, S.; Chang, Y.; Zhou, H.; Duan, Y.; Zhu, L.; Tian, Y.; Yan, L. NER-Net+: Seeing Motion at Nighttime with an Event Camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4768–4786. [CrossRef]
6. Arici, T.; Dikbas, S.; Altunbasak, Y. A Histogram Modification Framework and Its Application for Image Contrast Enhancement. *IEEE Trans. Image Process.* **2009**, *18*, 1921–1935. [CrossRef]
7. Celik, T.; Tjahjadi, T. Contextual and Variational Contrast Enhancement. *IEEE Trans. Image Process.* **2011**, *20*, 3431–3441. [CrossRef]
8. Guo, X.; Li, Y.; Ling, H. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Process.* **2017**, *26*, 982–993. [CrossRef]
9. Fu, X.; Zeng, D.; Huang, Y.; Zhang, X.P.; Ding, X. A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 2782–2790.
10. Luo, F.; Li, Y.; Zeng, G.; Peng, P.; Wang, G.; Li, Y. Thermal Infrared Image Colorization for Nighttime Driving Scenes with Top-Down Guided Attention. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15808–15823. [CrossRef]
11. Ashiba, M.I.; Tolba, M.S.; El-Fishawy, A.S.; El-Samie, F.E.A. Hybrid Enhancement of Infrared Night Vision Imaging System. *Multimed. Tools Appl.* **2019**, *79*, 6085–6108. [CrossRef]
12. Zain Ul Abideen, M.; An, N.; Ma, L.; Liu, R. SNRD-Net: SNR-Aware Dual Enhancement Network for Low-Light Images. *Comput. Vis. Image Underst.* **2026**, *268*, 104776. [CrossRef]
13. Wang, R.; Xu, X.; Fu, C.W.; Lu, J.; Yu, B.; Jia, J. Seeing Dynamic Scene in the Dark: A High-Quality Video Dataset with Mechatronic Alignment. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 9700–9709.
14. Wang, J.; Fu, H.; Zheng, W.; Wang, X.; Wang, X.; Zhang, H.; Ma, H. Rethinking the Low-Light Video Enhancement: Benchmark Datasets and Methods. *IEEE Trans. Image Process.* **2025**, *34*, 6799–6812. [CrossRef]
15. Zhang, S.; Zhang, Y.; Jiang, Z.; Zou, D.; Ren, J.; Zhou, B. Learning to See in the Dark with Events. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 666–682.
16. Liu, L.; An, J.; Liu, J.; Yuan, S.; Chen, X.; Zhou, W.; Li, H.; Wang, Y.F.; Tian, Q. Low-Light Video Enhancement with Synthetic Event Guidance. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 1692–1700. [CrossRef]

17. Liang, J.; Yang, Y.; Li, B.; Duan, P.; Xu, Y.; Shi, B. Coherent Event Guided Low-Light Video Enhancement. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 10581–10591.
18. Feng, Y.; Lv, H.; Liu, H.; Zhang, Y.; Xiao, Y.; Han, C. Event Density Based Denoising Method for Dynamic Vision Sensor. *Appl. Sci.* **2020**, *10*, 2024. [CrossRef]
19. Mueggler, E.; Bartolozzi, C.; Scaramuzza, D. Fast Event-based Corner Detection. In *Proceedings of the British Machine Vision Conference (BMVC)*; University of Zurich: Zürich, Switzerland, 2017.
20. Baldwin, R.W.; Almatrafi, M.; Kaufman, J.R.; Asari, V.; Hirakawa, K. Inceptive Event Time-Surfaces for Object Classification Using Neuromorphic Cameras. In *Proceedings of the Image Analysis and Recognition*; Springer: Cham, Switzerland, 2019; pp. 395–403.
21. Guo, S.; Delbruck, T. Low Cost and Latency Event Camera Background Activity Denoising. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 785–795. [CrossRef] [PubMed]
22. Fang, H.; Wu, J.; Hou, Q.; Dong, W.; Shi, G. Fast Window-Based Event Denoising with Spatiotemporal Correlation Enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1381–1394. [CrossRef]
23. Lin, W.; Li, Y.; Xu, C.; Liu, L. A motion denoising algorithm with Gaussian self-adjusting threshold for event camera. *Vis. Comput.* **2024**, *40*, 6567–6580. [CrossRef]
24. Baldwin, R.; Almatrafi, M.; Asari, V.; Hirakawa, K. Event Probability Mask (EPM) and Event Denoising Convolutional Neural Network (EDnCNN) for Neuromorphic Cameras. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 1698–1707.
25. Fang, H.; Wu, J.; Li, L.; Hou, J.; Dong, W.; Shi, G. AEDNet: Asynchronous Event Denoising with Spatial-Temporal Correlation among Irregular Data. In *Proceedings of the the 30th ACM International Conference on Multimedia, Lisboa Portugal, 10–14 October 2022*; ACM: New York, NY, USA, 2022; pp. 1427–1435.
26. Rios-Navarro, A.; Guo, S.; Gnaneswaran, A.; Vijayakumar, K.; Linares-Barranco, A.; Aarrestad, T.; Kastner, R.; Delbruck, T. Within-Camera Multilayer Perceptron DVS Denoising. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 3933–3942.
27. Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.J.; Conradt, J.; Daniilidis, K.; et al. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 154–180. [CrossRef] [PubMed]
28. Rebecq, H.; Gehrig, D.; Scaramuzza, D. ESIM: An Open Event Camera Simulator. In *Proceedings of the 2nd Conference on Robot Learning (CoRL)*; PMLR: Cambridge, MA, USA, 2018; pp. 969–982.
29. Gehrig, D.; Loquercio, A.; Derpanis, K.G.; Scaramuzza, D. End-to-End Learning of Representations for Asynchronous Event-Based Data. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 2019*; IEEE: New York, NY, USA, 2019; pp. 5632–5642.
30. Xu, L.; Xu, W.; Golyanik, V.; Habermann, M.; Fang, L.; Theobalt, C. EventCap: Monocular 3D Capture of High-Speed Human Motions Using an Event Camera. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 4967–4977.
31. Zou, S.; Guo, C.; Zuo, X.; Wang, S.; Xiaoqin, H.; Chen, S.; Gong, M.; Cheng, L. EventHPE: Event-based 3D Human Pose and Shape Estimation. In *Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 11613–11622.
32. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. Mediapipe: A framework for Building Perception Pipelines. *arXiv* **2019**, arXiv:1906.08172.
33. Ding, S.; Chen, J.; Wang, Y.; Kang, Y.; Song, W.; Cheng, J.; Cao, Y. E-MLB: Multilevel Benchmark for Event-Based Camera Denoising. *IEEE Trans. Multimed.* **2024**, *26*, 65–76. [CrossRef]
34. Shi, C.; Guo, S.; Wei, B.; Liu, H.; Zhang, Y.; Song, N.; Jin, J. A Label-Free and Non-Monotonic Metric for Evaluating Denoising in Event Cameras. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 669–684. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.