IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Ensemble-Based Event Camera Place Recognition Under Varying Illumination

Therese Joseph Tobias Fischer Michael Milford

Abstract— **Compared to conventional cameras, event cameras provide a high dynamic range and low latency, offering greater robustness to rapid motion and challenging lighting conditions. Although the potential of event cameras for visual place recognition (VPR) has been established, developing robust VPR frameworks under severe illumination changes remains an open research problem. Here, we introduce an ensemble-based approach to event camera place recognition that combines sequence-matched results from multiple event-to-frame reconstructions, VPR feature extractors, and temporal resolutions. Unlike previous event-based ensemble methods, which only utilise temporal resolution, our broader fusion strategy delivers significantly improved robustness under varied lighting conditions (e.g., afternoon, sunset, night), achieving up to 77% relative improvement in Recall@1 across day-night transitions. We evaluate our approach on two long-term driving datasets (with 8 km per traverse) without metric subsampling, thereby preserving natural variations in speed and stop duration that influence event density. We also conduct a comprehensive analysis of key design choices, including binning strategies, reconstruction methods, and feature extractors, to identify the most critical components for robust performance. Additionally, we propose a modification to the standard sequence matching framework that enhances performance at longer sequence lengths. To facilitate future research, we release our codebase and benchmarking framework [1].**

Index Terms— **Localization, Computer Vision for Transportation**

I. Introduction

Event cameras are a class of vision sensors that record changes in brightness asynchronously and independently for each pixel, producing a stream of events that encode the time, pixel location, and polarity of intensity changes. Unlike conventional cameras, they offer microsecond temporal resolution, high dynamic range, and low latency, making them particularly suitable for high-speed, low-power applications and operation in challenging lighting conditions [7]–[9].

These characteristics have generated growing interest in applying event cameras to autonomous navigation tasks [8], [10]–[14], including visual place recognition (VPR), a core component of long-term localisation and map-based re-localisation. Traditional VPR systems rely on frame-based imagery and aim to extract features invariant to appearance changes, dynamic elements, viewpoint shifts, and lighting variation [15]–[19]. However, they often falter under extreme illumination changes (e.g., day-night) [20], where event cameras can still provide a useful signal. Despite their promise, event-based VPR remains underexplored, especially in conditions involving severe illumination variation. Key open research questions include: 1) how to best represent and reconstruct event data for downstream place recognition, and 2) whether combining complementary representations or VPR models can improve robustness under challenging conditions.

We present an ensemble-based approach to event camera VPR that enhances robustness through late score fusion. Specifically, we propose three ensemble strategies: (i) aggregating across different event-to-frame reconstruction methods, both classical (e.g., event count, time surface [1]) and learned (E2VID [2]); (ii) aggregating across multiple VPR feature extractors, including NetVLAD [3], CosPlace [4], MixVPR [5], and MegaLoc [6], and (iii) aggregating all configurations with temporal resolutions, event-to-frame reconstruction methods and VPR feature extractors.

Unlike prior event-based VPR ensembling that fuses predictions only across temporal resolutions [11], our strategy evaluates ensembles of reconstruction methods, feature extractors and temporal resolutions, yielding improved robustness under challenging appearance changes, including illumination shifts across the day. This broader fusion combines complementary cues from different representations, leading to greater robustness across diverse visual and motion conditions. We also use complete traverses without metric subsampling in our evaluations, which preserves natural variations in speed that are challenging in event-based settings where event density is motion-dependent. Finally, we apply a modified sequence matching strategy to each ensemble member, leveraging temporal consistency for improved performance.

In this work, we contribute:

1) An ensemble-based event VPR framework that aggregates predictions across reconstruction methods, feature extractors and temporal resolutions, achieving up to 77% relative gain in Recall@1 under day–night transitions.
2) A comprehensive analysis of event-based VPR design choices, including binning strategies (time- vs. count-based), polarity inclusion, reconstruction methods and feature extractor performance.
3) A modification to sequence matching from [21] and [22] with dynamic history length and z-score normalisation for improved performance at longer sequence lengths.
4) Extensive evaluation on two challenging long-term driv-

The authors are with the QUT Centre for Robotics, School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, QLD 4000, Australia. Email: t2.joseph@connect.qut.edu.au

This research was partially supported by the QUT Centre for Robotics and funding from ARC Laureate Fellowship FL210100156 to MM and ARC DECRA Fellowship DE240100149 to TF.

[1] https://github.com/theresejoseph/ensemble_event_vpr_bench

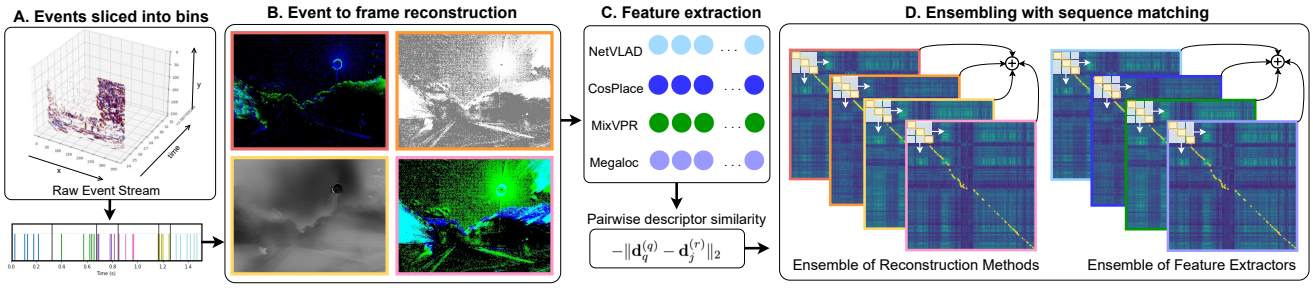

Fig. 1. **Overview of our ensemble-based event camera place recognition pipeline.** A) Each bin is reconstructed into a 2D frame using one of several methods: event count (with/without polarity), time surface [1], or the learned E2VID model [2]. (B) These reconstructed frames are processed by an ensemble of visual place recognition (VPR) feature extractors (NetVLAD [3], CosPlace [4], MixVPR [5], MegaLoc [6]) to generate global descriptors. (C) Pairwise descriptor similarities are computed, followed by sequence matching on each resulting similarity matrix. (D) Finally, sequence scores are aggregated across the multiple reconstruction methods, feature extractors, and temporal resolutions to enhance recognition performance.

ing datasets, *BrisbaneEventVPR* [11] and *NSAVP* [23] under varied lighting conditions.

II. Related Works

In this section, we review related works, focusing on three core areas that inform our contributions: conventional VPR methods using frame-based imagery (Section II-A), event-based VPR approaches, which adapt place recognition pipelines to event camera data (Section II-B), and ensemble techniques in VPR (Section II-C), which aim to improve robustness and performance through fusion of models or representations.

A. Place Recognition

Place recognition matches a query to a reference database via feature extraction and similarity search, enabling global re-localisation and loop closure [15]–[19], [24]. Visual place recognition (VPR) uses vision-based sensors such as conventional frame or event cameras. Beyond vision, LiDAR-based methods learn or extract geometry-aware point cloud descriptors [25], radar-based methods explore robustness to weather and illumination [26], and multimodal methods combine camera frames with LiDAR or events [27]–[29].

Modern VPR relies on deep networks: NetVLAD (VGG-16 + VLAD) [3], CosPlace (ResNet-50 with disjoint classes) [4], MixVPR (MLP mixing on flattened features) [5], and MegaLoc (DINOv2-Base + SALAD) [6]. In our work, we adopt these methods for event-based place recognition, evaluating both their performance and the benefits of ensembling.

B. Event-Based VPR

Event cameras output a continuous stream of sparse, asynchronous events rather than conventional image frames. Compared to conventional frame-based cameras, they provide microsecond-level temporal resolution, very low latency, high dynamic range, and low power consumption, which makes them well-suited to mobile sensing under fast motion and challenging illumination conditions such as night driving or headlight glare. These sensing properties are directly beneficial for visual place recognition, where robustness to motion blur, abrupt lighting changes, and low-light scenes is essential for reliable matching along long vehicle traverses [30], [31].

To leverage standard vision pipelines, particularly those based on deep learning, these events are typically aggregated into frame-like representations or voxel tensors. An early study on event-based visual SLAM [32] demonstrated the feasibility of place recognition over a 2.7km route using SeqSLAM [21] applied to polarity-removed event count reconstructions. Fischer et al. [11] later introduced an ensemble-based approach that fused multiple temporal windows, reconstructing event frames with E2VID [2] and extracting features using NetVLAD. In follow-up work [13], they proposed a lightweight method using a sparse subset of varying pixels as descriptors and sum of absolute differences (SAD) for matching, achieving competitive performance with minimal compute cost.

Subsequent works have adopted learning-based paradigms. Lee et al. [10] introduced a semi-supervised network that reconstructs denoised edge images using GRUs and convolutional modules. Ev-ReconNet [12] is a CNN-based autoencoder for edge reconstruction, which was later converted into a spiking neural network (SNN) for neuromorphic deployment. Both methods use NetVLAD for retrieval. Kong et al. [14] proposed a weakly supervised, end-to-end VPR framework that converts event bins into event spike tensor (EST) voxel grids [33], extracts features with a ResNet-34 backbone, and applies a VLAD layer with triplet ranking loss. While their dataset suite included night traverses, they only reported explicit day–night evaluation on the synthetic Oxford RobotCar and Carla simulator dataset.

Although Ensemble-Event-VPR [11] and Event-VPR [14] conducted ablations on windowing strategies, reconstruction methods, network backbones and loss functions, several aspects remain underexplored. In particular, prior work offers limited comparison of time surface reconstructions, event binning methods, and the performance of various SoTA VPR feature extractors when applied to event data, especially under day–night transitions. Our work addresses these gaps through an evaluation of these design decisions, demonstrating their significant impact on VPR performance under varying illumination conditions. Beyond analysing these factors individually, we show that combining them through score-level fusion further enhances robustness, as complementary representations contribute distinct but reinforcing cues under appearance change.

Other efforts explored hardware optimisations and deployment. One study applied a closed-loop bias controller for brightness-adaptive VPR, using SAD matching on event count reconstructions [34]. Most recently, Hines et al. [8] demonstrated event-based place recognition using a spike-based encoder with event count frames accumulated over a second, deployed on ultra-low-power neuromorphic hardware.

C. Ensemble Methods

Ensembles improve accuracy when base models are both accurate and diverse. If individual models have uncorrelated errors and perform better than random guessing, the probability that a majority vote is incorrect becomes lower than the individual error rate [35], [36].

In VPR, ensembles enhance robustness at different stages of the pipeline. Hierarchical Multi-Process Fusion applies late re-ranking across pipelines [37], while Patch-NetVLAD aggregates multi-scale patch features for viewpoint changes [38]. In neuromorphic settings, ensembles of spiking networks learn non-overlapping regions and combine spike responses at inference [39]. For event-based VPR, ensembling across time resolutions has shown gains [11]. This method identified that fusion based on the mean of similarity or difference matrices yielded the highest performance across product, median, min, max, trimmed mean and weighted strategies.

We extend these ideas by ensembling across reconstruction methods, feature extractors, and temporal resolutions, by using late score fusion to combine the varied similarity responses of these heterogeneous configurations. This extension enables complementary modalities to reinforce each other, improving matching consistency across changes in lighting, texture, and motion dynamics.

III. Methodology

Our event-based VPR pipeline, illustrated in Fig. 1, consists of five main stages. Section III-A details how asynchronous event data (x, y, t, p) is segmented into discrete bins. Section III-B describes the reconstruction of each event bin into a 2D frame using one of several methods, including learned methods and methods with and without polarity. Section III-C outlines how the reconstructed frames are processed by SoTA place recognition (VPR) models to extract global descriptors and compute pairwise descriptor similarities. Section III-D introduces a modified sequence matching algorithm with dynamic history length and z-score normalisation to improve robustness for longer sequences. Section III-E presents our general ensembling strategy, which aggregates similarity scores across reconstruction methods and feature extractors.

A. Event Slicing

Event cameras record a continuous stream of events where each event $e_j = (x_j, y_j, t_j, p_j)$ contains the pixel location (x_j, y_j), the timestamp t_j, and the polarity $p_j \in \{-1, +1\}$, representing a positive or negative brightness change.

Following standard practice in event-based vision [11], [32], [34], we evaluate two binning strategies: count-based binning and time-based binning. In the count-based binning strategy, each bin contains a fixed number of events, N. Specifically, the i-th bin $\mathbf{B}_i$ is defined as:

$$\mathbf{B}_i^N = \{e_j \mid j \in [iN, (i+1)N)\}. \tag{1}$$

In the time-based binning strategy, each bin spans a fixed duration temporal window Δt, starting from a reference time t_0. The i-th bin is given by:

$$\mathbf{B}_i^t = \{e_j \mid t_j \in [t_0 + i\Delta t,\ t_0 + (i+1)\Delta t)\}. \tag{2}$$

B. Event to Frame Reconstruction

Using the binned event data $\mathbf{B}_i = \{e_j\}$, we construct frame-based representations using several established reconstruction methods [1], [2].

The first method is a learned reconstruction model, E2VID [2], which produces an image estimate from a sequence of events using an encoder-decoder architecture with spatiotemporal fusion. We denote the reconstructed frame from bin $\mathbf{B}_i$ as $\mathbf{I}_i^{\text{E2VID}} = \mathcal{R}_{\text{E2VID}}(\mathbf{B}_i)$, where $\mathcal{R}_{\text{E2VID}}$ is the pretrained E2VID model applied to the events in bin $\mathbf{B}_i$.

For classical, non-learned reconstructions, we implement the event count reconstruction. In the polarity-aware variant, we maintain two separate channels, one for each polarity:

$$\mathbf{I}_i^{\text{count}}(x, y, c) = \sum_{e_j \in \mathbf{B}_i} \mathbb{1}(x_j = x \wedge y_j = y \wedge p_j = c), \tag{3}$$

where $c \in \{-1, +1\}$ denotes the polarity channel, and $\mathbb{1}(\cdot)$ is the indicator function. In the polarity-agnostic variant, the channels are summed into a single grayscale count image.

We also construct polarity-aware time surfaces [1] with exponential decay. For each pixel and polarity, we store the timestamp of the most recent event and compute

$$\mathbf{I}_i^{\text{TS}}(x, y, c) = \exp\left(-\frac{t_{\text{ref}} - t(x, y, c)}{\lambda}\right), \tag{4}$$

where $t(x, y, c)$ is the timestamp of the most recent event for this pixel location and polarity, λ is a decay constant, and $t_{\text{ref}} = t_{\text{end}} + \lambda$ is the adjusted frame reference time, where t_{end} is the timestamp of the last event in bin $\mathbf{B}_i$. The offset in t_{ref} ensures that even the most recent events are partially decayed, resulting in a smoother, more expressive time surface.

For each valid bin (i.e., containing at least one event), events are passed to the selected reconstruction method, which produces either a grayscale or multi-channel image. All reconstructions are computed at the native sensor resolution. For E2VID, hot pixel suppression is applied due to its sensitivity to structured noise, which can cause artifacts by interpreting hot pixels as meaningful features. Classical methods are more robust to such noise, as their simple aggregation tends to smooth over spurious events.

All classical reconstructed frames are normalised to produce image-like inputs by applying a per-pixel hyperbolic tangent to compress the dynamic range, followed by min–max scaling to the 8-bit grayscale range $[0, 255]$. For polarity-aware reconstructions, events are split into separate positive and negative arrays. Each is normalised using the same $\tanh$ and scaling procedure, then assigned to separate colour

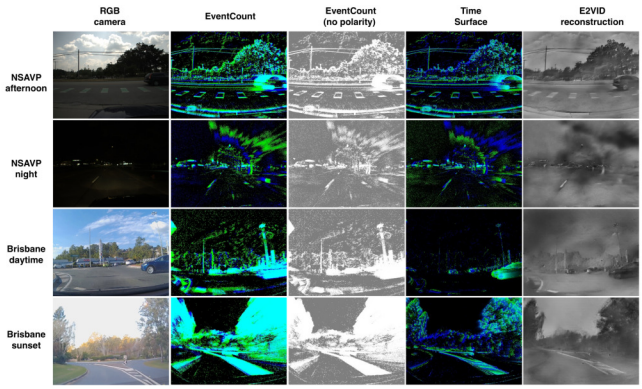

Fig. 2. Reconstructed event frames under varying illumination conditions. Column 1 shows the corresponding standard camera image. Columns 2–5 show event-based reconstructions: (2) Two-channel Event Count (polarity-separated), (3) Single-channel Event Count (polarity-combined), (4) Two-channel Time Surface (polarity-separated), and (5) Single-channel E2VID reconstruction (learned method). Rows 1–2 show afternoon and night conditions from a Gen4 Prophesee sensor in the NSAVP dataset [23], while rows 3–4 show daytime and sunset conditions from a DAVIS 346 sensor in the Brisbane Event dataset [11].

channels (e.g., green for positive, blue for negative), with the third channel set to zero. This results in a three-channel image compatible with standard learned feature extractors. An example of these reconstructions is shown in Figure 2, illustrating the information captured by each method.

C. Feature Extraction and Descriptor Matching

We use the reconstructed frames to generate feature descriptors using models originally trained for conventional frame-based VPR. We evaluate several feature extractors, including the classical model NetVLAD [3], a large-scale training approach CosPlace [4], a lightweight architecture MixVPR [5], and the current SoTA model, MegaLoc [6] (see Section II-A).

Let $\mathbf{I}_q^Q$ and $\mathbf{I}_j^R$ denote the q-th query frame and the j-th reference frame, respectively. A descriptor extractor $f(\cdot)$ maps each frame to a descriptor vector in a common embedding space: $\mathbf{d}_q^Q = f(\mathbf{I}_q^Q)$, $\mathbf{d}_j^R = f(\mathbf{I}_j^R)$, $\mathbf{d} \in \mathbb{R}^D$, where D is the descriptor dimension, Q is the size of the query set and R is the size of the reference set. We compute the similarity between each query and reference descriptor using the negative L2 distance, and define the similarity matrix $\mathbf{S} \in \mathbb{R}^{Q \times R}$ as:

$$\mathbf{S}[q, j] = -\|\mathbf{d}_q^Q - \mathbf{d}_j^R\|_2, \quad j_q = \arg\max_j \mathbf{S}[q, j], \quad (5)$$

where higher values of $\mathbf{S}[q, j]$ indicate greater similarity, and j_q denotes the predicted reference match for query q.

D. Sequence Matching

In trajectory matching tasks with repeat traversals, the similarity matrix typically exhibits a strong diagonal, with local warping due to speed changes, stops, or route deviations. SeqSLAM [21] takes advantage of this structure by applying temporal consistency across the similarity matrix. *SeqMatchNet* [22] adapted a simplified implementation of SeqSLAM by convolving a diagonal kernel across the similarity matrix, and scoring sequences based on their diagonal alignment. We use this implementation in our work.

To improve robustness near sequence boundaries and mitigate local biases caused by visual aliasing, we extend *SeqSLAM* by introducing a variable-length matching kernel and dual-axis normalisation. While *SeqSLAM* computes per-row diagonal scores using a fixed-size convolutional kernel, our method adapts the effective kernel size near sequence boundaries. Let R denote the number of reference frames and L the desired sequence length. For each query index q, we construct a submatrix $\mathbf{C}$ of the similarity matrix $\mathbf{S} \in \mathbb{R}^{Q \times R}$ by extracting up to L preceding query frames (the available history in a real time deployment setting), without exceeding the matrix boundary: $\mathbf{C}_q \in \mathbb{R}^{N \times R}$, where $N = \min(L, q+1)$.

We apply z-score normalisation to $\mathbf{C}$ across columns and then across rows, to reduce local bias in the similarity submatrix. To compute the sequence-matched score at each reference index j, we calculate the trace of the most recent $k \times k$ (where $k = \min(N, j+1)$) diagonal block:

$$\tilde{\mathbf{S}}[q, j] = \text{trace}\left(\mathbf{C}_{[N-k:N,\ j-k+1:j+1]}\right). \quad (6)$$

Using the trace is mathematically equivalent to convolving with a diagonal identity kernel, but allows for a more efficient implementation, as only the diagonal elements are computed rather than the full square kernel. This two-stage normalisation and variable-length kernel enable sequence matching to be applied even at the boundaries and improve robustness to local noise and visual aliasing. We apply this sequence matching procedure independently to each method configuration—i.e., each combination of reconstruction method and feature extractor—producing a set of sequence-matched similarity matrices that serve as inputs to our ensembling strategy.

E. Ensembling Strategies

Each $\tilde{\mathbf{S}}^{(m)} \in \mathbb{R}^{Q \times R}$ is the sequence-matched similarity matrix produced by the m-th method configuration. We treat these as individual ensemble members: $\mathcal{S} = \left\{\tilde{\mathbf{S}}^{(1)}, \ldots, \tilde{\mathbf{S}}^{(M)}\right\}$.

To aggregate predictions across configurations, we perform element-wise summation of all aligned similarity matrices, followed by maximum similarity selection as per the ensembling strategy from [11]:

$$\mathcal{E} = \sum_{m=1}^{M} \tilde{\mathbf{S}}^{(m)}, \quad \bar{j}_q = \arg\max_j \mathcal{E}[q, j]. \quad (7)$$

Unlike prior event-based VPR ensembling, which fuses predictions across temporal resolutions only [11], our strategy evaluates ensembles of reconstruction methods, ensembles of feature extractors and a combined ensemble of feature extractors, reconstruction and temporal ensembles. This broader fusion leads to significantly improved robustness under challenging appearance changes, including illumination shifts across the day.

IV. Experimental Setup

A. Datasets

In this paper, we evaluate our method on two outdoor datasets: BrisbaneEventVPR [11] and NSAVP [23], recorded during real-world vehicle traversals across day and night conditions.

BrisbaneEventVPR is a widely used benchmark dataset for event-based visual place recognition, introduced in 2020 and evaluated in several works [8], [10], [12], [13]. It consists of six traverses of an 8km route in Brisbane, Australia, recorded under various weather and illumination conditions, including daytime, night, sunrise, sunset, and morning on single and multilane roads. Data was captured using a 346×260 resolution iniVation DAVIS 346 event camera mounted on the vehicle's windshield, along with GPS and 1080p RGB video from a consumer-grade camera.

Novel Sensors for Autonomous Vehicle Perception (NSAVP) is a more recent dataset collected in Michigan, USA, with a focus on evaluating modern sensor technologies in diverse and realistic environments. It comprises two distinct routes (8.3 km and 8.6 km) traversed in both directions under varying lighting conditions. Data was collected using a 1280×720 resolution Prophesee Gen4 HD event sensor and high-precision 200 Hz ground-truth from an Applanix POS-LV 420 system, yielding a higher-resolution, roof-mounted configuration from a different manufacturer than the DAVIS 346 used in BrisbaneEventVPR. Compared to BrisbaneEventVPR, NSAVP also introduces greater environmental diversity, including suburban neighbourhoods and dense urban scenes with multi-story buildings, and offers higher fidelity sensing for large-scale benchmarking.

To ensure fair comparison across illumination changes, we evaluate five day-to-day and five day-to-night pairs. On BrisbaneEventVPR [11], we use `sunset1` as the reference against `sunrise`, `morning`, `daytime`, `sunset2`, and `night`. On NSAVP [23], we adopt the notation `R0-[F/R]-[A/S/N]` (Route 0, Forward/Reverse, Afternoon/Sunset/Night) to evaluate five pairs: FA0 vs. FS0, FA0 vs. FN0, FN0 vs. FS0, RA0 vs. RN0, and RS0 vs. RN0. Datasets such as DDD20 [40] and NYC Events [41] were excluded as they lack explicit repeat traversals.

B. Implementation Details

We adopt *Recall@1* as our primary evaluation metric, following standard practice in the VPR literature. This metric measures the ratio of correct top-1 matches to the total number of query frames.

To evaluate event binning strategies, we experiment with both count-based and time-based binning. All subsequent experiments use time-based binning, which consistently yields stronger performance. For each reconstructed frame, we linearly interpolate available GPS data to determine its ground-truth position. We extract features using the open-source *VPR-Methods-Evaluation* repository, applying a 25-metre tolerance to define correct matches [42]. This tolerance is smaller than previous methods [11], [23] for the Brisbane Event dataset, which used 70m, so our reported recall results are more conservative. No feature extractor or reconstruction model is retrained; all models are used as released. For comparison to LENS [8], we use 1 second temporal windows with event count reconstruction, a ground truth tolerance of +/- 3 seconds and a sequence length of 10.

For ensemble evaluation, we follow the protocol of [11], which requires temporal alignment between ensemble members. We consider time resolutions of 0.1, 0.25, 0.5, and 1.0 seconds, sampled at 1 Hz to match the coarsest resolution. Each single ensemble strategy, based on time resolution, reconstruction method, or feature extractor, is evaluated across 480 configurations ($4 \times 4 \times 3 \times 10$). These configurations are generated by combining 4 feature extractors, 4 reconstruction methods, 4 time resolutions, 3 sequence lengths (10, 20, 30), and 10 reference-query traverse pairs. We also evaluate a configuration that aggregates across all combinations of the individual methods (a total of 64) and report on these results as the combined ensembles.

V. Results and Discussion

This section presents and compares the performance of our event-based VPR system. Section V-A contains our evaluation of each ensemble strategy across reconstruction methods and feature extractors. Section V-B investigates the effect of our modified sequence matching approach (Section III-D) compared to the original SeqSLAM methodology. The influence of event-binning strategies on recognition accuracy and robustness is then examined in Section V-C. Finally, Section V-D compares key design choices in event-based VPR, including binning strategies, polarity, reconstruction methods, and feature extractors, critically evaluating the impact of each component on performance.

A. Event VPR via Ensembling

The effectiveness of the proposed ensemble strategies across reconstruction methods, feature extractors and the combination of feature extractors, reconstruction and temporal ensembles was evaluated against two baselines: (i) the best-performing individual ensemble member and (ii) the temporal resolution only ensemble from [11]. The results exhibit a bimodal distribution, reflecting the expected performance gap between day–day and day–night pairs due to significant domain shifts. To highlight these trends, results are grouped into day and night conditions. As shown in Figure 3, across 10 diverse traverse combinations, the proposed ensembles consistently outperformed both baselines.

Table I quantifies comparisons to existing methods along with the relative improvement over the temporal ensemble baseline [11]. Under severe illumination variation on BrisbaneEventVPR (day to night), the proposed ensembling improves average recall@1 (AR@1) from 1% for LENS [8] and 37.41% for Temporal Ensembles [11] to 44.16% and 66.39% for the ensemble of feature extractors and combined ensembles. On NSAVP day–night, the feature extractor and combined ensembles improve AR@1 from 28.40% to 32.18% and to 42.55%, respectively. Therefore, the

TABLE I

AVERAGE RECALL@1 AND GAIN OVER TEMPORAL ENSEMBLE [11].

Condition	Method	AR@1 (%)	Gain (%)
D–D (Brisbane Event)	LENS [8]	73.00	-11.23
	Temporal Window [11]	82.24	—
	Reconstruction	82.98	0.91
	Feature Extractor	83.16	1.13
	Combined Ensemble	84.72	3.02
D–N (Brisbane Event)	LENS [8]	1.00	-97.33
	Temporal Window [11]	37.41	—
	Reconstruction	42.06	12.44
	Feature Extractor	44.16	18.06
	Combined Ensemble	66.39	77.48
D–D (NSAVP)	Temporal Window [11]	56.88	—
	Reconstruction	58.09	2.13
	Feature Extractor	58.31	2.51
	Combined Ensemble	61.79	8.63
D–N (NSAVP)	Temporal Window [11]	28.40	—
	Reconstruction	30.47	7.31
	Feature Extractor	32.18	13.34
	Combined Ensemble	42.55	49.83

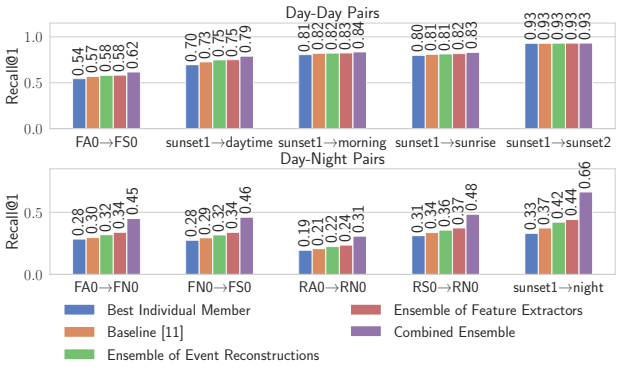

Fig. 3. Average Recall@1 (AR@1) for each reference–query pair across ensembling strategies and best individual method, shown separately for day and night conditions. Results are averaged over sequence lengths of 10, 20, and 30 at 1 Hz sampling. The combined ensemble aggregates predictions from varied feature extractors, event-to-frame reconstructions and temporal resolutions.

combined ensembles have a relative improvement of up to 77% in day-night transitions. Gains in daytime conditions are smaller but consistent, with up to 3.02% relative gain on BrisbaneEventVPR and 8.63% relative gain on NSAVP. Additionally, figure 4 illustrates the fusion of predictions from multiple event-to-frame reconstructions, yielding a higher recall than the best individual method and highlights the robustness gained through aggregation.

B. Sequence Matching

Table II compares the original SeqSLAM implementation (via SeqMatchNet) to the proposed sequence matching variant, which introduces an adaptive matching kernel and additional score normalisation. Results, averaged over 800 combinations of routes, reconstruction methods, feature extractors, and temporal resolutions, indicate improved Recall@1 at longer sequence lengths. At sequence lengths 20 and 30, the proposed method achieves Recall@1 gains of 0.0312 (p = 0.0554) and 0.0403 (p = 0.0096), respectively. Performance at length 10

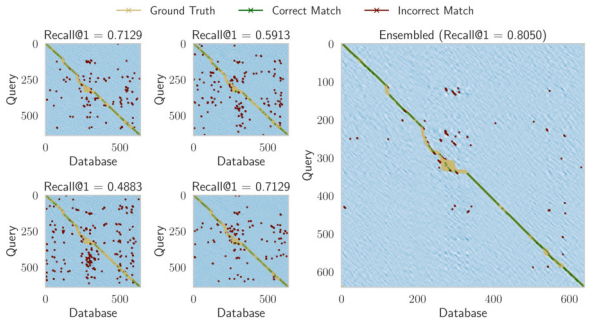

Fig. 4. Similarity matrices for a sunset–daytime pair with Megaloc feature extractor and varied reconstructions. Columns 1 and 2 show results from E2VID, EventCount, EventCount (no polarity) and time surface reconstructions. Column 3 shows the ensemble of reconstruction methods. Ground-truth, correct, and incorrect matches are annotated.

TABLE II

AR@1 ACROSS SEQUENCE LENGTHS AND P-VALUES FROM PAIRED T-TEST

Seq Len	SeqSLAM	Ours	p-value
10	0.3473	0.3410	0.6874
20	0.4301	**0.4613**	0.0554
30	0.4593	**0.4996**	0.0096

is similar across both methods (p = 0.6874). The highest average Recall@1 is observed at a sequence length of 30 using the proposed approach.

C. Event Binning Strategies

Two common event binning strategies are evaluated: fixed-count (B^N) and fixed-duration (B^t). As shown in Figure 5, fixed-duration binning consistently yields better Recall@1 in urban driving scenarios with standard VPR feature extractors. The figure also illustrates the trade-off between performance and binning resolution, where smaller bins improve accuracy but increase computational cost due to a higher number of reference and query frames. While a similar comparison was presented in [11], only minimal performance differences were reported, likely due to the limited temporal resolution range (0.02–0.14 seconds) used. In contrast, the broader range considered here (0.1–1.0 seconds) captures motion dynamics more effectively. To further investigate this, Figure 6 examines the distance travelled within individual bins across methods and resolutions, showing that time-based binning results in lower variance relative to the mean distance, supporting more consistent place representations across repeated traversals with varied illumination.

D. Reconstructions and Feature Extractors for Event VPR

Figure 7 presents Recall@1 performance across individual event-to-frame reconstruction methods, averaged over five temporal resolutions.

Under daytime conditions, the learned reconstruction method E2VID achieves the highest performance, with a mean Recall@1 of 63.3%, outperforming all classical reconstructions by 7–18%. This is consistent with its supervised training on well-lit scenes. In contrast, under nighttime conditions, E2VID records the lowest performance (Recall@1 of 8.6%),

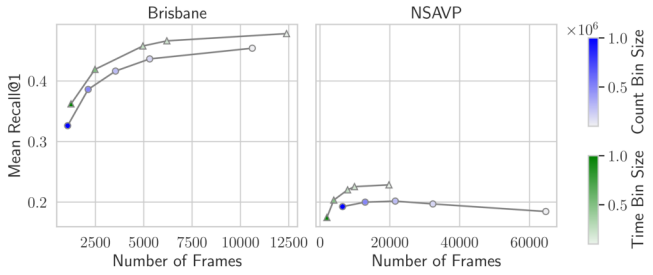

Fig. 5. Average Recall@1 versus the number of reference and query frames, evaluated across two binning strategies and varying event slice resolutions. Time-based binning consistently yields the highest recall, with performance improving at higher resolutions (i.e., smaller bin sizes). The total number of frames serves as a proxy for computational cost, as both event-to-frame reconstruction and feature extraction scale linearly with frame count. Consequently, higher-resolution bins incur greater compute cost.

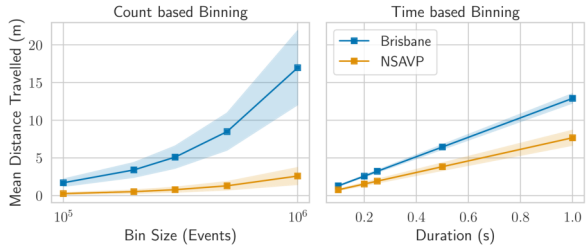

Fig. 6. Distance travelled per frame averaged across repeat traverses for each binning type and resolution. Time-based binning shows the lowest variance in mean distance, indicating greater consistency for place representation.

approximately 3–8% lower than classical methods, which demonstrate greater robustness in low-light environments.

Two variants of event count reconstruction are also evaluated: with and without polarity. Including polarity improves daytime performance by 6.8% Recall@1 (52.2% vs. 45.4%), while the difference at night is minimal (11.8% vs. 11.0%). Figure 8 shows VPR performance across different feature extractors. The observed trends are consistent with those reported in standard camera-based VPR benchmarks. Earlier methods, such as CosPlace and NetVLAD, achieve mean Recall@1 scores of 24–25%, while more recent approaches like MixVPR and MegaLoc reach 38–42%. MegaLoc, the current state-of-the-art, performs particularly well at night, achieving 16.9% Recall@1—at least 5% higher than all other evaluated methods.

E. Computational Tradeoff

Table III reports the end-to-end runtime for a single query and AR@1 for different ensemble configurations, illustrating the accuracy versus compute tradeoff. The temporal-window [11] and reconstruction ensembles improve AR@1 by around 6–8% over the single model; while the feature extraction ensemble is most suitable for real-time application running at 13Hz with 10% increase in AR@1. The combined ensemble achieves the highest AR@1 but with substantially higher latency, making it most suitable for sparse global relocalisation or offline evaluation. A complementary correlation analysis of similarity matrices across reconstruction methods and feature extractors (included in our public code repository)

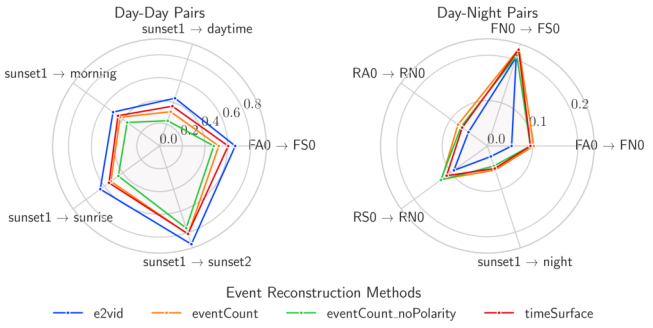

Fig. 7. Recall@1 across individual reconstruction methods, separated by day and night conditions, evaluated without sequence matching. These results show that the learned method E2VID performs best during the day but is detrimental at night. Encoding polarity in a separate channel also improves performance in daytime conditions for classical reconstructions.

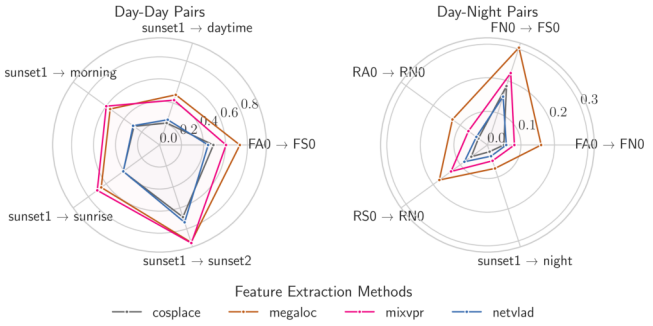

Fig. 8. Recall@1 across individual feature extraction methods, separated day and night conditions, evaluated using single-frame VPR without sequence matching. Event-based VPR shows similar performance trends to conventional frame-based methods, with Megaloc achieving the highest overall recall.

shows that some representations, such as event count and event count without polarity, are strongly correlated, suggesting that future work could prioritise more complementary combinations to achieve better accuracy–runtime trade-offs.

VI. Conclusions and Future Work

Despite growing interest in event-based place recognition, the full capabilities of event cameras remain underutilised. With the emergence of new long-term driving datasets, we present an ensemble-based method that fuses multiple VPR feature extractors, event reconstruction and temporal resolutions, consistently outperforming prior baselines. Beyond performance gains, our work addresses a gap in the event VPR literature by evaluating key design choices. We show that time-based binning outperforms count-based alternatives, and that the learned E2VID reconstruction, while effective during the day, performs poorly at night. Furthermore, we find that VPR performance trends across feature extractors mirror those observed in conventional frame-based settings.

While our study includes a diverse set of feature extractors, reconstructions, and temporal strategies, further gains may be achieved by learning event-specific feature extractors and developing reconstructions tailored for VPR in night conditions. Notably, none of the deep learning-based feature extractors used were trained on event data, suggesting a

TABLE III
RUNTIME VS PERFORMANCE ANALYSIS

Ensemble Type	Runtime (ms)	AR@1 (%)
None	47.63	47.07
Temporal Windows [11]	190.5	52.95
Event Reconstructions	190.5	54.61
VPR Feature Extractors	78.38	56.64
Combined Ensemble	1254.0	64.37

promising direction for developing event-native architectures. Future work could also explore real-world deployment of our method, expanding the utility of event cameras in challenging environments.

REFERENCES

[1] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A Hierarchy of Event-Based Time-Surfaces for Pattern Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1346–1359, 2017.

[2] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1964–1980, 2019.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, 2018.

[4] G. Berton, C. Masone, and B. Caputo, "Rethinking Visual Geo-Localization for Large-Scale Applications," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4878–4888.

[5] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "MixVPR: Feature Mixing for Visual Place Recognition," in *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2997–3006.

[6] G. Berton and C. Masone, "Megaloc: One retrieval to place them all," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2025, pp. 2861–2867.

[7] D. Falanga, K. Kleber, and D. Scaramuzza, "Dynamic obstacle avoidance for quadrotors with event cameras," *Sci. Robot.*, vol. 5, no. 40, Mar. 2020.

[8] A. D. Hines, M. Milford, and T. Fischer, "A compact neuromorphic system for ultra–energy-efficient, on-device robot localization," *Sci. Robot.*, vol. 10, no. 103, p. eads3968, Jun. 2025.

[9] G. Gallego *et al.*, "Event-Based Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, 2022.

[10] A. J. Lee and A. Kim, "EventVLAD: Visual Place Recognition with Reconstructed Edges from Event Cameras," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2021, pp. 2247–2252.

[11] T. Fischer and M. Milford, "Event-Based Visual Place Recognition With Ensembles of Temporal Windows," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6924–6931, 2020.

[12] H. Lee and H. Hwang, "Ev-ReconNet: Visual Place Recognition Using Event Camera With Spiking Neural Networks," *IEEE Sens. J.*, vol. 23, no. 17, pp. 20 390–20 399, 2023.

[13] T. Fischer and M. Milford, "How Many Events Do You Need? Event-Based Visual Place Recognition Using Sparse But Varying Pixels," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 12 275–12 282, 2022.

[14] D. Kong *et al.*, "Event-VPR: End-to-End Weakly Supervised Deep Network Architecture for Visual Place Recognition Using Event-Based Vision Sensor," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–18, 2022.

[15] C. Masone and B. Caputo, "A Survey on Deep Visual Place Recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021.

[16] S. Lowry *et al.*, "Visual Place Recognition: A Survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, 2016.

[17] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 113, p. 107760, 2021.

[18] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual Place Recognition: A Tutorial," *IEEE Robotics & Automation Magazine*, vol. 31, no. 3, pp. 139–153, 2024.

[19] S. Garg, T. Fischer, and M. Milford, "Where Is Your Place, Visual Place Recognition?" in *Int. Joint Conf. Artif. Intell.*, vol. 5, 2021, pp. 4416–4425.

[20] A. Lengyel, S. Garg, M. Milford, and J. C. van Gemert, "Zero-Shot Day-Night Domain Adaptation With a Physics Prior," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4399–4409.

[21] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE Int. Conf. Robot. Autom.*, 2012, pp. 1643–1649.

[22] S. Garg, M. Vankadari, and M. Milford, "SeqMatchNet: Contrastive Learning with Sequence Matching for Place Recognition & Relocalization," in *Proc. Conf. Robot. Learn.*, 2022, pp. 429–443.

[23] S. Carmichael, A. Buchan, M. Ramanagopal, R. Ravi, R. Vasudevan, and K. A. Skinner, "Dataset and Benchmark: Novel Sensors for Autonomous Vehicle Perception," *Int. J. Robot. Res.*, vol. 44, no. 3, pp. 355–365, 2025.

[24] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "The Revisiting Problem in Simultaneous Localization and Mapping: A Survey on Visual Loop Closure Detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, 2022.

[25] Y. Zhang, P. Shi, and J. Li, "LiDAR-Based Place Recognition For Autonomous Driving: A Survey," *ACM Comput. Surv.*, vol. 57, no. 4, pp. 106:1–106:36, 2024.

[26] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter Wave FMCW RADARs for Perception, Recognition and Localization in Automotive Applications: A Survey," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 533–555, Sep. 2022.

[27] A. Melekhin, D. A. Yudin, I. Petryashin, and V. Bezuglyj, "MSSPlace: Multi-Sensor Place Recognition With Visual and Text Semantics," *IEEE Access*, vol. 13, pp. 177 098–177 110, 2025.

[28] J. Komorowski, M. Wysoczańska, and T. Trzcinski, "MinkLoc++: Lidar and Monocular Image Fusion for Place Recognition," in *Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[29] K. Hou, D. Kong, J. Jiang, H. Zhuang, X. Huang, and Z. Fang, "FE-Fusion-VPR: Attention-Based Multi-Scale Network Architecture for Visual Place Recognition by Fusing Frames and Events," *IEEE Robot. Autom. Lett.*, vol. 8, no. 6, pp. 3526–3533, 2023.

[30] H. Wang *et al.*, "Event Camera Meets Resource-Aware Mobile Computing: Abstraction, Algorithm, Acceleration, Application," 2025, arXiv:2503.22943 [cs].

[31] B. Chakravarthi, A. A. Verma, K. Daniilidis, C. Fermuller, and Y. Yang, "Recent Event Camera Innovations: A Survey," in *Eur. Conf. Comput. Vis. Workshops*, 2025, pp. 342–376.

[32] M. Milford, H. Kim, S. Leutenegger, and A. Davison, "Towards Visual SLAM with Event-based Cameras," *The problem of mobile sensors workshop in conjunction with RSS*, 2015.

[33] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-End Learning of Representations for Asynchronous Event-Based Data," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5633–5643.

[34] G. B. Nair, M. Milford, and T. Fischer, "Enhancing Visual Place Recognition via Fast and Slow Adaptive Biasing in Event Cameras," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 3356–3363.

[35] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, 1990.

[36] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[37] S. Hausler and M. Milford, "Hierarchical Multi-Process Fusion for Visual Place Recognition," in *IEEE Int. Conf. Robot. Autom.*, 2020, pp. 3327–3333.

[38] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 141–14 152.

[39] S. Hussaini, M. Milford, and T. Fischer, "Ensembles of Compact, Region-specific & Regularized Spiking Neural Networks for Scalable Place Recognition," in *IEEE Int. Conf. Robot. Autom.*, 2023, pp. 4200–4207.

[40] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "DDD20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *IEEE Int. Conf. Intell. Transp. Syst.*, 2020.

[41] T. Pan, J. He, C. Chen, Y. Li, and C. Feng, "Nyc-event-vpr: A large-scale high-resolution event-based visual place recognition dataset in dense urban environments," in *IEEE Int. Conf. Robot. Autom.*, 2025, pp. 4657–4664.

[42] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11 080–11 090.