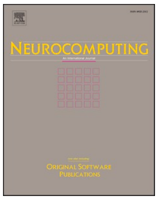
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey paper

Visual SLAM in the era of heterogeneous intelligence coexistence: A survey

Sa Su [a,b,c,1], Xu He [a,b,c,1,2], Xiaolin Meng [a,b,c,*], Youdong Zhang [a,b,c], Wenxuan Yin [a,b,c], Lingfei Mo [a,c], Fangwen Yu [d]

[a] The School of Instrument Science and Engineering, Southeast University, Nanjing, China
[b] The China-UK Centre on Intelligent Mobility, Southeast University, Nanjing, China
[c] State Key Laboratory of Comprehensive PNT Network and Equipment Technology, Southeast University, Nanjing, China
[d] The Center for Brain Inspired Computing Research, Department of Precision Instrument, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
VSLAM
heterogeneous intelligence
computing paradigms
framework
challenges and opportunities

ABSTRACT

Towards the era of Heterogeneous Intelligence (HI) coexistence, this paper reviews the latest progress of Visual Simultaneous Localization and Mapping (VSLAM) and explores the pathway of multiple HI integration-driven VSLAM systems. This work analyzes over 220 selected publications, with a literature cut-off date of September 2025, with papers distributed across the evolution of frontend Visual Odometry (VO), Loop Closure Detection (LCD), backend optimization and mapping. Moreover, it also discusses the support of heterogeneous hardware, including state-of-the-art sensors and processors. Finally, it analyzes the challenges and opportunities, proposes a novel VSLAM framework from the view of HI integration, and provides forward-looking suggestions. This study indicates that the cross-paradigm HI integration has the potential to transform current VSLAM technologies from "tool-oriented" to "cognition-oriented," providing new ideas and pathways for the next-generation VSLAM development.

1. Introduction

1.1. Background

Nowadays, the rise of the third wave of Artificial Intelligence (AI) is an undeniable reality. Advances in Machine Learning (ML) and Deep Learning (DL), coupled with explosive growth in computing power and data availability, have spurred widespread AI applications across various domains [1]. Current DL, however, is data-driven and dominant in specific tasks, yet far from human-level general intelligence. Towards the target of Artificial General Intelligence (AGI) development, brain-inspired intelligence is increasingly recognized as a pivotal approach to bridge this gap [2]. Meanwhile, the global push for quantum intelligence highlights the hybrid nature of the Heterogeneous Intelligence (HI) coexistence era.

Navigation has always been an important starting and end point of machine intelligence, underpinning the emergence of embodied intelligence [3,4]. Since localization and mapping are essential tasks for navigation, Simultaneous Localization and Mapping (SLAM) has been studied for over three decades. Visual SLAM (VSLAM), a key implementation of SLAM, consists of five main components: data processing, Visual Odometry (VO), Loop Closure Detection (LCD), backend

Abbreviations: AI, Artificial Intelligence; ML, Artificial Intelligence.; DL, Deep Learning; AGI, Artificial General Intelligence; HI, Heterogeneous Intelligence; SLAM, Simultaneous Localization and Mapping; VSLAM, Visual SLAM; VO, Visual Odometry; LCD, Loop Closure Detection; CANNs, Continuous Attractor Neural Networks; VPR, Visual Place Recognition; SNNs, Spiking Neural Networks; CPUs, Central Processing Units; GPUs, Graphics Processing Units; NeRFs, Neural Radiance Fields; GS, Gaussian Splatting; PnP, Perspective-n-Point; BA, Bundle Adjustment; DSO, Direct Sparse Odometry; SVO, Semidirect Visual Odometry; CNNs, Convolutional Neural Networks; GNN, Graph Neural Networks; IMUs, Inertial Measurement Units; VIO, Visual-Inertial Odometry; HDR, High Dynamic Range; FPS, Frame Per Second; RMSE, Root Mean Square Error; BoW, Bag-of-Words; RANSAC, Random Sample Consensus; VLM, Visual-and-Language Model; STDP, Spike-Timing-Dependent Plasticity; MCN, Mini-Column Network; ESNs, Echo State Networks; NLS, Nonlinear Least Squares; DVS, Dynamic Vision Sensor; ATIS, Asynchronous Time-based Image Sensor; DAVIS, Dynamic and Active-pixel Vision Sensor; BPUs, Brain Processing Units; IPUs, Intelligent Processing Units; IBM, International Business Machine; VLN, Vision-and-Language Navigation.

[*] Corresponding author at: The School of Instrument Science and Engineering, Southeast University, Nanjing, China.
E-mail address: xiaolin_meng@seu.edu.cn (X. Meng).

[1] Xu He & Sa Su contributed equally,

[2] Xu He is the co-first author.

https://doi.org/10.1016/j.neucom.2025.132458
Received 12 July 2025; Received in revised form 1 December 2025; Accepted 17 December 2025
Available online 18 December 2025

optimization, and mapping [5]. Their joint optimization and simultaneous operation form a complete SLAM system.

Traditional VSLAM builds on mathematical paradigms, with a robust theoretical basis, yielding milestone solutions such as ORB-SLAM [6]. However, it struggles in complex settings due to issues like feature extraction failure, scale drift, and limited scene understanding [7]. Data-driven AI paradigms greatly mitigate these issues and excel in VO, LCD, and mapping [8]. Examples include UnDeepVO [9] for DL-based monocular VO, Airloop [10] for lifelong learning-based LCD, and DS-SLAM [11] for dynamic semantic mapping.

Brain-inspired SLAM, emerging beyond classical methods, emulates the brain's navigation mechanisms via neurodynamic models [12]. Advances like RatSLAM [13] and NeuroSLAM [14] use Continuous Attractor Neural Networks (CANNs) to mimic the brain's path integration logic, rebuilding the backend optimization while fitting VSLAM frameworks. Progress also includes brain-inspired Visual Place Recognition (VPR) [15–18] and ANN2SNN methods such as SpikingJelly [19] for converting DL to Spiking Neural Networks (SNNs).

Traditional VSLAM systems employ mature visual sensors, providing practical solutions, such as the ORB-SLAM series [6,20,21]. AI cameras also warrant attention [22]. Recently, bionic cameras have been used to enhance robustness in low-texture environments [23]. Moreover, neuromorphic cameras [24] excel in challenging settings (e.g., motion blur and latency), offering high efficiency, low latency and power consumption, expanding VSLAM capabilities [25,26].

Despite advances in Central Processing Units (CPUs), Graphics Processing Units (GPUs) and AI processors, von Neumann architectures still lag behind the brain's spatiotemporal representation and generalization. Drawing on neuroscience insights, the state-of-the-art neuromorphic chips, such as the Tianjic [27] and SpiNNaker2 [28], support a hybrid HI integration of AI and brain-inspired paradigms. This represents the chip designers' response in the era of HI coexistence.

1.2. The connotation of the "Era of HI Coexistence" and its manifestation to VSLAM

As known, machine's intelligence is realized through both software and hardware. Software's intelligence relies on advanced algorithms and computational paradigms, while hardware provides perceptual information, computing resources, serving as the carrier. Today's era is marked by the coexistence of multiple paradigms, including mathematical computing, AI, brain-inspired intelligence, and even quantum intelligence. For brevity, this paper terms this the "Era of Coexistence of HI." In this context, we firstly need to determine the manifestations of these different intelligence paradigms with obviously heterogeneous natures (collectively referred to as HI) in VSLAM research.

On the one hand, although multiple HI paradigms coexist in the current era, not all VSLAM research has integrated more than one HI paradigm. This means that a VSLAM system can benefit from a single paradigm's contributions but may also face certain challenges. *On the other hand*, a VSLAM system can also benefit from the integration of different HI paradigms. Specifically, within the VSLAM framework, components can be implemented by diverse computing paradigms. Furthermore, a VSLAM system can configure heterogeneous sensors and processors for complementary environmental perception and performance optimization. In other words, a VSLAM system can be improved through cross-paradigm integration (Fig. 1).

1.3. Motivation and innovation declaration

In the above two manifestations, the former is common, while the latter is emerging. Therefore, this unique backdrop offers the VSLAM community significant opportunities for innovation and may even prompt a transformation in its research paradigm.

However, no existing work has systematically surveyed VSLAM technology in the era of HI coexistence, explored cross-paradigm HI integration pathways, or analyzes challenges and opportunities. This gap forms our motivation. Therefore, this paper not only provides a comprehensive review of VSLAM advances benefited from different HI paradigms, but also dissects the trend of cross-paradigm HI integration-empowered VSLAM. It analyzes over 220 selected publications, with a literature cut-off date of September 2025, with papers distributed across the evolution of VO, LCD, backend optimization and mapping.

To clarify the unique positioning and innovation of this work, a comparison with representative surveys is detailed in Table 1. The *contributions* are as follows:

1) This paper, towards the HI coexistence era, systematically dissects the VSLAM progress from a multi-dimensional perspective and carries out in-depth discussions and analyses.
2) This paper summarizes the roadmap of VSLAM systems empowered by cross-paradigm HI integration and proposes a unified framework for VSLAM system (See Section VII).
3) This paper prospectively analyzes the opportunities and challenges of VSLAM technology in the era of HI coexistence, offering the forward-looking perspectives and suggestions.

Note. Since quantum intelligence is still in its germination stage, it is

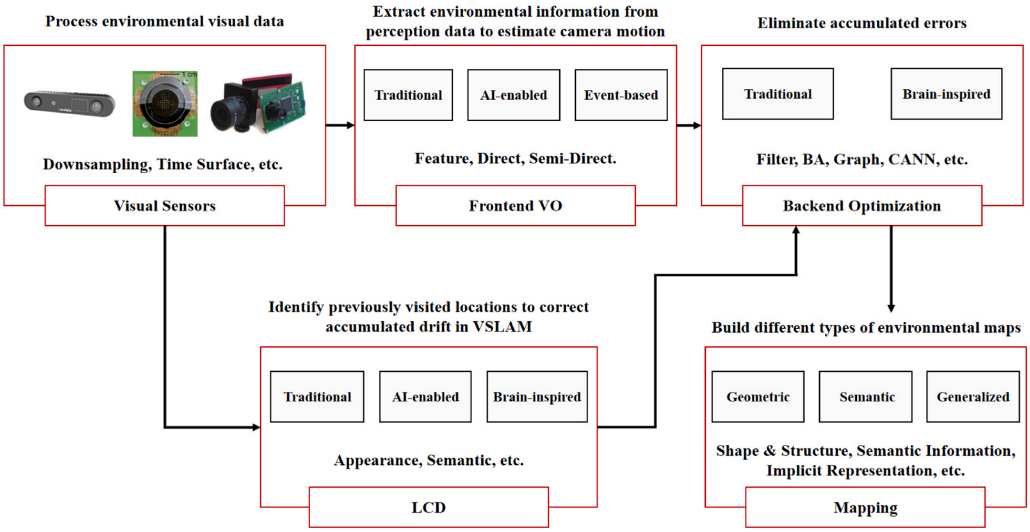

Process environmental visual data
Downsampling, Time Surface, etc.
Visual Sensors
Extract environmental information from perception data to estimate camera motion
Traditional
AI-enabled
Event-based
Feature, Direct, Semi-Direct.
Frontend VO
Eliminate accumulated errors
Traditional
Brain-inspired
Filter, BA, Graph, CANN, etc.
Backend Optimization
Identify previously visited locations to correct accumulated drift in VSLAM
Traditional
AI-enabled
Brain-inspired
Appearance, Semantic, etc.
LCD
Build different types of environmental maps
Geometric
Semantic
Generalized
Shape & Structure, Semantic Information, Implicit Representation, etc.
Mapping

Fig. 1. The VSLAM framework.

Table 1
Comparison with existing representative surveys.

Works	Year	Main Focus / Scope	Distinction from Our Work
[7]	2016	A review charting SLAM evolution towards the “Robust-Perception Age.”	■ Its focus is on robust perception in SLAM, with a different purpose from ours. ■ Our work covers a broader and more comprehensive scope.
[29]	2017	A review of VSLAM advances within a specific timeframe (2010–2016).	■ Review from both technical and historical points of views. ■ Its main effort is devoted to the review of traditional methods.
[30]	2022	A comprehensive survey of state-of-the-art on VSLAM before 2022	■ Its focus is on comprehensive review of feature-based VSLAM with simulations. ■ Its scope includes traditional and DL methods, forming a subset of this work.
[31]	2019	A problem-specific review focused on VSLAM for dynamic environments.	■ Its main focus is on a specific issue in VSLAM applications. ■ Our work is a comprehensive survey with a forward-looking perspective.
[32]	2024	A review on the impacts of Neural Radiance Fields (NeRFs) and 3D Gaussian Splatting (GS) to SLAM.	■ Its focus is on specific, groundbreaking technologies’ contribution to SLAM. ■ Our work is a comprehensive survey with a forward-looking perspective.
[8] [33]	2024 2023	Two focused surveys on the application of DL techniques across the VSLAM pipeline.	■ Their purposes and scopes are dedicated on pure DL-based VSLAM research. ■ Our work covers a broader and more comprehensive scope.
[24] [34]	2024 2022	Two specialized surveys on event-based vision and VSLAM.	■ Their purposes and scopes are dedicated on pure event-based VSLAM research. ■ Our work covers a broader and more comprehensive scope.
[35] [36] [37]	2022 2021 2015	Three reviews of VPR/LCD, a specific module within the VSLAM framework.	■ Their scopes are component-specific surveys, forming a subset of this work. ■ Our work covers a broader and more comprehensive scope.

not included in the scope of this study.

1.4. Outlines

Sections II to V systematically review and analyze the key advances. Section VI summarizes hardware support for system-level VSLAM development. Section VII proposes a unified VSLAM framework suitable for HI integration and discusses opportunities and challenges. Section VIII offers forward-looking perspectives and concludes the paper.

2. VO progress

The VO task involves extracting environmental information from images and estimating camera motion between adjacent images based on the geometric relationship between the camera and spatial points. This section systematically reviews VO progress, with corresponding discussion and analysis.

2.1. Traditional VO

Traditional VO methods are often categorized into feature-based and direct methods. Feature-based methods detect salient points using handcrafted descriptors, compute their matching relationships, and estimate camera motion via the Perspective-n-Point (PnP) or Bundle Adjustment (BA) methods [38]. The ORB-SLAM series, utilizing point feature-based VO, are widely recognized in the VSLAM community. In addition, line and edge features are stable and easily extracted in structured settings, reduce complexity and can be combined with point features [39–41]. For example, StructSLAM [42] employs architectural lines to reduce drift error. MonoSLAM [43] addresses tracking failure in texture-less environments by extracting points, lines, and vanishing points for feature complementarity. Cai et al. [44] review common handcrafted descriptors (e.g., point, line, edge, corner, and region features).

Feature-based methods dominate VO but discard most image information. In contrast, direct methods estimate camera motion by minimizing photometric errors using pixel grayscale information from two frames, relying on grayscale invariance and nonlinear optimization. Direct methods can be categorized into sparse, semi-dense, and dense forms based on the number of pixels used. For example, DTAM [45], the progenitor of direct methods, generates dense maps and camera poses by aligning the entire image. LSD-SLAM [46] is a typical semi-dense direct method. Direct Sparse Odometry (DSO) [47] is a sparse direct method, estimating camera motion by minimizing sparse photometric errors for efficient computation. FD-SLAM [48] is a dense method, using frame-to-model to align input frames with active submaps via joint optimization of geometric and photometric errors. Beyond the above, semi-direct VO solutions combine the strengths of feature-based and direct methods to balance computational efficiency and accuracy [49]. Semidirect Visual Odometry (SVO) [50] is a typical example.

2.2. AI-enabled VO

Unlike traditional methods, AI-enabled VO can learn robust representations (e.g., depth, optical flow, feature points) automatically, and show promise in overcoming limitations of handcrafted features and environmental adaptivity [8,51]. This has led to renovation of the VSLAM frontend. Following AlexNet’s breakthrough in ImageNet [52], the powerful ability of Convolutional Neural Networks (CNNs) led directly to the creation of PoseNet [53], marking the beginning of data-driven VO paradigms.

Supervised learning drives end-to-end training with ground-truth values. For example, DeepVO [54] uses a supervised CNN-LSTM network to capture spatiotemporal dependencies in image sequences for predicting VO trajectories. DytanVO [55] introduces a dynamic perception module optimized via curriculum learning for dynamic object segmentation and pose estimation. GANVO [56] enhances pose estimation accuracy with optical flow consistency constraints in its discriminator, while the generator produces depth maps, creating a mutually reinforcing optimization mechanism. STDN-VO [57] mimics the human visual system’s dual-stream mechanism, extracting spatial and temporal features with different networks and fusing them to predict poses, significantly enhancing VO accuracy.

Recently, unsupervised and self-supervised VO methods have gained prominence due to the scarcity of labeled data. Unsupervised methods, pioneered by works like Zhou et al. [58] and GeoNet [59], use inherent geometric constraints in multi-view imagery as a supervisory signal, eliminating the need for external labels. Wang et al. [60] resolve scale ambiguity in monocular VO via joint training of depth, optical flow, and scale networks without annotated data. Self-supervised methods like [61] report a Graph Neural Networks (GNN)-based solution with positional constraints for robust feature matching in harsh environments. D3VO [62], a self-supervised VO, can jointly estimate depth, pose, and uncertainty for high-precision pose estimation. As a self-supervised semantic VO method, InstanceVO [63] performs motion estimation, depth prediction, and instance segmentation using a shared encoder. In addition, some studies have explored weak-supervised [64] and semi-supervised [65] VO methods.

Beyond the above, recent hybrid VO methods developed through the joint optimization of traditional and learning-based paradigms show promise. Lu et al. [66] incorporate pose graph and BA optimization into DL network training for unsupervised monocular VO, preventing pose drift via joint optimization. DF-VO [67] enforces physical consistency between CNN-predicted depth and feature-based optical flow using a dual-branch architecture and differentiable BA. GraphAVO [68] fuses pixel motion information with graph-assisted optimization and cascaded dilated convolutions to enhance pose estimation accuracy and robustness. DPVO [69], building on DROID-SLAM [70], replaces dense optical flow tracking with a sparse strategy that tracks random tile subsets. It significantly reduces computational load and demonstrates higher accuracy for monocular VO without dense optical flow tracking. DPV-SLAM [71] then extends DPVO to form a complete, real-time, low-memory monocular VSLAM system.

2.3. Event-based VO

Neuromorphic event cameras capture pixel-level brightness changes instead of fixed-frame-rate intensity images, offering advantages in low-light and high-speed motion scenarios where traditional cameras struggle [34]. This has led to the emergence of event-based VO solutions.

For example, EventPointNet [72] converts event data into time surfaces, extracts Harris corner features, and trains a network for keypoint detection, achieving event-based VO through feature matching and pose estimation. Hadviger et al. [73] and Zhou et al. [74] both proposed event-based stereo VO methods. The former relies on time surfaces for feature detection and pose estimation by minimizing reprojection error. The latter uses time surfaces to create spatiotemporal data, estimates inverse depth with nonlinear optimization, fuses depth into a semi-dense map, and tracks the camera in real time. Similar methods are found in [75–77]. In 2022, Hidalgo-Carrio et al. [77] proposed an event-aided direct sparse odometry method that tracks camera motion by fusing event and grayscale frames, enabling accurate 6-DoF estimation.

In addition, several studies have integrated traditional visual data with event data to develop hybrid VO with heterogeneous camera data. For instance, RAMP-VO [78] fuses event and image data using a pixel-level asynchronous feature extractor, integrates features across scales with a multi-scale fusion module, and optimizes state estimation with differentiable BA constraints. Zhu et al. [79] enhance event-based VO using adaptive time surface to select distinctive pixels and design a nonlinear pose optimization method combining RGB-D and event data to improve pose estimation accuracy and robustness. In addition, several works integrate Inertial Measurement Units (IMUs) to develop event-based Visual-Inertial Odometry (VIO) solutions [80–82].

2.4. Periodic discussion

The essence of VO lies in using changes in the camera's perspective to design effective rules to infer its pose changes. Traditional VO methods, based on mathematical paradigms, have a solid theoretical foundation and have achieved practical success in engineering, building on over 30 years of research. Currently, most mathematical paradigm-based VO solutions ensure real-time efficiency on conventional commercial-grade edge devices. Moreover, new solutions such as 360 VO [83] continue to emerge. However, they face limitations such as poor adaptability and robustness in low-texture settings, sensitivity to lighting changes and motion blur in dynamic scenes. These factors form the driving force behind the development of AI-enabled VO methods in recent years.

AI-enabled VO methods excel in representation learning for adaptive feature representation and multi-task optimization. When computing power is not a constraint, they can obviously overcome limitations of handcrafted features, outperforming traditional VO in unstructured and low-texture scenes. Recently, AI-enabled VO research has trended toward label-free approaches. Some hybrid VO methods that integrate the strengths of learning-based and traditional optimization methods are promising. However, the limitations of AI-enabled VO solutions ought to be highlighted as well. They still face challenges and may fail in dynamic scenes and out-of-distribution conditions due to poor generalization, weak real-time performance, and invalid representations from motion blur and lighting changes.

Event-based VO has gained momentum recently, due to its High Dynamic Range (HDR) and event-driven characteristics, which can counteract dynamic blur. Current event-based VO solutions are rapidly developing with diverse ideas coexisting. However, the asynchronous spiking nature of event data requires additional processing steps, with the time surface method being widely adopted. Moreover, although event cameras are highly sensitive to dynamic changes, they suffer significant texture loss. Moreover, there have been no new advances in event cameras capturing depth like depth cameras. Thus, pure event camera-based VO cannot fully replace traditional VO. Therefore, some researchers have also reported strategies that combine the strengths of event cameras with traditional cameras, like [78]. However, this area remains underdeveloped. For example, integrating event camera's asynchronous spiking outputs with traditional visual frames still faces challenges, as hard synchronization issues may need to be considered.

The above discussions mainly focus on VO progress with individual HI paradigms. Building on this, we identify several open challenges. Apart from AI's generalization, this paper attempts to address other issues from the perspective of HI integration, hoping to inspire the related community.

1. How to ensure generalization of AI-enabled VO algorithms? The generalization ability of AI-enabled VO is constrained by its data-driven logic. Since data scarcity and distribution bias are objective reality, training data for VO is often limited and unlabeled, making it hard for DL models to learn universal representations from incomprehensive samples.

   Yet, in LCD, lifelong learning-based solutions like AirLoop [10] show promise in cross-domain generalization, whilst the AI-enabled VO solutions like DPVO [69] focus on zero-shot generalization. Maybe, we can consider to combining lifelong learning, few-shot/zero-shot learning, and even meta learning to keep the performance and robustness of AI-enabled VO methods under cross-domain or out-of-distribution conditions.
2. How can we balance computational efficiency and performance? Taking [84], it integrates three parallel threads into ORB-SLAM3 for dynamic disturbance elimination and background completion, equipping traditional VO with AI capabilities to improve accuracy and adaptability. However, this also causes latency and increases computational demands. Similar issues are common in many works reviewed in [31].

   Regarding this, we believe that focusing solely on a single paradigm is somewhat limited. Integrating multiple HI paradigms with software-hardware considerations might bring new insights. For example, integrating approaches like Spiking-Yolo [85] into traditional VO and deploying it on neuromorphic processors (see Section VI) could reduce power consumption and latency while maintaining accuracy in dynamic object segmentation and filtering.
3. How can we seamlessly integrate the advantages of multiple HI paradigms into VO research? Some studies like [66–68] combine learning-based feature representation with traditional optimization-based correction, forming hybrid strategies. Some others like [63] and [84] integrate AI paradigms into traditional VO systems for dynamic noise filtering. These examples reflect the coupling of multiple HI paradigms at the algorithmic/software level.

Studies like [78] present hardware-level fusion of different HI paradigms.

However, in current research, complementary integration of multiple HI paradigms at the software-hardware co-design level is rare. Can we achieve cross-paradigm fusion of different HI paradigms through this perspective to develop novel hybrid VO solutions? For example, how can we use SNNs to learn features from event cameras, and DL to learn features from standard frames? This approach allows us to replace the current observation-level fusion logic with representation-level fusion, and further integrate differentiable optimization methods into the hybrid network more seamlessly, forming a novel hybrid VO solution. However, this remains an open question requiring further exploration, without a conclusion on this matter.

Note. To facilitate quick access to the essential information of representative VO methods, this paper constructs Table 2. However, its compilation is challenging for the substantial variability in test benchmarks, hardware configurations, and evaluation metrics across studies, as well as the frequent omission of reports on comparative algorithm performance, computational efficiency, and hardware specifications. To address this issue, we have taken publication quality, citation metrics, timeliness, and reproducibility into consideration to provide a concise summary in Table 2, with metrics like accuracy, robustness, and efficiency.

3. LCD progress

In VSLAM, the LCD task is to identify previously visited locations to correct accumulated drift in VSLAM, typically by calculating scene similarity using VPR techniques [35]. This section reviews LCD progress, with corresponding discussion and analysis.

3.1. Traditional LCD

Common keyframe detection methods include the Bag-of-Words (BoW) model, geometric consistency verification, and spatial neighborhood constraint. The BoW model quantizes visual features (e.g., SIFT, ORB) into word frequency vectors for similarity matching [88]. Geometric consistency verification filters mismatches by analyzing feature points' spatial distribution. For example, ORB-SLAM and FAB-MAP 2.0 [89] use Random Sample Consensus (RANSAC) to enhance keyframe detection accuracy despite its computational intensity. Spatial neighborhood constraint leverages camera motion continuity and locality to filter keyframes, constructing a topological graph for efficient retrieval with spatial indexing structures like octrees for fast search [90]. Moreover, combining spatial constraints with appearance-based retrieval can enhance LCD robustness in large-scale environments, especially in repetitive structures, reducing mismatches [91].

After keyframe retrieval, rules are needed to measure scene similarity. Many methods assess the scene similarity using metrics like match count, spatial uniformity, and geometric consistency verification [35, 36]. When scenes are vectorized, their similarities can be quantified using Euclidean distance, Hamming distance, and cosine similarity, etc. Notably, most RatSLAM-derived brain-inspired SLAM methods use visual template matching to achieve LCD [13].

3.2. AI-enabled LCD

Essentially, LCD involves designing rules to evaluate the similarity between multiple scene description features to judge loop closure occurrence. AI-enabled methods can extract and match features automatically through representation learning, becoming valuable in LCD to mitigate the limitations of handcrafted features. Visual and semantic feature descriptions are commonly designed to describe scenes.

Some studies focus on using DL-designed feature descriptors to

Table 2
Summary of VO methods.

	Category	Methods	Year	Contributions
Traditional	Feature-Based	ORB-SLAM [6]	2015	Using ORB features for high-precision pose estimation and mapping. Accuracy: > LSD-SLAM [46], ≈PTAM [87]; Robustness: low failure rate; Efficiency: 25–30 Frame Per Second (FPS) @ low-cost, business-grade CPU.
		StructSLAM [42]	2015	Using architectural line features to reduce drift error in visual SLAM. Accuracy: > > MonoSLAM [43]; Robustness: stable under low-texture conditions; Efficiency: ∼40 FPS @ common commercial-grade CPU.
		Xu et al. [39]	2023	Using point-line flow feature for monocular Visual-Inertial SLAM. Accuracy: ≈ORB-SLAM3 [21]; Robustness: stable under low-texture conditions; Efficiency: ∼17 FPS @ common commercial-grade CPU.
	Direct	DTAM [45]	2011	A pioneering dense direct method that tracks and maps by aligning the entire image. Accuracy: ≈ PTAM [87]; Robustness: > PTAM under motion blur; Efficiency: depends on GPU.
		LSD-SLAM [46]	2015	A representative semi-dense direct SLAM method. Accuracy: ∼2.5 % Root Mean Square Error (RMSE) (KITTI); Robustness: stable to lighting changes; Efficiency: ∼145 FPS (pixel 154 ×46) @ relatively basic commercial GPU.
		DSO [47]	2018	A sparse direct VO method estimating motion by minimizing sparse photometric errors. Accuracy: ∼ORB-SLAM [6]; Robustness: >ORB-SLAM; Efficiency: ∼55 FPS @ common commercial-grade CPU.
AI-Enabled	Supervised	DeepVO [54]	2017	Using a supervised CNN-LSTM network to predict VO trajectories. Accuracy: < ORB-SLAM [6] (KITTI); Robustness: stable under motion blur, lighting changes, low-texture; Efficiency: not report.
		DytanVO [55]	2023	Using curriculum learning to optimize dynamic object segmentation and pose estimation. Accuracy: > > DeepVO [54]; Robustness: stable under dynamic scenes; Efficiency: ∼6 FPS @ 2 high-end commercial-grade GPUs.
		STDN-VO [57]	2025	Mimicking the human visual system's dual-stream mechanism, extracting spatial and temporal features with different networks and fusing them to predict poses. Accuracy: > > DeepVO [54]

(*continued on next page*)

Table 2 (*continued*)

	Category	Methods	Year	Contributions
				and ORB-SLAM [6] (KITTI); Robustness: good generalization; Efficiency: ~26 FPS @ high-end commercial-grade GPU.
	Unsupervised	GANVO [56]	2019	Unsupervised monocular VO where discriminator enhances pose accuracy through optical flow consistency while generator produces depth maps. Accuracy: > ORB-SLAM [6]; Robustness: > ORB-SLAM in both complex and dynamic scenes; Efficiency: ~30 FPS @ computing power-rich commercially-grade GPU.
		Zhou et al. [58]	2017	Pioneering unsupervised learning of depth and ego-motion from video. Accuracy: ≈ORB-SLAM [6] (KITTI); Robustness: ≈ORB-SLAM; Efficiency: not reported.
		GeoNet [59]	2018	An unsupervised learning framework for jointly estimating dense depth, optical flow and camera pose. Accuracy: > [58] and ORB-SLAM [6]; Robustness: stable in occluded and texture-ambiguous regions; Efficiency: < 16 FPS @ relatively basic commercial GPU.
		Kannapiran et al. [61]	2023	A self-supervised stereo VO using a GNN and positional constraints for robust feature matching. Accuracy: ~20 cm RMSE (synthetic dataset); Robustness: stable with scene and lighting changes; Efficiency: ~7 FPS @ common commercial-grade GPU.
		D3VO [62]	2020	Using self-supervised learning to jointly estimate depth, pose, and uncertainty for a monocular VO. Accuracy: > > [58] and ORB-SLAM [6]; Robustness: stable with dynamic blur and lighting changes; Efficiency: not reported.
	Hybrid	DF-VO [67]	2020	Enforcing physical consistency between CNN-predicted depth and feature-based optical flow using differentiable BA. Accuracy: > [58] and ORB-SLAM2 [20]; Robustness: stable under scale-drift mitigation, scale ambiguity resolution; Efficiency: not report.
		Liu et al. [65]	2024	An adaptive learning framework for hybrid VO. Accuracy: > DF-VO [67]; Robustness: different disparity distributions; Efficiency: ~9 FPS @ dedicated workstation-grade GPU.
		GraphAVO [68]	2024	Enhancing pose estimation by fusing pixel motion with graph-assisted optimization. Accuracy: > [58] and ORB-SLAM2 [20]; Robustness: stable under motion blur; Efficiency: ~194 FPS @ relatively basic commercial GPU.
Event-Based	Event-Only	ESVO [74]	2021	An event-based stereo VO for 3D reconstruction via spatiotemporal consistency optimization and probabilistic depth fusion. Accuracy: > ORB-SLAM2 [20]; Robustness: stable in low light and HDR conditions; Efficiency: ~20 FPS (DAVIS 346) @ common commercial-grade CPU.
		EVIO [75]	2022	A monocular event-based VO using event-corner with sliding windows graph-based optimization. Accuracy: > ORB-SLAM3 [21]; Robustness: > ORB-SLAM3, stable in low-light and HDR conditions; Efficiency: ~40 FPS (DAVIS346) @ computing power-rich commercially-grade CPU.
		Wang et al. [76]	2023	Achieving event-based stereo VO with native temporal resolution via continuous-time Gaussian process regression. Accuracy: > ESVO [74]; Robustness: stable in complex motions and HDR conditions; Efficiency: not report.
	Hybrid	Hidalgo-Carrio et al. [77]	2022	Tracking a DAVIS240C event camera’s motion by combining events and grayscale frames, estimating motion by minimizing brightness increment error. Accuracy: >ESVO [74] and ORB-SLAM [6]; Robustness: stable under low frame rates with depth noise and contrast noise; Efficiency: not report.
		RAMP-VO [78]	2024	Fusing event and image data using pixel-level asynchronous feature extraction and multi-scale fusion with differentiable BA. Accuracy:> DPVO [69], ORB-SLAM2 [20], ORB-SLAM3 [21]. Robustness: stable in low-light and HDR scenarios; Efficiency: the training relies on a dedicated workstation-grade GPU without time-cost reports.
		Zhu et al. [79]	2023	Using adaptive time surface to select distinctive pixels and combines RGB-D with event data for improved pose estimation. Accuracy: >ESVO [74]; Robustness: reliable under complex dynamic motion conditions.; Efficiency: ~12 FPS (RGB-D & DVXplorer Lite) / ~80 FPS (RGB-D only) @ common commercial-grade CPU.
		ESVO2 [86]	2024	A direct event-based VO approach using a stereo event camera. Accuracy: >ESVO

(*continued on next page*)

Table 2 (*continued*)

Category	Methods	Year	Contributions
			[74]; Robustness: stable in low light, and HDR scenes; Efficiency: mapping efficiency increased by 5x compared to ESVO @ computing power-rich commercially-grade CPU.

enhance LCD [92]. Many others have developed specialized DL strategies for representation learning of local or global visual features to create effective scene descriptions for similarity estimation. For instance, NetVLAD [93], an upgrade of VLAD [94], uses a CNN to extract global image features and map images into compact vectors, improving scene recognition accuracy over traditional BoW models. Furthermore, without considering computational cost, some researchers use Visual-and-Language Model (VLM) to create scene description, such as [95]. It provides human-readable failure traceability and has interpretability and real-world application potential.

Recent DL-assisted visual LCD has focused on improving effectiveness, robustness, and real-time performance. Ma et al. [96] proposed a fast LCD method, combining an image-to-sequence candidate selection strategy and a feature matching algorithm with motion consistency constraints. Memon et al. [97] used VGG16 for feature extraction and moving object recognition, introducing a super dictionary combined with an AE for quick scene revisit determination. GOReloc [98] employs semantic topology graph matching and graph-kernel vector similarities to efficiently extract candidate subgraphs, surpassing ORB-SLAM2 in real-time performance. LoopNet [99] is an LCD method for dynamic settings, fusing feature maps and highlighting key landmarks through a multi-scale attention-based Siamese convolutional network. Zhou et al. [100] proposed a lightweight Siamese capsule network for LCD, employing depthwise separable and dilated convolutions with pruning layers to enhance real-time performance. AirLoop [10] is a lightweight lifelong LCD method, combining memory-aware synapses and relational knowledge distillation to adapt to new environments. VIPeR [101] improves AirLoop through adaptive mining, multi-stage memory, and probabilistic distillation, reducing catastrophic forgetting and boosting benchmark performance, thereby enhancing VPR in terms of adaptability and robustness. I2KEN [102] is also a lifelong LCD method, solving cross-domain adaptability and catastrophic forgetting via single- and cross-domain knowledge augmentation, and lifelong adaptive fusion.

Additionally, semantic descriptions are also effective for the LCD task and can be combined with visual feature methods. Semantic descriptions are primarily learning-based methods. For example, Singh et al. [103] designed a hierarchical semantic-geometric descriptor to fuse global scene categories and their geometric distribution, using semantic labels to filter out dynamic interference, enhancing LCD performance. Similarly, PlaceNet [104] extends LoopNet by learning to ignore dynamic objects to create landmark-focused semantic descriptions, robust to dynamic scenes and scale variations. AEGIS-Net [105] and CGiS-Net [106] construct global descriptors by fusing low-level color and geometric cues with high-level semantic features, showcasing superior robustness compared to appearance-only methods like NetVLAD. TextSLAM [107] models textual objects as texture-rich planar patches, using their semantic information as landmarks to match text semantics for keyframe detection, achieving robust LCD.

Semantic features, beyond forming scene descriptions, can pair with visual features to reduce matching uncertainty in LCD. SLC^2-SLAM [108] enhances LCD in NeRF SLAM for better reconstruction quality using latent codes as local features and aggregating them with semantic information. Chen et al. [109] addressed instance-level inconsistencies to enhance LCD for dynamic scenes by integrating visual-semantic geometric verification. SemanticLoop [110] creates a 3D semantic graph via instance-level embedding and uncertainty detection, achieving robust LCD by geometric matching. SymbioLCD2 [111], building on SymbioLCD [112], combines semantic and visual features in a graph structure, performing LCD with the Weisfeiler-Lehman kernel under temporal constraints.

3.3. Brain-inspired VPR

While LCD is strictly a subset of VPR applications, both aim to determine whether a particular scene has been visited. Therefore, this paper reviews brain-inspired VPR technologies to investigate progress beneficial to VSLAM’s implementation of LCD from the brain-inspired computing paradigm.

Fischer et al. [113] proposed an energy-efficient event camera-based VPR method that extracts sparse features and uses feature count differences for rapid localization. Ev-ReconNet [114], LoCS-Net [115] and VPRTempo [116] are all SNN-based VPR models. Ev-ReconNet processes event streams directly to improve accuracy in extreme lighting. LoCS-Net uses ANN2SNN conversion for fast VPR, enhancing real-time performance. VPRTempo uses temporal encoding linked to pixel intensity, trained with Spike-Timing-Dependent Plasticity (STDP) and a supervised delta learning rule, ensuring each output spike neuron responds to a unique location. Hussaini et al. developed a series of SNN-based VPR methods, from regularized neuron allocation [16] to a modular region-specific ensemble system [17], and finally a modular architecture with geographical tiling and ensemble learning to enhance accuracy and generalization [18].

Some unique approaches also warrant attention. For example, Zhu et al. [117] developed a spatiotemporal memory algorithm inspired by insect mushroom body neural circuits, using neuromorphic computing to encode spikes and store visual sequence memories for real-time visual familiarity assessment in complex environments. Neubert et al. [15] employed a Mini-Column Network (MCN) model inspired by the brain neocortex for VPR tasks, simulating sequence memory and cell predictive connections. They also reported combining MCN with a grid cell-inspired model to enhance VPR [118]. Ozdemir et al. [119] focused on Echo State Networks (ESNs) for capturing temporal relationships in data, combining ESNs with preprocessed CNNs for VPR tasks, surpassing some sequence matching models.

3.4. Periodic discussion

The essence of LCD is to design machine-computable rules for describing environments and assessing scene similarity. LCD and VPR technologies are largely similar, but LCD in VSLAM must consider computational timeliness.

Traditional LCD methods identify keyframes and perform feature matching for similarity comparison with historical scenes. While effective in structured environments, these methods struggle with insufficient robustness due to lighting changes, appearance variations, and viewing angle differences. They also face challenges related to heavy storage requirements for visual templates.

AI-enabled LCD methods automatically learn environmental descriptions through representation learning, assessing scene similarity with higher accuracy and robustness within computational timeliness constraints. The roles of AI in LCD and VO are somewhat similar, both involving the extraction and description of scene features, followed by application-specific utilization. Thus, to a large extent, the previous analysis of the strengths and weaknesses of AI-enabled VO methods in Section II is largely applicable to AI-enabled LCD paradigms. However, while VO tasks focus on the offset representation of scene features, LCD focuses on their similarity. In VO, sparse semantic features are rarely used for pose estimation and mostly serve as an auxiliary in geometric constraints and dynamic noise filtering. In LCD, however, semantic features are often used to enhance scene descriptions or mitigate the negative impact of distracting backgrounds and dynamic objects.

Brain-inspired VPR methods have introduced new sensor types (e.g., event cameras) and shifted computing paradigms. For instance, appearance/feature-based solutions using SNN and ESN have brought new changes to VSLAM's LCD [116,119]. Notably, some studies have explored novel mechanisms inspired by animal and insect brains neural mechanisms, offering new insights for VPR [117,118]. These emerging methods show benefits in computational efficiency, interpretability, and adaptability to neuromorphic deployment, yet further exploration and improvement are still needed in terms of models' training effectiveness and accuracy.

The preceding contents examine LCD advancements through the lens of individual HI paradigms. From this foundation, we pinpoint several open challenges. Besides AI generalization, this paper then delves into these challenges, through the lens of HI integration. It is worth noting that, as analyzed above, AI applications in VO and LCD tasks share many commonalities in their underlying logic. Thus, while the perspective on the following open challenges is similar to that in Section II, there are differences, and some insights may be mutually beneficial.

1. How to ensure generalization of AI-enabled LCD algorithms? As discussed in Section II, expanding the scale of high-quality training data clearly benefits AI-enabled solutions. However, this is challenging, especially for numerous public benchmarks that are already established and unchangeable.

   Therefore, the previous discussions in Section II about using lifelong learning, few-shot/zero-shot learning, and meta-learning to enhance the generalization and usability of AI-enabled LCD methods are equally applicable here.
2. How can we balance computational efficiency and performance? The insights here differ from those in the VO section. Since training an SNN with complex network structure is difficult, there is almost no research on using SNNs for continuous dynamic pose estimation in VO from complex traditional image data. However, for static scene description tasks, brain-inspired VPR methods, particularly SNN-based solutions, can be effective. They may have huge advantages in efficiency and power consumption on neuromorphic hardware.

Moreover, SNN-based semantic recognition solutions have demonstrated reliable performance [85,120]. However, they still fall short of DL in descriptor representation for complex visual environments. Therefore, we suggest introducing brain-inspired paradigms to facilitate neuromorphic acceleration into pure appearance-based visual feature description methods (traditional or AI-enabled) to build visual-semantic feature descriptors, which is a worthwhile approach.

3. How can we seamlessly integrate the advantages of multiple HI paradigms into LCD research? Like VO, most existing LCD progress has only preliminarily integrated multiple HI paradigms at either the algorithmic level or the hardware level. For example, combining cross-modal data from traditional and event cameras shows benefits in overcoming single-modal limitations, enabling more diverse and reliable environmental description rules [117]. Given the natural compatibility of SNNs with event camera data and the common use of CNNs for traditional images, relevant advances have been made in hybrid VPR like [119]. Moreover, the brain's neural mechanisms are valuable to inspire novel VPR ideas [118], potentially shifting VSLAM's LCD from feature-based to episodic memory-based approaches.

Nevertheless, in the existing VPR works, it is rare to find a study like NeuroGPR [121] that integrates multiple HI paradigms through software-hardware co-design. Thus, future research may explore NeuroGPR as a foundation for capability enhancement or practical application, such as integrating it as an LCD module within VSLAM systems. By the way, no similar breakthrough has been seen in the VO field. It may have reference value for the VO progress.

Note. Table 3 offers rapid access to the essential information of representative LCD progress. Given the huge variances in test benchmarks, evaluation metrics, and hardware configurations across different

Table 3
Summary of LCD/VPR methods (A, B, and C correspond to the properties of summarized methods, representing Traditional, AI-enabled, and Brain-inspired method types, respectively.).

	A	B	C	Methods	Year	Contributions
Appearance -Only	√			Bow [88]	2003	Quantizing visual features (e.g., SIFT, ORB) into word vectors. Precision & Robustness: depends on scenes; Efficiency: depends on CPU.
	√			FAB-MAP 2.0 [89]	2009	Using RANSAC to enhance robustness and improve LCD accuracy. Precision & Robustness: depends on scenes; Efficiency: depends on CPU.
	√			RatSLAM [13]	2013	Using a local view cell module to store and match visual templates for LCD. Precision & Robustness: depends on parameter configuration; Efficiency: depends on parameter configuration.
	√			NeuroSLAM [14]	2019	Performing LCD as well as RatSLAM. Accuracy, robustness, and efficiency are all on par with RatSLAM [13].
		√		NetVLAD [93]	2016	Using a CNN to extract global features and map images into compact vectors. Precision: ~74 % (average) Recall@1 (multiple datasets from [99]); Robustness: stable to illumination and viewpoint changes, and occlusion; Efficiency: depends on GPU.
		√		Ma et al. [96]	2022	Using a convolutional AE and motion consensus with a super dictionary. Precision: > 80 % @ maximum recall (KITTI); Robustness: stable in complex environments; Efficiency: ~105ms per inference @ an entry-level-priced GPU.
		√		LoopNet [99]	2022	A multi-scale attention-based Siamese convolutional network for LCD. Precision: > NetVLAD [93]; Robustness: stable to scene, viewpoint, and illumination variations; Efficiency: 2x faster than NetVLAD.
		√		AirLoop [10]	2022	A lightweight lifelong learning LCD method. Precision: ~92 %

(continued on next page)

Table 3 (*continued*)

	A	B	C	Methods	Year	Contributions
						Recall@1 (Nordland); Robustness: stable to appearance changes; Efficiency: 97–290ms per inference (hardware configuration undisclosed).
			√	Fischer et al. [113]	2022	Event camera-based VPR that extracts features with significant changes and uses feature count differences for rapid VPR. Precision: ~64 % Recall@1 (self-collected dataset); Robustness: robust to moderate speed variations; Efficiency: ~1ms per inference (DAVIS346) @ a high-end commercial CPU.
			√	VPRTempo [116]	2024	Employing temporal encoding linked to pixel intensity, trained via STDP and a supervised delta learning rule. Precision: ~56 % Recall@1 (Nordland), > NetVLAD [93]; Robustness: stable to seasonal/lighting changes; Efficiency: > 50 Hz @ a common commercial-grade CPU.
			√	Hussaini et al. [16]	2022	A regularized weighted neuron allocation scheme for SNN-based VPR. Precision: 47.5 % Recall@1 (Nordland), > NetVLAD [93]; Robustness: stable under lighting/seasonal changes; Efficiency: ~0.2 s per inference (~81 watts) @ a GPU (configuration undisclosed).
			√	Hussaini et al. [17]	2023	A modular, region-specific SNN ensemble system for VPR. Precision: ~52.6 % Recall@1 (Nordland), > NetVLAD [93] & [16]; Robustness: stable to seasonal/lighting changes; Efficiency: not report.
			√	Hussaini et al. [18]	2025	A SNN-based VPR architecture integrating a geographical tiling mechanism and ensemble learning. Precision: > [16]; Robustness: stable to seasonal/lighting changes; Efficiency: 1–2 s per inference @ a high-end commercial CPU.
			√	Neubert et al. [15]	2019	Employing MCN inspired by the human neocortex for VPR. Precision: > 70 % @ average precision (Nordland); Robustness: stable to lighting/seasonal changes; Efficiency: ~2.1 s per inference @ a common commercial-grade CPU.
			√	Ozdemir et al. [119]	2022	Combining ESNs with preprocessed CNNs for VPR. Accuracy: > > NetVLAD [93]; Robustness: depends on the parameter configuration; Efficiency: not report.
		√	√	NeuroGPR [121]	2023	Integrating both neuromorphic and traditional cameras and combining AI-enabled and brain-inspired hybrid computing paradigms. Precision: relevant to different environments.; Robustness: robust to environmental uncertainties like appearance ambiguity and lighting changes; Efficiency: 10.5x lower latency & 43.6 % lower power consumption than Jetson Xavier NX @ Tianjic.
Semantic-Assisted		√		TextSLAM [107]	2024	Modeling textual objects as texture-rich landmarks, using text semantic matching to detect keyframes and search for point-level correspondences. Precision: >ORB-SLAM and NetVLAD [93]; Robustness: > NetVLAD, stable to motion blur, illumination changes; Efficiency: > NetVLAD @ entry-level-priced CPU.
		√		SemanticLoop [110]	2023	Constructing a 3D semantic graph via instance-level embedding and uncertainty detection with geometric graph matching. Precision: > 90 %@recall (TUM); Robustness: stable against appearance changes and complex scenes; Efficiency: < 0.4 ms per matching @ entry-level-priced CPU.

(continued on next page)

Table 3 (*continued*)

A	B	C	Methods	Year	Contributions
	√		SymbioLCD2 [111]	2022	Constructing a graph structure to fusion semantic and visual features with temporal constraints. Precision: > 93 %@recall (TUM), > ORB-SLAM2 [20]; Robustness: robust to dynamic disturbances; Efficiency: not report.
	√		PlaceNet [104]	2023	Expending LoopNet, generates robust feature representations through multi-scale feature learning and semantic fusion. Precision: > 95 % @recall (multiple benchmarks); Robustness: stable against dynamic scenes, illumination changes, and viewpoint variations; Efficiency: > NetVLAD [93], ∼5ms per matching @ relatively basic commercial GPU.

works, as well as the frequent omission of key comparable metrics, we have adopted the same principles used for compiling Table 2 to prepare Table 3.

4. Backend optimization progress

In VSLAM, backend optimization integrates the frontend's local perception cues to prevent mapping failures from accumulated errors. Traditional methods model it as state estimation or nonlinear optimization. In contrast, brain-inspired SLAM methods simulate navigational cells and path integration via brain-inspired models [122], building a spatial experience map by integrating local cues continuously.

4.1. *Traditional backend optimization*

Traditional backend optimization methods are divided into filtering and optimization methods. Filtering methods, based on Bayesian theory, process data in real time through iterative prediction and updates. They are suitable for dynamic environments but face high computational complexity and limitations of the Markov assumption [38]. EKF-SLAM [123] is an early filter-based solution. In contrast, optimization methods reformulate SLAM as a Nonlinear Least Squares (NLS) problem, using all historical data to achieve high accuracy.

Despite their computational demands, optimization methods have become mainstream, supported by advances in hardware and optimization theory. Among them, BA is the most classic technique. It typically uses Gauss-Newton or Levenberg-Marquardt methods to solve the NLS problem. PTAM [87] separates tracking and mapping into parallel threads, optimizing recent keyframes via local BA. The ORB-SLAM series, based on PTAM, is a prime example. LSD-SLAM [46] maintains consistency in large-scale settings by combining semi-dense direct methods with BA. DS-SLAM [11] and RTAB-Map [124] integrate semantic information and memory management strategies in dynamic and large-scale settings, respectively, reducing VO drift and updating the map via BA.

Beyond BA, graph optimization models SLAM problems as a graph structure, with nodes as poses or landmarks and edges as constraints. Currently, general-purpose graph optimization frameworks like g2o and GTSAM have significantly advanced SLAM standardization and application. In addition, the recently reported PyPose [125] has also been proven to support the backend optimization of SLAM with high efficiency. These frameworks provide flexible graph structure definitions and efficient interfaces, lowering SLAM development barriers.

Moreover, pose graph optimization, often applied in global optimization, employs camera poses as nodes and relative measurements as edges, reducing computational load compared to global BA. For example, LDSO [126] uses a pose graph to correct errors post-loop closure by optimizing only camera poses. RGB-D SLAM [127] integrates RGB features and depth into a pose graph to minimize optimization variables. In addition, factor graph optimization models SLAM as a bipartite graph, decomposing the joint probability distribution into factors. Its modularity aids integration of multi-sensor data and prior knowledge. Representative cases like iSAM [128], iSAM2 [129], and their improved versions [130–132] support incremental optimization, efficiently processing new observations without re-optimizing the entire graph.

4.2. *Brain-inspired backend path integration*

Unlike traditional ideas, the goal of brain-inspired SLAM is to replicate the brain's ability to encode spatial experience, integrate local environmental cues from the SLAM frontend, and use stored spatial memories (visual templates) to suppress cumulative error drift during long-term mapping [13,133,134]. Brain-inspired SLAM methods receive self-motion cues from sensors (e.g., VO, Sonar, Lidar), use CANNs to simulate the brain's spatial cue encoding and path integration, and obtain spatial representations by decoding neural activity patterns. LCD is performed via visual template matching, mimicking the brain's mechanism of correcting path integration errors with similar spatial memories [122]. These commonalities transform traditional backend optimization into a problem of optimal spatial experience encoding and decoding.

Taking RatSLAM as an example, it used a CANN-based pose cell network, inspired by hippocampal place cells, to encode path integration by extracting self-motion cues from VO, with visual template matching for LCD. The spatial experience was subsequently decoded to construct an experience map [135]. Afterwards, as neuroscience advanced, grid cell mechanisms in the entorhinal cortex were elucidated and integrated into algorithms like NeuroBayesSLAM [133]. Zeng et al. [136], inspired by the entorhinal cortex's joint encoding mechanism, proposed a combined encoding CANN model of grid cells and head-direction cells to replace RatSLAM's pose cell network. Furthermore, they simplified CANN's neurodynamics using a Bayesian probabilistic framework, creating the more efficient NeuroBayesSLAM.

Currently, research is increasingly focusing on 3D navigation cells. DolphinSLAM [137] integrates RatSLAM and FABMap, using a CANN-based 3D place cell network to build experience maps for underwater scenes. Yu proposed NeuroSLAM [14]. It constructs a joint pose cell model using 3D grid cells and multilayer head-direction cells, replacing RatSLAM's pose cell network to achieve 3D path integration and build multilayer experience maps. Thereafter, Shen et al. [138] proposed ORB-NeuroSLAM, incorporating ORB features to improve NeuroSLAM's LCD and enhance experience map accuracy.

4.3. *Periodic discussion*

In backend optimization, both traditional and brain-inspired pathways, despite differing principles and computing paradigms, share the same goal of effectively integrating the frontend's local cues to reduce accumulated errors and avoid mapping failure.

Traditional methods, rooted in early probabilistic computation, have evolved from filtering methods to optimization approaches. These methods, now mature after decades of development, still face challenges in computational efficiency on edge devices. Co-design of software and hardware to enhance computational efficiency is a promising solution [139,140].

Brain-inspired SLAM converts backend optimization into optimal spatial experience encoding and decoding, supported by spatial memory. It simulates the brain's path integration using brain-inspired models and memory matching for LCD to correct experience maps. However, it currently has lower mapping accuracy and limited ability to describe complex environments compared to traditional methods, restricting its practical applications. Research on navigation neural circuits is still in its early stages, with limited understanding, making it challenging to fully replicate the powerful path integration capabilities of animal brains.

Therefore, research on navigation neural mechanisms and the collaborative mechanisms of heterogeneous navigation cells is essential for advancing brain-inspired SLAM studies. To address the low computational efficiency of CANN models, some researchers have improved efficiency using a Bayesian framework [133], while others have accelerated CANNs by converting them to SNNs, utilizing neuromorphic computing solutions [141,142].

Moreover, given that brain-inspired SLAM reconstructs the underlying logic of traditional backend optimization, the two seem to be in competition in terms of global optimization in the backend. However, notably, the ORB-NeuroSLAM [138] system attempts to introduce local BA optimization in the frontend VO based on ORB features, which does not conflict with the brain-inspired path integration in the backend.

5. Mapping progress

In VSLAM, backend mapping's function design depends on frontend processing and application requirements, not a fixed algorithmic framework. Thus, VSLAM/s mapping methods range from sparse to dense and from geometric to semantic, constrained by frontend feature extraction. This paper categorizes backend mapping into geometric, semantic, and generalized mapping based on different scene representation.

5.1. Geometric mapping

Geometric mapping focuses on scene shape and structure, including depth information, mesh representation, and topological representation. Depth information, obtained from stereo vision, depth cameras, or DL methods, can reflect the scene's geometric structure. For example, CNN-SLAM [143] uses a CNN to predict per-pixel depth and integrates it into the VSLAM system for dense reconstruction.

Mesh representation constructs a mesh map by estimating the height or depth of each mesh cell and is widely used in navigation and path planning. For example, Gmapping [144] builds high-precision 2D occupancy grid maps widely adopted by Robot Operating System (ROS) for indoor navigation. Adding height information to a 2D grid creates an elevation map (2.5D map) suitable for uneven terrain navigation.

Voxel maps divide 3D space into regular grids (voxels) to record occupancy status for 3D environmental modeling. SpOctA [145] improves 3D voxel map construction efficiency using octree encoding. Topological representation focuses on the environment's topological structure and is used by most brain-inspired SLAM systems to build experience maps due to CANN's decoding characteristics. RatSLAM creates a 2D topological experience map [135], later extended to 2.5D by Milford et al. [146] and to 3D by NeuroSLAM.

5.2. Semantic mapping

Semantic mapping aims to construct maps with geometric and semantic information, focusing on semantic extraction and mapping [147]. Common semantic extraction methods in VSLAM include object detection and semantic segmentation using techniques like SSD, YOLO series, and other learning-based approaches.

For instance, DS-SLAM [11] removes dynamic objects using SegNet and motion consistency checks, reducing localization errors in dynamic environments and aiding dense semantic octree map construction. DynaSLAM [148] performs initial semantic segmentation with a Mask R-CNN based on ORB-SLAM2 and tracks unsegmented dynamic objects by minimizing photometric reprojection errors. Detect-SLAM [149] and Dynamic-SLAM [150] improve SSD detectors for specific tasks, with similar works including YOLO-SLAM [151] and CubeSLAM [152].

In fact, VSLAM technology can be combined with many advanced object detection and semantic segmentation methods beyond the common solutions. For example, Blitz-SLAM [153] uses BlitzNet (based on ResNet-50) for object detection and semantic segmentation. Reviews of object detection and semantic segmentation research over the past 20 years are in [154–156]. Latest methods based on GNN, SNN, etc., can be found in [157–167] and are expected to positively impact semantic mapping.

Semantic mapping integrates semantic data (e.g., object categories and locations) with scene geometry to enhance map interpretability [168]. For instance, SemanticFusion [169] fuses multi-view CNN semantic predictions using SLAM-derived correspondences and probabilistic methods, producing accurate and real-time 3D semantic maps. TextSLAM [107] embeds geometric parameters and semantic content (text strings) of textual objects into a 3D map synchronously. QuadricSLAM [170] represents objects with quadratic surfaces, integrating geometric constraints and semantic information for a flexible and compact representation. More related progress can be found in [147, 171,172].

5.3. Generalized mapping

Generalized mapping utilizes implicit scene representations via DL to encode scenes compactly for reconstruction or pose estimation. For instance, CodeSLAM [173] employs a deep AE to convert images into an optimization-friendly format, enhancing VSLAM efficiency and accuracy in camera pose tracking and scene reconstruction.

Recently, NeRF has advanced 3D scene representation, enabling implicit mapping in VSLAM. iNeRF [174] first applied NeRF for pose estimation through re-localization using a pre-trained model. iMAP [175] then integrated NeRF into VSLAM for joint optimization of the embedded scene map. NICE-SLAM [176] expanded this by using hierarchical and neural implicit representations to model larger scenes. SLC^2-SLAM [108] used semantic-guided LCD tailored for NeRF SLAM with graph optimization and BA, delivering superior reconstruction quality, especially in large-scale scenes. Vox-Fusion [177] combines the sparse-voxel octree with neural implicit representations, yielding a memory-efficient, dynamically extensible, real-time dense SLAM framework.

Additionally, recent advances like mixed spiking NeRF [178], event camera-based E-NeRF [179] and E^{2}NeRF [180], which integrates an event camera with a standard RGB camera, offer efficient SNN-based NeRF solutions that may boost neuromorphic VSLAM development. However, NeRF-assisted generalized mapping faces challenges such as over-smoothing and catastrophic forgetting, despite strengths in feature mapping, tracking, and novel view synthesis.

Moreover, 3D GS have attracted attention for their efficient rendering, explicit representation, and robust optimization. GS-SLAM [181] combines 3D Gaussians with splat rendering, encapsulating scene geometry and appearance using 3D Gaussians, opacity, and spherical harmonics. This approach significantly enhances rendering speed and map optimization efficiency compared to NeRF-based methods. Photo-SLAM [182], SplaTAM [183], and GS-SLAM all model scenes with 3D GS, representing each point as a Gaussian distribution with direction, elongation, color, and opacity.

Not only that, event-driven 3D GS progress based on event cameras is increasing. For example, EOGS [184] optimizes rendering using brightness changes from an event camera with an event brightness loss function, enabling high-quality 3D GS reconstruction under motion blur and low-light conditions. Ev-GS [185] infers 3D GS from monocular event camera data, excelling in reducing blurring, improving visual quality, and offering computational and memory efficiency. E2GS [186] integrates event camera data with GS, using both blurry images and event data to enhance image deblurring and novel view synthesis quality while achieving faster training and rendering speeds. These methods offer efficient rendering, explicit representation, and rich optimization capabilities, while utilizing submaps to prevent catastrophic forgetting and maintain computational efficiency.

5.4. Periodic discussion

Backend mapping is determined by frontend processing and application needs. The frontend dictates input data quality and type, while the application defines the map's functionality and form. For instance, when the frontend uses feature-based methods, VSLAM systems typically construct sparse maps (e.g., PTAM [87], ORB-SLAM). Direct methods can produce sparse, semi-dense (e.g., DSO [47]), or dense maps (e.g., DTAM [45]). Semi-direct methods, like semi-direct multimap SLAM [187], can combine both approaches for robust real-time reconstruction in dynamic scenes. Brain-inspired SLAMs (e.g., RatSLAM, NeuroSLAM) create geometric topological maps via CANNs for spatial experience encoding and decoding.

This paper classifies backend mapping strategies into three categories. Geometric mapping focuses on depth information, mesh representation, and topological structure. Semantic mapping integrates object detection and semantic segmentation methods to enrich environment representations through semantic extraction and mapping. Generalized mapping transitions to implicit representations and neural rendering, enabling lightweight storage and enhanced expressiveness. State-of-the-art techniques like NeRF and 3D GS redefine mapping paradigms [32].

From an algorithmic perspective, brain-inspired SLAM excels at capturing environmental topology, while AI-enabled SLAM is proficient at extracting semantic information. Humans can simultaneously encode both topological and semantic information into abstract the cognitive map during exploration. Therefore, it is not difficult to envision that integrating AI-based semantic understanding with brain-inspired topological descriptions could emulate human spatial cognition, driving the development of cognition-driven VSLAM for robust, large-scale VSLAM systems. However, how to adapt the current VSLAM framework to incorporate neural implicit map representations, especially event-driven approaches like [180,186], etc., requires further investigation.

6. Hardware support for VSLAM systems

Hardware enables the practical application of intelligence. Traditional visual sensors face challenges like motion blur and light sensitivity, despite improvements in resolution, sensitivity, and dynamic range. Event cameras and bio-inspired visual sensors address these issues by providing novel visual perception capabilities for VSLAM systems. Hardware computational power remains a critical limiting factor for intelligence, especially in AI, as evidenced by the 20-year dormancy of DL due to computational constraints.

Moreover, the evolution of backend optimization pathway in SLAM also highlights the significant impact of hardware computing power on system-level SLAM development. From a system-level perspective, beyond algorithms, it is essential to summarize the heterogeneous visual sensors and computing hardware that benefit VSLAM technology in the era of HI coexistence. This will facilitate more researchers in generating innovative ideas.

6.1. Visual sensors

Traditional visual sensors like monocular and stereo cameras estimate depth using multi-view geometric constraints but are limited by feature matching accuracy and adaptability to dynamic environments. RGB-D cameras, which directly capture depth via structured light or time-of-flight technology, perform well in low-texture scenes. ORB-SLAM2 is a prime example that supports monocular, stereo, and RGB-D cameras. Panoramic cameras, which use multi-lens stitching or fish-eye lenses to expand the field of view, enhance global scene understanding and have been demonstrated in NeuroSLAM [14]. Moreover, multi-camera setups are employed in VSLAM systems like BE-SLAM [188] and BEV-SLAM [189].

Bionic cameras, modeled after insects' optic flow navigation mechanism, offers a wider field of view, stronger moving object detection, and higher light sensitivity. Despite limited research focus, recent progress demonstrates significant advantages in enhancing VSLAM performance in low-texture environments. Specifically, Liu et al. [23, 190] developed a VSLAM system with a bionic eye that actively searches for texture, thereby improving system robustness. For research progress on bio-inspired visual sensors, refer to [191].

Event cameras, inspired by primate retinal structures. The Dynamic Vision Sensor (DVS) series, Asynchronous Time-based Image Sensor (ATIS) series, Dynamic and Active-pixel Vision Sensor (DAVIS) series, mimic the peripheral retinal structure by detecting brightness changes and outputting event streams [34]. Moreover, exemplified by Vidar [192] developed by Huang's team, uses foveal photoreceptors and proposes an integrative visual sampling model. Beyond them, SCAMP-5 [193] is a novel event camera that integrates sensing and computing by processing optical signals on-chip and synchronously parallelizing all pixels within the same clock cycle, unlike DVS's asynchronous output. Benefit from event-based DVS, Kreiser et al. [141] promoted the development of neuromorphic SLAM. Research on event -based VPR benefits LCD [117,194]. In addition, NeuroGPR [121] integrates both RGB-D and DAVIS346 cameras for place recognition.

6.2. Chips and processors

In non-mobile environments, high-performance CPU/GPU workstations support intensive AI processing but are unsuitable for edge applications in unmanned systems due to latency and energy constraints. This has driven advancements in dedicated AI processors for mobile devices [195]. Low-cost edge devices like Raspberry Pi and Orange Pi are used for prototyping but lack power for complex VSLAM tasks [196]. Moderate-capability edge AI modules, like NVIDIA's Jetson TX2 and Rockchip's RK3588, are suitable for moderately complex VSLAM tasks. Dedicated AI-accelerated devices, including edge GPUs (e.g., NVIDIA's Jetson NX and Orin series), Google's Coral TPU, and Cambricon's NPU, enhance energy efficiency. Jetson-SLAM [140] achieves over 60 FPS on Jetson NX and exceeds 200 FPS on desktop GPUs. Hybrid AI computing boxes can integrate multi-core CPUs, GPUs, and NPUs for real-time mapping and localization [197]. Other notable processors include Horizon Robotics' Brain Processing Units (BPUs) [198] and Graphcore's Intelligent Processing Units (IPUs) [199], etc.

Despite advances in AI-specific accelerators, von Neumann architecture-based computing units face diminishing returns from Moore's Law. Amid the era of HI co-development, neuromorphic computing has emerged as a solution. Successful achievements include the Neurogrid [200] and Braindrop [201], the BrainScaleS series [202], the SpiNNaker series [28,203], SynSense's DYNAPs [204], Dynap-SEL [205] and Dynap-SE2 [206], Intel's Loihi series [207,208], International Business Machines (IBM)'s TrueNorth [209], China's Darwin series [210,211], and Tianjic series chips [2,27], ect. Yoon et al. [142] demonstrated a 65-nanometer NeuroSLAM accelerator IC based on neuromorphic computing, achieving ultra-low-power VSLAM functionality via mixed-signal oscillators. Theoretically, ANN2SNN technology can theoretically convert DL solutions into SNNs, facilitating ultra-low-power SLAM through neuromorphic computing. However, this requires a system-level coordinated solution.

Additionally, quantum intelligence has introduced new hardware computing solutions in the era of HI coexistence, such as IBM's Flamingo processor, Google's Willow chip, and PsiQuantum's Omega chip [212–215]. However, quantum computing chips and sensors have not yet been applied in SLAM and thus are not detailed here. They may potentially benefit the system-level development of VSLAM in the future.

6.3. Periodic discussion

In the era of HI coexistence, manifestations of intelligence are diverse, including heterogeneous sensors, computing methods, and processors. They offer opportunities for next-generation VSLAM technologies and systems. As previously noted, hardware support, including sensors and chips, embodies intelligence to meet practical application needs. Compared to traditional visual sensors, bio-inspired visual sensors and event cameras have equipped VSLAM systems with new perception capabilities in the era of HI coexistence. At present, a VSLAM system can even integrate multi-cameras, taking BEV-SLAM [189] as an example.

Today, AI continues to innovate with increasingly mature AI-specific processors. However, constraints from the von Neumann architecture limit further progress in optimizing power consumption and improving computational efficiency. In contrast, brain-inspired intelligence, with substantial global investment, holds broad future prospects [216]. Despite incomplete understanding of the brain's architecture, research findings have inspired advanced neuromorphic chips like Intel's Loihi2, leading to the world's largest neuromorphic computing system, Hala Point [217]. It is designed to support advanced research in brain-inspired intelligence and address efficiency and sustainability challenges in current AI. This paper does not address quantum computing in-depth due to the lack of mature and practical quantum chips. Table 4 shows comparison of current neuromorphic chips.

Table 4
Comparison of large-scale neuromorphic chips.

Chips	Signals	On-chip learning	Process (nm)	Neurons / Synapses	Energy Efficiency (GSOPS/W)
Neurogrid	Mixed	No	180	64k/100 M	1.1
Braindrop	Mixed	Yes	28	4k/16 M	2630
BrainScaleS	Mixed	Yes	180	512/128k	10
BrainScaleS2	Mixed	Yes	65	512/131k	N. A.
Dynap-SEL	Mixed	Yes	28	1k/64k	N. A.
SpiNNaker	Digital	Yes	130	18k/18 M	0.064
SpiNNaker2	Digital	Yes	22	Configuration	N. A.
Loihi	Digital	Yes	14	128k/128 M	< 42.4
Loihi2	Digital	Yes	7	1 M/120 M	N. A.
TrueNorth	Digital	No	28	1 M/256 M	46–400
Darwin	Digital	No	180	2048/4.19 M	N. A.
Darwin3	Digital	Yes	22	2.3 M/-	N. A.
Tianjic	Digital	No	28	39k/9.75 M	649
TianjicX	Digital	Yes	28	160k/20 M	N. A.

7. Framework, challenges and opportunities

This section delineates the challenges and opportunities confronting VSLAM in the era of HI coexistence and proposes a systematic framework to catalyze the emerging trend of cross-paradigm HI integration for future community-wide innovation.

7.1. System-level development framework

Currently, multiple HI paradigms are at varying stages of development, playing diverse roles in VSLAM research. For example, mathematical computing-based VSLAM has developed over three decades with low costs and good scalability but struggles in complex environments due to reliance on human-designed rules. AI-enabled VSLAM, driven by data and representation learning, has improved accuracy by overcoming limitations of traditional computational rules designed by human experience, but it faces high-complexity computation and weak generalization. AI-specific chips enable some real-time solutions but their high costs are impractical for low-cost robots. Brain-inspired VSLAM can theoretically excel in computational efficiency and optimize power consumption on neuromorphic hardware, but it trails traditional methods in accuracy and faces high costs.

Each of these HI paradigms has strengths and weaknesses at the software and hardware levels. Therefore, formulating mutually beneficial fusion schemes is pivotal for promoting the practical application of HI integration in the future VSLAM research. Given this, this paper proposes a VSLAM framework from the perspective of multiple HI integration (Fig. 2). This framework divides the development of a VSLAM system into the input end, algorithm end, and deployment end. It can also serve as a template to expand the input end and guide the system-level development of general SLAM systems.

The input end includes heterogeneous visual sensors, from which the VSLAM system selects one or more categories to support the algorithm end. The algorithm end makes up with the computational layer (heterogeneous computing methods) and the functional layer (VSLAM functions like VO, LCD, backend optimization and mapping). During development, suitable computing paradigms in the computational layer are selected based on input data characteristics to meet the functional layer's requirements. The functional layer can leverage hybrid computing paradigms if key technologies are coordinated to support full VSLAM functionality. Taking semi-bionic SLAM [218], it employed feature-based VO and a AlexNet for LCD, with CANN-based head--direction and place cell network to construct experience map. Liu, et al. [219] combined Yolov3 with RatSLAM to create a semantic-embedded topological experience map.

Additionally, the deployment end can use hybrid processors as needed. Some hybrid AI computing boxes have been designed to improve efficiency with joint AI computing power. The Tianjic chip and SpiNNaker2 both support the integration of AI and brain-inspired models. NeuroGPR [121] is a Tianjic-empowered example, using hybrid computing paradigms.

7.2. Challenges and opportunities

1) In the VSLAM systems, environment perception relies on various sensors, with the fusion of heterogeneous sensors enhancing robustness in complex settings. Bionic cameras excel in low-texture and varying lighting conditions, while event cameras handle high-dynamic scenes and motion blur, presenting opportunities for VSLAM system development. However, integrating heterogeneous sensors (e.g., RGB-D, panoramic, event, bionic) poses challenges due to distinct data characteristics and the need for time synchronization, algorithm adaptation, and efficient system operation.
2) Recently, the VSLAM field has seen innovative solutions empowered by advanced technologies like GNNs, SNNs, NeRF, and 3D GS, which outperform traditional AI-based methods in VO, LCD, and mapping. This has brought significant opportunities for enhancing VSLAM systems. However, most research focuses on improving a single key technology within the VSLAM framework. How can we break down the barriers between these key technologies to promote the system-level development and pragmatic applications of cross-paradigm HI integration-empowered VSLAM? Despite the continuous emergence of new technologies and ideas, research reports that have overcome this challenge are still lacking.
3) Cognitive neuroscience has revealed many neural mechanisms underlying spatial cognition and navigation. These insights drive the development of brain-inspired SLAM technologies and provide new foundations for advancing key VSLAM technologies. STDN-VO [57] simulates the dual-stream processing of the human visual system,

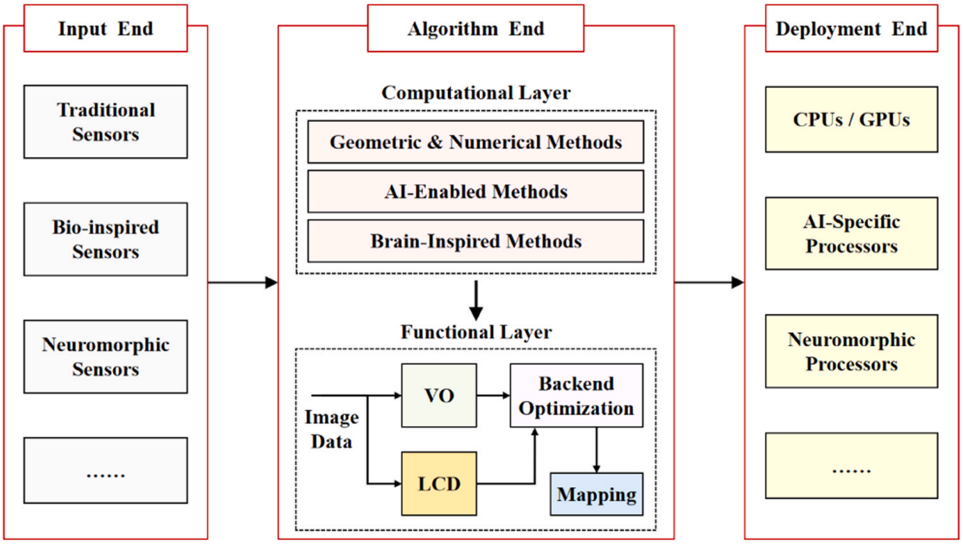


Fig. 2. The proposed system-level development framework.

NeuroSLAM leverages the path integration mechanism of navigation cells [14], and some VPR technologies are inspired by the brain's memory mechanisms [15]. However, translating these findings into reliable VSLAM applications remains challenging, involving high technical barriers in interdisciplinary research and uncertainties in cross-boundary collaboration.

4) Currently, Neumann architecture-based hardware, like Graphcore's Colossus IPU (23.6 billion transistors), is nearing its limits but faces high energy consumption. Meanwhile, neuromorphic computing offers high efficiency and power consumption advantages. Quantum processors in development also provide opportunities for VSLAM through improved computing power and energy efficiency. However, challenges remain, including incomplete toolchains for neuromorphic and quantum hardware, the huge gap from usability to practicality, and the difficulties of maintaining their ecosystems.

8. Perspectives and conclusion

8.1. Perspectives

1) This century is dubbed the "century of the brain" [220]. We propose paying close attention to neuroscience research on spatial cognition and navigation, as well as advancements in computational neuroscience, to drive the transformation of VSLAM. Animal brains explore environments and construct cognitive maps in a way that closely mirrors the SLAM process. However, animal brains rely significantly on spatial memory during place recognition, preventing erroneous results due to lighting or scene changes. Current VPR technologies struggle to replicate this capability, which is highly valuable for VSLAM systems. Similarly, the innate hierarchical map memory and adaptive navigation strategy regulation capabilities of brains are worth emulating in VSLAM.

Can the brain's adaptive navigation strategies inspire dynamic adaptive regulation when integrating various computational paradigms into a complete VSLAM system? Is it possible to construct non-single-type environmental map descriptions? For instance, in simple scenes with clear structural features, traditional mathematical methods could enhance VSLAM efficiency and create simple map descriptions. In low-texture, feature-degraded scenes, complex computational paradigms and advanced sensors could ensure stable VSLAM performance and build detailed scene maps. By reconstructing the VSLAM framework using the brain's spatial cognition mechanisms, we can promote the transition of VSLAM from tool-oriented to cognition-oriented, applicable to broader SLAM technology transformations.

2) We endorse the "dual-brain fusion" concept proposed by the Tianjic team [2]. Assuming current hardware computing power is not a limiting factor, we advocate fully leveraging heterogeneous hybrid computing in the development of next-generation SLAM systems. For example, researchers should employ suitable AI technologies to enhance VO performance, improve the LCD module, and combine these with conventional backend optimization and AI-enabled mapping methods to form a comprehensive AI-based VSLAM solution. This solution can be deployed on general-purpose or AI-specific processors without considering power consumption.

Alternatively, AI algorithms can be converted to SNNs using ANN2SNN technologies like SpikingJelly [19] and deployed on neuromorphic processors for low-power solutions. Moreover, the VO, LCD, and mapping stages are not restricted to a single computational paradigm. Traditional mathematical methods, AI-enabled methods, and brain-inspired methods can be combined as needed to form a HI integration-driven VSLAM system using platforms like Tianjic or SpiNNaker2.

3) Traditional SLAM, which requires human intervention and lack autonomy, has led to the development of active SLAM, integrating decision-making, planning, localization, and mapping but remaining SLAM-focused [221]. Embodied AI, emphasizing learning through environmental interaction and physical embodiment [4], shares common needs with active VSLAM, creating a natural connection. Developing Embodied AI, especially embodied neuromorphic intelligence [3], that mimic the brain's cognitive navigation mechanisms could offer new pathways for active VSLAM.

For example, when LCD information is missing, the system could rely more on internal state estimation (e.g., path integration) and dynamically adjust perception and action strategies based on uncertainty. This transformation requires VSLAM systems to evolve from geometric re-constructors to embodied intelligent agents integrating perception, action, memory, and adaptive learning loops, leveraging neuromorphic computing's low power consumption and real-time capabilities for more robust and intelligent active exploration and mapping.

4) SLAM is essentially a self-consistent joint estimation of metric-topological structure: it simultaneously "maps" and "localizes" in one shot, and then stops. The spatial prior becomes immutable as soon as the robot begins its moment. Dynamic obstacles or environmental changes cannot revise it, forcing the system into passive

localization and preventing any online update. To bridge this gap, the future SLAM should evolve into a task-level loop where mapping, localization, navigation and feedback run continuously. The map should grow or prune like living cells as obstacles appear or disappear, enabling the agent to instantly re-interpret space. This process is much closer to human navigation mechanism and is precisely what spatial intelligence aims to achieve [222–224].

Recently, the research of Vision-and-Language Navigation (VLN) has attracted intense academic interest. It has shifted research focus from “geometrically correct” to “cognitively plausible” pathways, attempting to instill human-like spatial intelligence into navigation agents [225]. In fact, if we regard the map itself as a special form of natural language, an interpretable semantic artifact [226], so the instruction-understanding and environment-understanding stages in VLN already exhibit the hallmarks of task-level SLAM. Consequently, contemporary VLN and active SLAM are highly overlapping endeavors. Unlike traditional SLAM, VLN pursues not a static geometric description but a continuously evolving map that is both semantic and task-oriented.

Spatial intelligence builds upon this living map a representational model that mirrors the real world, performs logical inference, and enables explanation and decision-making. The core idea behind this pipeline closely resembles the human psychological process of using natural language to specify goals and constraints, then navigating through a continuous visual stream. For instance, SLAM handles low-level geometric-topological measurement (analogous to the hippocampal-entorhinal circuit), while VLN realizes high-level semantic interpretation and linguistic reasoning (analogous to the prefrontal-language network). Spatial intelligence couples the two via world models and predictive representations (akin to the predictive map theory [227]), allowing the agent not merely to “reach a location,” but to understand “why it should reach it” and to anticipate “what might be needed next.” Thus, SLAM is upgraded from a one-shot tool to a lifelong spatial memory system, and VLN evolves from “following a map” to “looking, thinking, and revising on the fly.” Together, in embodied navigation, they converge toward a brain-like mode of spatial cognition and navigation, aligning with the developmental trajectory advocated in this paper: from tool-oriented to cognition-oriented.

9. Conclusion

In the era of HI coexistence, VSLAM benefits from both individual HI paradigm and the emerging trend of cross-paradigm integration. This paper analyzes key progress in VSLAM from individual HI paradigm and proposes a system-level framework for HI integration-driven VSLAM systems.

The deeper significance of HI integration is to break through the cognitive limitations of a single computational paradigm and move towards integration and collaboration inspired by biological general intelligence. The core challenge involves not only technical integration but also constructing effective fusion strategies across paradigms (e.g., data formats, computing paradigms, and hardware support) to achieve a synergistic outcome. This integration is expected to endow VSLAM systems with new qualities of environmental understanding, prediction, and long-term adaptation, crucial for transforming robots from simple spatial perception tools to intelligent agents with true cognitive abilities and complex interaction and autonomous learning capabilities. Ultimately, it aims to realize a paradigm conversion from "tool-based" to "cognition-based."

Furthermore, this paper also analyzes the challenges and opportunities for VSLAM innovation in the HI coexistence era, as well as prospective suggestions. It is hoped that this paper can inspire innovative ideas for the development of next-generation VSLAM systems.

CRediT authorship contribution statement

Fangwen Yu: Writing – review & editing. **Lingfei Mo:** Writing – review & editing, Resources. **Wenxuan Yin:** Writing – review & editing, Writing – original draft, Validation. **Xu He:** Writing – review & editing, Writing – original draft, Conceptualization. **Sa Su:** Writing – review & editing, Writing – original draft. **Youdong Zhang:** Writing – review & editing, Writing – original draft, Visualization. **Xiaolin Meng:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangsu Province under Grant No. BK20243064. Xu He receives support from the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant No. SJCX24_0067 & SEU Innovation Capability Enhancement Plan for Doctoral Students under Grant No. CXJH_SEU 24204.

Data availability

Data will be made available on request.

References

[1] Y. Lyu, Artificial intelligence: enabling technology to empower society, Engineering 6 (3) (2020) 205–206.

[2] J. Pei, et al., Towards artificial general intelligence with hybrid Tianjic chip architecture, Nature 572 (7767) (2019) 106–111.

[3] C. Bartolozzi, et al., Embodied neuromorphic intelligence, Nat. Commun. 13 (1) (2022) 1415.

[4] T. Zhang, et al., A survey of visual navigation: From geometry to embodied AI, Eng. Appl. Artif. Intell. 114 (2022) 105036.

[5] J. Fuentes-Pacheco, et al., Visual simultaneous localization and mapping: a survey, Artif. Intell. Rev. 43 (1) (2015) 55–81.

[6] R. Mur-Artal, et al., ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.

[7] C. Cadena, et al., Past, present, and future of simultaneous localization and mapping: toward the robust-perception age, IEEE Trans. Robot. 32 (6) (2016) 1309–1332.

[8] C. Chen, et al., Deep learning for visual localization and mapping: a survey, IEEE Trans. Neural Netw. Learn. Syst. 35 (12) (2024) 17000–17020.

[9] R. Li, et al., UnDeepVO: monocular visual odometry through unsupervised deep learning, IEEE Int. Conf. Robot. Autom. (ICRA) (2018) 7286–7291.

[10] D. Gao, et al., AirLoop: lifelong loop closure detection, Int. Conf. Robot. Autom. (ICRA) (2022) 10664–10671.

[11] C. Yu, et al., DS-SLAM: a semantic visual SLAM towards dynamic environments, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2018) 1168–1174.

[12] Y. Bai, et al., "A review of brain-inspired cognition and navigation technology for mobile robots, Cyborg Bionic Syst. 5 (2024) 0128.

[13] D. Ball, et al., OpenRatSLAM: an open source brain-based SLAM system, Auton. Robots 34 (3) (2013) 149–176.

[14] F. Yu, et al., NeuroSLAM: a brain-inspired SLAM system for 3D environments, Biol. Cybern. 113 (5-6) (2019) 515–545.

[15] P. Neubert, et al., A neurologically inspired sequence processing model for mobile robot place recognition, IEEE Robot. Autom. Lett. 4 (4) (2019) 3200–3207.

[16] S. Hussaini, et al., Spiking neural networks for visual place recognition via weighted neuronal assignments, IEEE Robot. Autom. Lett. 7 (2) (2022) 4094–4101.

[17] S. Hussaini, et al., Ensembles of compact, region-specific & regularized spiking neural networks for scalable place recognition, IEEE Int. Conf. Robot. Autom. (ICRA (2023) 4200–4207.

[18] S. Hussaini, et al., Applications of spiking neural networks in visual place recognition, IEEE Trans. Robot. 41 (2025) 518–537.

[19] W. Fang, et al., "SpikingJelly: an open-source machine learning infrastructure platform for spike-based intelligence, Sci. Adv. 9 (40) (2023) eadi1480.

[20] R. Mur-Artal, J.D. Tardós, ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras, IEEE Trans. Robot. 33 (5) (2017) 1255–1262.

[21] C. Campos, et al., ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM, IEEE Trans. Robot. 37 (6) (2021) 1874–1890.

[22] DeepCamera: advanced AI-powered video analytics for your CCTV and NVR systems. [online] Available: https://medevel.com/deepcamera-ai/

[23] Y. Liu, et al., Robust active visual SLAM system based on bionic eyes, IEEE Int. Conf. Cyborg Bionic Syst. (CBS) (2019) 340–345.

[24] K. Huang, et al., Event-based simultaneous localization and mapping: a comprehensive survey (2024). ArXiv, abs/2304.09793v2, 2024.

[25] D. Liu, et al., Spatiotemporal registration for event-based visual odometry, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2021) 4935–4944.

[26] J. Zhang, et al., Event-based sensor fusion and application on odometry: a survey, IEEE 6th Int. Conf. Image Process. Appl. Syst. (IPAS) (2025) 1–6.

[27] L. Deng, et al., Tianjic: a unified and scalable chip bridging spike-based and continuous neural computation, IEEE J. SolidState Circuits 55 (8) (2020) 2228–2246.

[28] H. Gonzalez, et al., SpiNNaker2: a large-scale neuromorphic system for Event-Based and asynchronous machine learning (2024). ArXiv, abs/2401.04491.

[29] T. Taketomi, et al., Visual SLAM algorithms: a survey from 2010 to 2016, IPSJ Trans. Comput. Vis. Appl. 9 (1) (2017) 16.

[30] I.A. Kazerouni, et al., A survey of state-of-the-art on visual SLAM, Expert Syst. Appl. 205 (2022) 117734.

[31] M.R.U. Saputra, et al., Visual SLAM and structure from motion in dynamic environments: a survey, ACM Comput. Surv. 51 (2) (2018) 1–36.

[32] F. Tosi, et al., How Nerfs and 3D gaussian splatting are reshaping SLAM: a survey, arXiv, arxiv:2402.13255 (2024).

[33] S. Mokssit, et al., Deep learning techniques for visual SLAM: a survey, IEEE Access 11 (2023) 20026–20050.

[34] G. Gallego, et al., Event-based vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2022) 154–180.

[35] K. Tsintotas, et al., The revisiting problem in simultaneous localization and mapping: a survey on visual loop closure detection, IEEE Trans. Intell. Transp. Syst. 23 (11) (2022) 19929–19953.

[36] X. Zhang, et al., Visual place recognition: a survey from deep learning perspective, Pattern Recognit. 113 (2021) 107760.

[37] S. Lowry, et al., Visual place recognition: a survey, IEEE Trans. Robot. 32 (1) (2015) 1–19.

[38] J. Bongard, "Probabilistic robotics, Artif. Life 14 (2) (2008) 227–229.

[39] L. Xu, et al., EPLF-VINS: real-time monocular visual-inertial SLAM with efficient point-line flow features, IEEE Robot. Autom. Lett. 8 (2) (2023) 752–759.

[40] G. Zhang, et al., Building a 3-D line-based map using stereo SLAM, IEEE Trans. Robot. 31 (6) (2015) 1364–1377.

[41] Q. Wang, et al., Line flow based simultaneous localization and mapping, IEEE Trans. Robot. 37 (5) (2021) 1416–1432.

[42] H. Zhou, et al., StructSLAM: visual SLAM with building structure lines, IEEE Trans. Veh. Technol. 64 (4) (2015) 1364–1375.

[43] A.J. Davison, et al., MonoSLAM: real-time single camera SLAM, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1052–1067.

[44] D. Cai, et al., A comprehensive overview of core modules in visual SLAM framework, Neurocomputing 590 (2024) 127760.

[45] R.A. Newcombe, et al., DTAM: dense tracking and mapping in real-time. International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 2320–2327.

[46] J. Engel, et al., Large-scale direct SLAM with stereo cameras, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2015) 1935–1942.

[47] J. Engel, et al., Direct sparse odometry, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2018) 611–625.

[48] X. Yang, et al., FD-SLAM: 3-D reconstruction using features and dense matching, Int. Conf. Robot. Autom. (ICRA) (2022) 8040–8046.

[49] P. Tanskanen, et al., Semi-direct EKF-based monocular visual-inertial odometry, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2015) 6073–6078.

[50] C. Forster, et al., SVO: semidirect visual odometry for monocular and multicamera systems, IEEE Trans. Robot. 33 (2) (2017) 249–265.

[51] K. Wang, et al., Approaches, challenges, and applications for deep visual odometry: toward complicated and emerging areas, IEEE Trans. Cogn. Dev. Syst. 14 (1) (2022) 35–49.

[52] A. Krizhevsky, et al., ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[53] A. Kendall, et al., PoseNet: a convolutional network for real-time 6-DOF camera relocalization, IEEE Int. Conf. Comput. Vis. (ICCV) (2015) 2938–2946.

[54] S. Wang, et al., DeepVO: towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks, IEEE Int. Conf. Robot. Autom. (ICRA) (2017) 2043–2050.

[55] S. Shen, et al., DytanVO: joint refinement of visual odometry and motion segmentation in dynamic environments, IEEE Int. Conf. Robot. Autom. (ICRA) (2023) 4048–4055.

[56] Y. Almalioglu, et al., GANVO: unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks, Int. Conf. Robot. Autom. (ICRA) (2019) 5474–5480.

[57] C. Xu, et al., Spatiotemporal dual-stream network for visual odometry, IEEE Robot. Autom. Lett. 10 (4) (2025) 3867–3874.

[58] T. Zhou, et al., Unsupervised learning of depth and ego-motion from video, IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2017) 6612–6619.

[59] Z. Yin, et al., GeoNet: unsupervised learning of dense depth, optical flow and camera pose, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2018) 1983–1992.

[60] Z. Wang, et al., Unsupervised scale network for monocular relative depth and visual odometry, IEEE Trans. Instrum. Meas. 73 (2024) 1–16.

[61] S. Kannapiran, et al., Stereo visual odometry with deep learning-based point and line feature matching using an attention graph neural network, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2023) 3491–3498.

[62] N. Yang, et al., D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2020) 1278–1289.

[63] Y. Xie, et al., InstanceVO: self-supervised semantic visual odometry by using metric learning to incorporate geometrical priors in instance objects, IEEE Robot. Autom. Lett. 9 (11) (2024) 10708–10715.

[64] A. Abouee, et al., Weakly supervised End2End deep visual odometry, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW) (2024) 858–865.

[65] Z. Liu, et al., Adaptive learning for hybrid visual odometry, IEEE Robot. Autom. Lett. 9 (8) (2024) 7341–7348.

[66] G. Lu, Deep unsupervised visual odometry via bundle adjusted pose graph optimization, IEEE Int. Conf. Robot. Autom. (ICRA) (2023) 6131–6137.

[67] H. Zhan, et al., Visual odometry revisited: what should be learnt? IEEE Int. Conf. Robot. Autom. (ICRA) (2020) 4203–4210.

[68] R. Song, et al., GraphAVO: self-supervised visual odometry based on graph-assisted geometric consistency, IEEE Trans. Intell. Transp. Syst. 25 (12) (2024) 20673–20682.

[69] Z. Teed, et al., Deep patch visual odometry, Adv. Neural Inf. Process. Syst. (NeurIPS) (2022).

[70] Z. Teed, et al., DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras, Adv. Neural Inf. Process. Syst. (NeurIPS) (2021).

[71] L. Lipson, et al., Deep patch visual SLAM, Eur. Conf. Comput. Vis. (ECCV) (2024).

[72] H. Jun, et al., EventPointNet: robust keypoint detection with neuromorphic camera data, Int. Conf. Control Autom. Syst. (ICCAS (2022) 829–834.

[73] A. Hadviger, et al., Feature-based event stereo visual odometry, Eur. Conf. Mob. Robots (ECMR) (2021) 1–6.

[74] Y. Zhou, et al., Event-based stereo visual odometry, IEEE Trans. Robot. 37 (5) (2021) 1433–1450.

[75] W. Guan, et al., Monocular event visual inertial odometry based on event-corner using sliding windows graph-based optimization, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2022) 2438–2445.

[76] J. Wang, et al., Event-based stereo visual odometry with native temporal resolution via continuous-time Gaussian process regression, IEEE Robot. Autom. Lett. 8 (10) (2023) 6707–6714.

[77] J. Hidalgo-Carrió, et al., Event-aided direct sparse odometry, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2022) 5771–5780.

[78] R. Pellerito, et al., Deep visual odometry with events and frames, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2024) 8966–8973.

[79] S. Zhu, et al., Event camera-based visual odometry for dynamic motion tracking of a legged robot using adaptive time surface, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2023) 3475–3482.

[80] R. Yuan, et al., EVIT: event-based visual-inertial tracking in semi-dense maps using windowed nonlinear optimization, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2024) 10656–10663.

[81] G. Gong, et al., TEVIO: thermal-aided event-based visual inertial odometry for robust state estimation in challenging environments, IEEE Transactions on Instrumentation and Measurement.

[82] P. Chen, et al., ESVIO: event-based stereo visual inertial odometry, IEEE Robot. Autom. Lett. 8 (6) (2023) 3661–3668.

[83] H. Huang, et al., 360VO: visual odometry using a single 360 camera, Int. Conf. Robot. Autom. (ICRA) (2022) 5594–5600.

[84] Y. Luo, et al., FD-SLAM: a semantic SLAM based on enhanced fast-SCNN dynamic region detection and DeepFillv2-Driven background inpainting, IEEE Access 11 (2023) 110615–110626.

[85] S. Kim, et al., Spiking-YOLO: spiking neural network for energy-efficient object detection, AAAI Conf. Artif. Intell. (AAAI) (2019).

[86] J. Niu, et al., ESVO2: direct visual-inertial odometry with stereo event cameras, IEEE Transactions on Robotics.

[87] G. Klein, et al., Parallel tracking and mapping for small AR workspaces, IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007, pp. 225–234.

[88] Sivic, et al., Video google: a text retrieval approach to object matching in videos, IEEE Int. Conf. Comput. Vis. (ICCV) 2 (2003) 1470–1477.

[89] M. Cummins, et al., Appearance-only SLAM at large scale with FAB-MAP 2.0, Robotics Science Systems (2009).

[90] G. Grisetti, et al., A tutorial on graph-based SLAM, IEEE Intell. Transp. Syst. Mag. 2 (4) (2010) 31–43.

[91] D. Galvez-López, et al., Bags of binary words for fast place recognition in image sequences, IEEE Trans. Robot. 28 (5) (2012) 1188–1197.

[92] A. Loquercio, et al., Efficient descriptor learning for large scale localization, IEEE Int. Conf. Robot. Autom. (ICRA) (2017) 3170–3177.

[93] R. Arandjelovic, et al., NetVLAD: CNN architecture for weakly supervised place recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1437–1451.

[94] Y. Huang, et al., VLAD-based loop closure detection for monocular SLAM, IEEE Int. Conf. Inf. Autom. (ICIA) (2016) 511–516.

[95] M. Aissi, et al., VIPER: visual perception and explainable reasoning for sequential decision-making (2025). ArXiv, abs/2503.15108.

[96] J. Ma, et al., Fast and robust loop-closure detection via convolutional auto-encoder and motion consensus, IEEE Trans. Ind. Inform. 18 (6) (2022) 3681–3691.

[97] A. Memon, et al., Loop closure detection using supervised and unsupervised deep neural networks for monocular SLAM systems, Robot. Auton. Syst. 126 (2020) 103470.

[98] Y. Wang, et al., GOReloc: graph-based object-level relocalization for visual SLAM, IEEE Robot. Autom. Lett. 9 (10) (2024) 8234–8241.
[99] H. Osman, et al., LoopNet: where to Focus? Detecting loop closures in dynamic scenes, IEEE Robot. Autom. Lett. 7 (2) (2022) 2031–2038.
[100] Y. Zhou, et al., A visual SLAM loop closure detection method based on lightweight siamese capsule network, Sci. Rep. 15 (1) (2025) 7644.
[101] Y. Ming, et al., VIPeR: visual incremental place recognition with adaptive mining and continual learning, IEEE Robot. Autom. Lett. 10 (3) (2025) 3038–3045.
[102] H. Qian, et al., I2KEN: intra-domain and inter-domain knowledge enhancement network for lifelong loop closure detection, IEEE Robot. Autom. Lett. 10 (9) (2025) 8722–8729.
[103] G. Singh, et al., Hierarchical loop closure detection for long-term visual SLAM with semantic-geometric descriptors, IEEE Int. Intell. Transp. Syst. Conf. (ITSC) (2021) 2909–2916.
[104] H. Osman, et al., PlaceNet: a multi-scale semantic-aware model for visual loop closure detection, Eng. Appl. Artif. Intell. 119 (2023) 105797.
[105] Y. Ming, et al., AEGIS-Net: attention-guided multi-level feature aggregation for indoor place recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2024, pp. 4030–4034.
[106] Y. Ming, et al., CGiS-Net: aggregating colour, geometry and implicit semantic features for indoor place recognition, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2022) 6991–6997.
[107] B. Li, et al., TextSLAM: sisual SLAM with semantic planar text features, IEEE Trans. Pattern Anal. Mach. Intell. 46 (1) (2024) 593–610.
[108] Y. Ming, et al., SLC^2-SLAM: semantic-guided loop closure using shared latent code for NeRF SLAM, IEEE Robot. Autom. Lett. 10 (5) (2025) 4978–4985.
[109] H. Chen, et al., Semantic loop closure detection with instance-level inconsistency removal in dynamic industrial scenes, IEEE Trans. Ind. Inform. 17 (3) (2021) 2030–2040.
[110] J. Yu, et al., SemanticLoop: loop closure with 3D semantic graph matching, IEEE Robot. Autom. Lett. 8 (2) (2023) 568–575.
[111] J. Kim, et al., Closing the loop: Graph networks to unify semantic objects and visual features for multi-object scenes, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2022) 4352–4358.
[112] J. Kim, et al., SymbioLCD: ensemble-based loop closure detection using CNN-extracted objects and visual Bag-of-Words, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2021), 5425-5425.
[113] T. Fischer, et al., How many events do you need? Event-Based visual place recognition using sparse but varying pixels, IEEE Robot. Autom. Lett. 7 (4) (2022) 12275–12282.
[114] H. Lee, et al., Ev-ReconNet: visual place recognition using event camera with spiking neural networks, IEEE Sens. J. 23 (17) (2023) 20390–20399.
[115] U. Akcal, et al., LoCS-Net: localizing convolutional spiking neural network for fast visual place recognition, Front. Neurorobotics 18 (2025) 1490267.
[116] A. Hines, et al., VPRTempo: a fast temporally encoded spiking neural network for visual place recognition, IEEE Int. Conf. Robot. Autom. (ICRA) (2024) 10200–10207.
[117] L. Zhu, et al., "Neuromorphic sequence learning with an event camera on routes through vegetation, Sci. Robot. 8 (82) (2023) eadg3679.
[118] S. Schubert, et al., Towards combining a neocortex model with entorhinal grid cells for mobile robot localization, Eur. Conf. Mob. Robots (ECMR) (2019) 1–8.
[119] A. Ozdemir, et al., EchoVPR: echo state networks for visual place recognition, IEEE Robot. Autom. Lett. 7 (2) (2022) 4520–4527.
[120] X. Luo, et al., Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection, Eur. Conf. Comput. Vis. (ECCV) (2024).
[121] F. Yu, et al., Brain-inspired multimodal hybrid neural network for robot place recognition, Sci. Robot. 8 (78) (2023) eabm6996.
[122] B. McNaughton, et al., Path integration and the neural basis of the 'cognitive map, Nat. Rev. Neurosci. 7 (8) (2006) 663–678.
[123] T. Bailey, et al., Consistency of the EKF-SLAM algorithm, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2006) 3562–3568.
[124] M. Labbé, et al., RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation, J. Field Robot. 36 (2) (2019) 416–446.
[125] C. Wang, et al., PyPose: a library for robot learning with physics-based optimization, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2023) 22024–22034.
[126] X. Gao, et al., LDSO: direct sparse odometry with loop closure, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2018) 2198–2204.
[127] F. Endres, et al., "3-D mapping with an RGB-D camera, IEEE Trans. Robot. 30 (1) (2014) 177–187.
[128] M. Kaess, et al., iSAM: incremental smoothing and mapping, IEEE Trans. Robot. 24 (6) (2008) 1365–1378.
[129] M. Kaess, et al., iSAM2: incremental smoothing and mapping with fluid relinearization and incremental variable reordering, IEEE Int. Conf. Robot. Autom. (ICRA) (2011) 3281–3288.
[130] X. Wang, et al., AprilSAM: real-time smoothing and mapping, IEEE Int. Conf. Robot. Autom. (ICRA) (2018) 2486–2493.
[131] M. Hsiao, et al., MH-iSAM2: mMulti-hypothesis iSAM using bayes tree and hypo-tree, Int. Conf. Robot. Autom. (ICRA) (2019) 1274–1280.
[132] Y. Zhang, et al., MR-iSAM2: incremental smoothing and mapping with multi-root bayes tree for multi-robot SLAM, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2021) 8671–8678.
[133] T. Zeng, et al., NeuroBayesSLAM: neurobiologically inspired Bayesian integration of multisensory information for robot navigation, Neural Netw. 126 (2020) 21–35.
[134] V. Edvardsen, Goal-directed navigation based on path integration and decoding of grid cells in an artificial neural network, Nat. Comput. 18 (1) (2019) 13–27.
[135] M. Milford, et al., Mapping a suburb with a single camera using a biologically inspired SLAM system, IEEE Trans. Robot. 24 (5) (2008) 1038–1053.
[136] T. Zeng, et al., Cognitive mapping based on conjunctive representations of space and movement, Front. Neurorobotics 11 (2017) 61.
[137] L. Silveira, et al., An open-source bio-inspired solution to underwater SLAM, IFAC-PapersOnLine 48 (2) (2015) 212–217.
[138] D. Shen, et al., ORB-NeuroSLAM: a brain-inspired 3D SLAM system based on ORB features, IEEE Internet Things J. 11 (7) (2024) 12408–12418.
[139] R. Eyvazpour, et al., Hardware implementation of SLAM algorithms: a survey on implementation approaches and platforms, Artif. Intell. Rev. 56 (7) (2023) 6187–6239.
[140] A. Kumar, et al., High-speed stereo visual SLAM for low-powered computing devices, IEEE Robot. Autom. Lett. 9 (1) (2024) 499–506.
[141] R. Kreiser, et al., A neuromorphic approach to path integration: A head-direction spiking neural network with vision-driven reset, IEEE Int. Symp Circuits Syst. (ISCAS) (2018) 1–5.
[142] J. Yoon, et al., NeuroSLAM: a 65-nm 7.25-to-8.79-TOPS/W mixed-signal oscillator-based SLAM accelerator for edge robotics, IEEE J. SolidState Circuits 56 (1) (2021) 66–78.
[143] K. Tateno, et al., CNN-SLAM: real-time dense monocular SLAM with learned depth prediction, IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2017) 6565–6574.
[144] G. Grisetti, et al., Improving grid-based SLAM with Rao-Blackwellized particle filters by adaptive proposals and selective resampling, IEEE International Conference on Robotics and Automation (ICRA, 2005, pp. 2432–2437.
[145] D. Lyu, et al., SpOctA: a 3D sparse convolution accelerator with octree-encoding-based map search and inherent sparsity-aware processing, IEEE/ACM Int. Conf. Comput. Aided Des. (ICCAD) (2023) 1–9.
[146] M. Milford, et al., Feature-based visual odometry and featureless place recognition for SLAM in 2.5D environments, Australas. Conf. Robot. Autom. (2011) 1–8.
[147] Y. Wang, et al., A survey of visual SLAM in dynamic environment: the evolution from geometric to semantic approaches, IEEE Trans. Instrum. Meas. 73 (2024) 1–21.
[148] B. Bescos, et al., DynaSLAM: tracking, mapping, and inpainting in dynamic scenes, IEEE Robot. Autom. Lett. 3 (4) (2018) 4076–4083.
[149] F. Zhong, et al., Detect-SLAM: making object detection and SLAM mutually beneficial, IEEE Winter Conf. Appl. Comput. Vis. (WACV) (2018) 1001–1010.
[150] S. Wen, et al., Dynamic SLAM: a visual SLAM in outdoor dynamic scenes, IEEE Trans. Instrum. Meas. 72 (2023) 1–11.
[151] W. Wu, et al., YOLO-SLAM: a semantic SLAM system towards dynamic environment with geometric constraint, Neural Comput. Appl. 34 (8) (2022) 6011–6026.
[152] S. Yang, et al., CubeSLAM: monocular 3-D object SLAM, IEEE Trans. Robot. 35 (4) (2019) 925–938.
[153] Y. Fan, et al., Blitz-SLAM: a semantic SLAM in dynamic environments, Pattern Recognit. 121 (2025) 108225.
[154] G. Csurka, et al., Semantic image segmentation: two decades of research (2022).
[155] Y. Mo, et al., Review the state-of-the-art technologies of semantic segmentation based on deep learning, Neurocomputing 493 (2022) 626–646.
[156] Z. Zou, et al., Object detection in 20 years: a survey, Proc. IEEE 111 (3) (2023) 257–276.
[157] B. Zhang, et al., Affinity attention graph neural network for weakly supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2022) 8082–8096.
[158] T. Zhou, et al., Group-wise learning for weakly supervised semantic segmentation, IEEE Trans. Image Process. 31 (2022) 799–811.
[159] Z. Xu, et al., Category-guided graph convolution network for semantic segmentation, IEEE Trans. Netw. Sci. Eng. 11 (6) (2024) 6080–6089.
[160] Z. Du, et al., MVF-GNN: multi-View fusion with GNN for 3D semantic segmentation, IEEE Robot. Autom. Lett. 10 (4) (2025) 3262–3269.
[161] S. Dong, et al., Semantic-context graph network for point-based 3D object detection, IEEE Trans. Circuits Syst. Video Technol. 33 (11) (2023) 6474–6486.
[162] C. Li, et al., DAGCN: dynamic and adaptive graph convolutional network for salient object detection, IEEE Trans. Neural Netw. Learn. Syst. 35 (6) (2024) 7612–7626.
[163] R. Zhang, et al., Accurate and efficient event-based semantic segmentation using adaptive spiking encoder-decoder network, IEEE Trans. Neural Netw. Learn. Syst. (2025).
[164] Z. Jia, et al., Event-based semantic segmentation with posterior attention, IEEE Trans. Image Process. 32 (2023) 1829–1842.
[165] B. Xie, et al., EISNet: a multi-modal fusion network for semantic segmentation with events and images, IEEE Trans. Multimed. 26 (2024) 8639–8650.
[166] Q. Su, et al., Deep directly-trained spiking neural networks for object detection, IEEE/CVF Int. Conf. Comput. Vis. (ICCV) (2023) 6532–6542.
[167] C. Iaboni, et al., Event-based spiking neural networks for object detection: a review of datasets, architectures, learning rules, and implementation, IEEE Access 12 (2024) 180532–180596.
[168] Y. Mehan, et al., QueSTMaps: queryable semantic topological maps for 3D scene understanding, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2024) 13311–13317.

[169] J. McCormac, et al., SemanticFusion: dense 3D semantic mapping with convolutional neural networks, IEEE Int. Conf. Robot. Autom. (ICRA) (2017) 4628–4635.

[170] L. Nicholson, et al., QuadricSLAM: dual quadrics from object detections as landmarks in object-oriented SLAM, IEEE Robot. Autom. Lett. 4 (1) (2019) 1–8.

[171] L. Xia, et al., A survey of image semantics-based visual simultaneous localization and mapping: application-oriented solutions to autonomous navigation of mobile robots, Int. J. Adv. Robot. Syst. 17 (3) (2020).

[172] K. Chen, et al., Semantic visual simultaneous localization and mapping: a survey (2022). ArXiv, abs/2209.06428.

[173] M. Bloesch, et al., CodeSLAM-Learning a compact, optimisable representation for dense visual SLAM, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2560–2568.

[174] L. Yen-Chen, et al., iNeRF: inverting neural radiance fields for pose estimation, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021) 1323–1330.

[175] E. Sucar, et al., iMAP: implicit mapping and positioning in real-time, IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 6209–6218.

[176] Z. Zhu, et al., NICE-SLAM: neural implicit scalable encoding for SLAM, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 12776–12786.

[177] X. Yang, et al., Vox-Fusion: dense tracking and mapping with voxel-based neural implicit representation, IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (2022) 499–507.

[178] K. Wang, et al., Mixed spiking NeRF: towards a more efficient neural radiance fields, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2025, pp. 1–5.

[179] S. Klenk, et al., E-NeRF: neural radiance fields from a moving event camera, IEEE Robot. Autom. Lett. 8 (3) (2023) 1587–1594.

[180] Y. Qi, et al., E2NeRF: event enhanced neural radiance fields from blurry images, IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 13208–13218.

[181] C. Yan, et al., GS-SLAM: dense visual SLAM with 3D Gaussian splatting, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 19595–19604.

[182] H. Huang, et al., Photo-SLAM: real-time simultaneous localization and photorealistic mapping for monocular, stereo, and RGB-D cameras, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 21584–21593.

[183] N. Keetha, et al., SplaTAM: splat, track & map 3D gaussians for dense RGB-D SLAM, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 21357–21366.

[184] J. Jeong, et al., EOGS: event only 3D gaussian splatting for 3D reconstruction, Int. Conf. Control Autom. Syst. (ICCAS (2024) 1550–1551.

[185] J. Wu, et al., EV-GS: event-based gaussian splatting for efficient and accurate radiance field rendering, IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2024, pp. 1–6.

[186] H. Deguchi, et al., E2GS: event enhanced gaussian splatting, IEEE Int. Conf. Image Process. (ICIP) (2024) 1676–1682.

[187] H. Xie, et al., Semi-direct multimap SLAM system for real-time sparse 3-D map reconstruction, IEEE Trans. Instrum. Meas. 72 (2023) 1–13.

[188] J. Luo, et al., BE-SLAM: BEV-enhanced dynamic semantic SLAM with static object reconstruction, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2024) 10105–10112.

[189] J. Ross, et al., BEV-SLAM: building a globally-consistent world map using monocular vision, IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) (2022) 3830–3836.

[190] Y. Liu, et al., Real-time robust stereo visual SLAM system based on bionic eyes, IEEE Trans. Med. Robot. Bionics 2 (3) (2020) 391–398.

[191] S. Wu, et al., Artificial compound eye: a survey of the state-of-the-art, Artif. Intell. Rev. 48 (4) (2017) 573–603.

[192] S. Dong, et al., Spike camera and its coding methods, Data Compress. Conf. (DCC) (2017) 437. -437.

[193] P. Dudek, et al., Sensor-level computer vision with pixel processor arrays for agile robots, Sci. Robot. 7 (76) (2022) eabl7755.

[194] T. Fischer, et al., Event-based visual place recognition with ensembles of temporal windows s, IEEE Robot. Autom. Lett. 5 (4) (2020) 6924–6931.

[195] M. Talib, et al., A systematic literature review on hardware implementation of artificial intelligence algorithms, J. Supercomput. 77 (2) (2021) 1897–1938.

[196] O. Silveira, et al., Evaluating a visual simultaneous localization and mapping solution on embedded platform. International Symposium on Industrial Electronics (ISIE), IEEE, 2020, pp. 530–535.

[197] A. Seewald, et al., RB5 low-cost explorer: implementing autonomous long-term exploration on low-cost robotic hardware, IEEE Int. Conf. Robot. Autom. (ICRA) (2024) 5977–5983.

[198] Brain processing unit—artificial brain tissue APIs. [online] Available: https://www.creativeapplications.net/robotics/brain-processing-unit-artificial-brain-tissue-apis/

[199] Introducing the Colossus? MK2 GC200 IPU. [online] Available: https://www.graphcore.ai/products/ipu

[200] B. Benjamin, et al., Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations, Proc. IEEE 102 (5) (2014) 699–716.

[201] A. Neckar, et al., Braindrop: a mixed-signal neuromorphic architecture with a dynamical systems-based programming model, Proc. IEEE 107 (1) (2019) 144–164.

[202] Y. Stradmann, et al., Demonstrating analog inference on the BrainScaleS-2 mobile system, IEEE Open J. Circuits Syst. 3 (2022) 252–262.

[203] E. Painkras, et al., SpiNNaker: a 1-W 18-Core system-on-chip for massively-parallel neural network simulation, IEEE J. SolidState Circuits 48 (8) (2013) 1943–1953.

[204] S. Moradi, et al., A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs), IEEE Trans. Biomed. Circuits Syst. 12 (1) (2018) 106–122.

[205] C. Thakur, et al., Large-scale neuromorphic spiking array processors: a quest to mimic the brain, Front. Neurosci. 12 (2018) 891.

[206] O. Richter, et al., DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor, Neuromorphic Comput. Eng. 4 (1) (2024) 014003.

[207] M. Davies, et al., Loihi: a neuromorphic manycore processor with on-chip learning, IEEE Micro 38 (1) (2018) 82–99.

[208] G. Brayshaw, et al., A neuromorphic system for the real-time classification of natural textures, IEEE Int. Conf. Robot. Autom. (ICRA) (2024) 1070–1076.

[209] F. Akopyan, et al., TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 34 (10) (2015) 1537–1557.

[210] D. Ma, et al., Darwin: a neuromorphic hardware co-processor based on spiking neural networks, J. Syst. Archit. 77 (2017) 43–51.

[211] D. Ma, et al., Darwin3: a large-scale neuromorphic chip with a Novel ISA and On-Chip Learning, Natl. Sci. Rev. 11 (5) (2023) nwae102.

[212] Q. Memon, et al., Quantum computing: navigating the future of computation, challenges, and technological breakthroughs, Quantum Rep. 6 (4) (2024) 627–663.

[213] Expanding the IBM quantum roadmap to anticipate the future of quantum-centric supercomputing. [online] Available: https://www.ibm.com/quantum/blog/ibm-quantum-roadmap-2025

[214] Meet Willow, our state-of-the-art quantum chip. [online] Available: https://blog.google/technology/research/google-willow-quantum-chip/

[215] PsiQuantum Team. A manufacturable platform for photonic quantum computing. Nature, Feb. 2025

[216] T. Lillicrap, et al., Backpropagation and the brain, Nat. Rev. Neurosci. 21 (6) (2020) 335–346.

[217] "Intel reveals world’s biggest ‘brain-inspired’ neuromorphic computer," [Online] Available: https://www.newscientist.com/article/2426523-intel-reveals-worlds-biggest-brain-inspired-neuromorphic-computer/

[218] D. Liu, et al., Semi-bionic SLAM Based on visual odometry and deep learning network, IEEE Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER, 2021, pp. 293–299.

[219] X. Liu, et al., Vision-IMU multi-sensor fusion semantic topological map based on RatSLAM, Measurement 220 (2023) 113335.

[220] K. Amunts, et al., The coming decade of digital brain research: a vision for neuroscience at the intersection of technology and computing, Imaging Neurosci. 2 (2024) 1–35.

[221] J. Placed, et al., A survey on active simultaneous localization and mapping: state of the art and new frontiers, IEEE Trans. Robot. 39 (3) (2023) 1686–1705.

[222] Y. Zhang, et al., Vision-and-language navigation today and tomorrow: a survey in the era of foundation models, Trans. Mach. Learn. Res. (2024).

[223] P. Anderson, et al., Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2018) 3674–3683.

[224] J. Feng, et al., A survey of large language model-powered spatial intelligence across scales: advances in embodied agents, smart cities, and earth science, ArXiv, abs/2504.09848 (2025).

[225] B. Yin, et al., Spatial mental modeling from limited views (2025). ArXiv, abs/2506.21458.

[226] R.A. Epstein, et al., The cognitive map in humans: spatial navigation and beyond, Nat. Neurosci. 20 (11) (2017) 1504–1513.

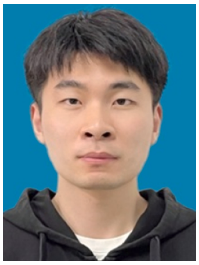
[227] K.L. Stachenfeld, et al., The hippocampus as a predictive map, Nat. Neurosci. 20 (11) (2017) 1643–1653.

Sa Su is with the School of Instrument Science and Engineering, Southeast University, Nanjing, China (e-mail: 220233684@seu.edu.cn). She is pursuing a M.Sc. degree with the School of Instrument Science and Engineering, Southeast University. Her research interests focus on AI and visual SLAM.

Xu He is with the School of Instrument Science and Engineering, Southeast University, Nanjing, China (e-mail: hexu@seu.edu.cn). He is pursuing a Ph.D. degree with the School of Instrument Science and Engineering of Southeast University. He has authored 6 JCR Q1/Q1 TOP papers as the first author. His research interests focus on brain-inspired navigation, intelligent PNT, and AGI. He is a student member of IEEE, and CCF.

Xiaolin Meng, is with the School of Instrument Science and Engineering, Southeast University, Nanjing, China, as Chair Professor of Intelligent Mobility (e-mail: xiaolin_meng@seu.edu.cn). He was Professor of Intelligent Mobility and a Theme Leader of positioning and navigation technologies with the University of Nottingham, Nottingham, U.K., until May 2020. He has authored more than 400 publications and undertaken research portfolio of more than £ 20 m. His research interests mainly include intelligent transportation systems and smart cities, connected and autonomous vehicles, structural health monitoring of large infrastructure, and precision agriculture/livestock farming. Dr. Meng is the first-ever Professor of Intelligent Mobility in the UK. He has been selected by the UK's Research Council as Academic Fellow, and is a Fellow of the RIN and IET, and a Chartered Engineer (CEng) of the UK's Engineering Council. He is also a Chinese National distinguished professor. He received the ION Best Paper Award in 1999 and the 2019 UK Engineer "Collaborate to Innovate" Award. He is the founder of the Sino-UK Geospatial Engineering Center and also the founder and chief scientist of UbiPOS UK LTD.

Youdong Zhang is with the School of Instrument Science and Engineering, Southeast University, Nanjing, China (e-mail: ydzhang@seu.edu.cn). He is pursuing a Ph.D. degree with the School of Instrument Science and Engineering of Southeast University. His research interests focus on brain-inspired navigation and brain-inspired computing.



Wenxuan Yin is with the School of Instrument Science and Engineering, Southeast University, Nanjing, China (e-mail: wenxuan_yin@seu.edu.cn). He is pursuing a Ph.D. degree with the School of Instrument Science and Engineering of Southeast University. His research interests focus on brain-inspired navigation and intelligent control.

Lingfei Mo is with the School of Instrument Science and Engineering, Southeast University, Nanjing, China (e-mail: lfmo@seu.edu.cn). He is an Associate Professor at the School of Instrument Science and Engineering, Southeast University, Nanjing, China. His research interests focus on IoT, AI, brain-inspired navigation and brain-like computing. He has authored over 100 technical publications in standard journals and conferences and has more than 40 China patents in the area of IoT and AI. Dr. Mo is also a member of the IEEE and an Associate Editor of the IEEE Journal of Radio Frequency Identification (RFID).

Fangwen Yu is currently an Assistant Research Fellow in the Center for Brain Inspired Computing Research and the Department of Precision Instrument, Tsinghua University, China (e-mail: yufangwen@tsinghua.edu.cn). He received his Ph.D. from the China University of Geosciences, Wuhan, China. His research interests include brain-inspired 3D navigation, neuromorphic robotics, etc. He is the first author of the cover article of "Science Robotics" (May. 2023), and the designer of NeuroSLAM and NeuroGPR. He has published over 30 papers in "Science Robotics", "Nature Communications", etc. and has been granted more than 10 China patents. Dr. Yu has served as a guest editor for international journals such as "Satellite Navigation", and as the chair of the brain-inspired navigation forum at international navigation conferences such as IPIN and UPINLBS. He is also a member of IEEE, the IEEE Robotics and Automation Society (RAS) and the Royal Institute of Navigation (RIN).