4.4 Localization and Mapping

\subsection{Localization and Mapping} While 3 D reconstruction focuses on recovering the spatial structure of the environment, autonomous UAV operation additionally requires continuous estimation of the vehicle state and consistent maintenance of the surrounding map. These capabilities are provided by localization and mapping systems, which estimate the UAV pose from onboard sensor observations while simultaneously constructing spatial representations of the environment. Traditional UAV navigation systems rely heavily on the Global Navigation Satellite System (GNSS) for global positioning. However, GNSS signals are often unreliable or unavailable in many real-world environments, such as urban canyons, dense forests, indoor spaces, or underground structures, where signal blockage and multipath effects degrade positioning accuracy (see Fig.~\ref{fig: gnss}). In these GNSS-denied scenarios, UAVs must estimate their motion using onboard sensors such as cameras, LiDAR, and inertial measurement units (IMU). By integrating multi-sensor observations, localization and mapping systems enable UAVs to estimate their pose relative to the environment while maintaining a consistent map to reduce long-term drift. \begin{figure} \centering \includegraphics[width=1\linewidth]{fig/GNSS. Pdf} \caption{Comparison between GNSS-available and GNSS-denied environments. While GNSS provides reliable global reference in open areas, signal blockage and multipath effects in urban, forest, indoor, and underground settings render it ineffective. This challenge necessitates a fundamental shift towards onboard perception systems using cameras, LiDAR, and IMUs for autonomous localization.} \label{fig: gnss} \end{figure} \subsubsection{State Estimation: From Single-modal Tracking to Multi-modal Synergy} State estimation aims to recover the trajectory of the UAV during flight. Existing approaches can be broadly divided into local odometry estimation and global geo-localization. \begin{figure} \centering \includegraphics[width=1\linewidth]{Localization/odometry. Pdf} \caption{The evolution of state estimation techniques. The development progresses from Visual-Inertial Odometry (VIO) to LiDAR-Inertial Odometry (LIO), and ultimately to LiDAR-Visual-Inertial (L-V-I) multi-modal fusion. Each paradigm shift is driven by the need to overcome specific single-sensor limitations—such as VIO's vulnerability to low-light/texture-less scenarios and LIO's susceptibility to geometric degeneracy. Currently, synergistic multi-modal fusion is the dominant trend for achieving highly robust state estimation.} \label{fig: odometry} \end{figure} \textbf{Visual and LiDAR Odometry.} Local odometry provides short-term motion estimation by integrating sequential sensor observations.\ Visual-Inertial Odometry (VIO) combines visual feature tracking with inertial measurements to estimate motion in real time. Early systems were mainly based on filtering frameworks, such as MSCKF~\cite{mourikis 2007 multi}, OKVIS~\cite{leutenegger 2015 keyframe}, and ROVIO~\cite{bloesch 2015 robust}. These methods integrate IMU measurements with visual observations through probabilistic filtering to provide efficient state estimation. Subsequent approaches adopt nonlinear optimization strategies to improve estimation accuracy. For example, VINS-Mono~\cite{qin 2018 vins} performs tightly coupled visual–inertial optimization over sliding windows, while ORB-SLAM

3~\cite{campos 2021 orb} extends the classical SLAM architecture with multi-map atlas representations to improve robustness and long-term consistency. Although VIO provides lightweight and accurate motion estimation, visual sensing remains sensitive to illumination changes, motion blur, and texture-less environments. To address these limitations, LiDAR-Inertial Odometry (LIO) integrates geometric measurements from LiDAR sensors with inertial observations. Early work such as LOAM~\cite{zhang 2014 loam} extracts edge and planar features from point clouds to estimate motion, while LeGO-LOAM~\cite{shan 2018 lego} improves efficiency by exploiting ground constraints for outdoor environments. More recent systems adopt factor-graph optimization frameworks. For example, LIO-SAM~\cite{shan 2020 lio} formulates LiDAR–IMU fusion as a smoothing-based optimization problem, enabling robust and globally consistent state estimation. For resource-constrained platforms such as UAVs, efficient implementations including FAST-LIO 2~\cite{xu 2022 fast} and Point-LIO~\cite{he 2023 point} further reduce computational overhead through incremental data structures and point-wise updates. Recent research increasingly focuses on LiDAR–Visual–Inertial (L-V-I) fusion, which combines complementary sensing modalities to improve robustness under diverse environmental conditions. Systems such as LVI-SAM~\cite{shan 2021 lvi}, R 3 LIVE~\cite{lin 2022 r}, and Fast-LIVO~\cite{zheng 2024 fast} tightly couple visual features, LiDAR geometry, and inertial measurements within unified optimization frameworks. By integrating both appearance and geometric information, these multi-modal approaches achieve more reliable state estimation in perceptually challenging or geometrically degenerate environments. \textbf{Global Cross-view Geo-localization.} While local odometry provides short-term motion estimation, accumulated drift inevitably leads to long-term positioning errors. To mitigate this issue, UAV systems often incorporate global geo-localization techniques that align onboard observations with external geographic references. Early approaches relied on handcrafted feature matching between aerial imagery and satellite maps. With the emergence of large-scale datasets such as University-1652~\cite{zheng 2020 university}, recent methods increasingly adopt deep learning–based cross-view representation learning. For example, SAFA~\cite{shi 2019 spatial} introduces spatial-aware feature aggregation to improve cross-view matching, while TransGeo~\cite{zhu 2022 transgeo} employs Transformer-based architectures to model global spatial relationships between UAV and satellite images. To further address the large viewpoint and scale variations inherent in aerial imagery, several studies introduce multi-scale and geometric priors. SUES-200~\cite{zhu 2023 sues} proposes a large-scale benchmark for UAV geo-localization and encourages multi-scale feature learning, while ATRPF~\cite{liao 2025 uav} incorporates ring-partitioning strategies to improve robustness under large altitude variations. Additional approaches explore geographic priors, such as heightmap gradient matching~\cite{werner 2025 kilometer} and wide-area satellite stitching~\cite{downes 2023 wide}, enabling kilometer-scale localization even in the absence of GNSS signals. \subsubsection{Mapping for Localization: Continuous and Semantic Paradigms} % \begin{figure} % \centering % \includegraphics[width=1\linewidth]{mapping for localization. Png} % \caption{Two main paradigms of mapping for localization. Dense Mapping utilizes

photometric loss to provide robust camera tracking against motion blur. Conversely, Semantic Mapping leverages high-level landmarks and concepts to solve perceptual aliasing and enable reliable loop closure. Together, they form a stable foundation for autonomous localization in complex environments.} % \label{fig: placeholder} % \end{figure} In addition to trajectory estimation, UAV systems require spatial maps that provide stable references for localization. Traditional SLAM systems typically construct sparse geometric maps based on feature landmarks. However, such representations often lack sufficient appearance information and may become unreliable in dynamic or repetitive environments. Recent research therefore explores richer mapping paradigms that improve localization robustness. \textbf{Neural and Gaussian Dense Mapping.} Following the emergence of neural implicit representations, several studies incorporate neural scene modeling into SLAM systems to enable dense mapping and photometric tracking. Early frameworks such as iMAP~\cite{sucar 2021 imap} and NICE-SLAM~\cite{zhu 2022 nice} adopt NeRF-based representations to jointly optimize camera poses and neural scene parameters. These methods leverage dense photometric consistency rather than sparse feature correspondences, improving tracking robustness in challenging visual conditions. More recently, the introduction of 3 D Gaussian Splatting (3 DGS) has enabled more efficient dense mapping. Systems such as SplaTAM~\cite{keetha 2024 splatam}, Photo-SLAM~\cite{huang 2024 photo}, and GS-SLAM~\cite{yan 2024 gs} represent scenes using Gaussian primitives and perform pose optimization through differentiable rasterization. Compared with neural volume rendering, Gaussian-based representations provide faster rendering and real-time tracking capabilities. For UAV-specific scenarios, Outdoor Monocular SLAM~\cite{cheng 2025 outdoor} addresses scale drift in monocular reconstruction, while EC 3 R-SLAM~\cite{hu 2025 ec 3 r} introduces submap-based reconstruction to maintain consistency in large-scale aerial environments. \textbf{Open-vocabulary Semantic Mapping.} While dense geometric maps improve photometric tracking, they often struggle with perceptual aliasing in environments containing repeated structures. To address this limitation, recent work incorporates semantic information into mapping systems. By leveraging Vision-Language Models (VLMs), open-vocabulary semantic mapping frameworks associate spatial locations with high-level semantic concepts. Systems such as ConceptFusion~\cite{jatavallabhula 2023 conceptfusion}, VLMaps, and FindAnything~\cite{laina 2025 findanything} integrate language-aligned visual features into spatial maps, enabling semantic reasoning and concept-based localization. In addition, uncertainty-aware approaches such as VI-SLAM~\cite{jung 2025 uncertainty} improve robustness in dynamic environments by identifying and filtering transient objects. For large-scale outdoor deployments, scalable mapping frameworks~\cite{laina 2024 scalable} and efficient submap exploration strategies~\cite{papatheodorou 2025 efficient} further improve long-term consistency and operational scalability.