CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Boosting Embodied Visual Localisation Through Multi-Granular Semantics and Multi-Robot Consensus

Wenshuai Wang[1,2] | Shiyu Liu[1] | Peifeng Jiang[1] | Hong Liu[1] | Runwei Ding[2]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, China | [2]Pengcheng Laboratory, Shenzhen, China

Correspondence: Hong Liu (hongliu@pku.edu.cn) | Runwei Ding (dingrw@pcl.ac.cn)

Received: 10 February 2026 | **Revised:** 12 April 2026 | **Accepted:** 27 April 2026

Keywords: image retrieval | multi-robot system | pose estimation | semantic consistency | visual localisation

ABSTRACT

Achieving high accuracy and synergy remains extremely difficult for multi-robot embodied visual localisation, which suffers from persistent real-world challenges such as viewpoint ambiguity, appearance variation and dynamic occlusion. Conventional optimisation-based methods often lead to incorrect feature matching without domain adaptation, whereas existing deep learning models are limited in both precision and generalisability for pose regression. Although semantic information can offer a cognitive-level scene representation, its potential has not been fully exploited due to inconsistent labelling and scaling. To address these issues, this paper proposes a novel semantic and synergic consensus framework (termed SSC-MR) that enhances image retrieval and pose estimation by leveraging multi-granular semantics and multi-robot consensus. Specifically, the spatial topology is attentively generated as a global scene graph for fast image retrieval across heterogeneous robots in a holistic workspace, whereas a local semantic-guided pose estimation strategy refines matched keypoints under semantic consistency constraints. To further improve the efficiency of multi-robot collaboration, an intra-robot minimum redundancy consensus module selects informative keyframes and compresses feature representations to reduce perceptual redundancy, and an inter-robot maximum correlation consensus module further establishes high-confidence state associations among interacting robots for knowledge sharing. Extensive experiments on five public datasets demonstrate that the proposed SSC-MR obtains superior effectiveness and robustness across diverse indoor and outdoor scenes, providing a reliable and scalable solution for embodied visual localisation in practical robotic applications. Notably, it achieves a 62.1% improvement in localisation accuracy with nearly 22× faster speed compared to a typical distributed architecture for 10 collaborative robots.

1 | Introduction

Visual localisation aims to estimate the 6-degree-of-freedom (6DoF) pose of an agent relative to pre-built 3D maps or a database of reference images, which is a fundamental capability for emerging embodied intelligence on service robots and autonomous vehicles. Extending this capability to a team of robots enables multi-robot visual localisation, where multiple agents jointly estimate their poses within a shared coordinate system through the fusion of inter-robot observations [1]. This task is essential for missions that demand collective spatial awareness, redundancy and extended coverage beyond the limits of a single platform. Representative downstream applications include large-scale environmental monitoring, disaster response, warehouse logistics and coordinated search-and-rescue operations [2].

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

Conventional monocular methods for visual localisation are broadly categorised into three main paradigms: absolute pose estimation (APE), scene coordinate regression (SCR) and relative pose estimation (RPE) [3]. APE relies on local keypoints extraction, 2D-3D correspondence matching, and epipolar geometry solving for extrinsic matrices, but often lacks robustness under significant appearance changes and can be computationally intensive. SCR obtains camera poses by indirectly regressing the panoramic scene coordinates for imagery pixels, yet its accuracy heavily depends on the quality and prior information of the scene representation. RPE typically decomposes the problem into similar image retrieval and pairwise pose estimation, but its accuracy is constrained by retrieval quality and remains sensitive to environmental changes. Recent advances in deep learning have fostered data-driven pose regression and enhanced modules like feature learning and matching [4]. However, their performance is highly dependent on visual similarity between query and database images regarding viewpoint, illumination and geometry [5], which often breaks down in practical inference due to varying camera viewpoints, scene appearances and robot platforms.

As a high-level abstraction of scene content, semantic information is considered a promising avenue to overcome low-level photometric and geometric variations during visual localisation. Figure 1 illustrates a multi-granular semantic parsing scheme for indoor and outdoor scenes, encompassing the following: (1) place-level identifies the macro environment or location of the entire image; (2) object-level detects independent, complete objects in the scene; (3) region-level focuses on representative components or surfaces of objects; and (4) pixel-level extracts the most fine-grained keypoints. This structured semantic understanding provides a reliable foundation for establishing correspondences across robots, even under significant spatio-temporal variations. Semantic consistency provides strong constraints by ensuring stable structural elements are interpreted identically across different viewpoints, making it a critical component for building a localisation system that is accurate, efficient and robust to environmental changes. However, semantic labels are highly dependent on subjective human cognition rather than measurable parameters, so the ambiguity is widespread in large-scale, long-term and multi-robot visual localisation.

A multi-robot system generally constitutes connectivity and optimisation problems on a dynamic multiview perception network [6], where each robot provides a distinct and often complementary viewpoint of the shared workspace. Therefore, when dealing with embodied interactions between multiple robots, additional intertwined challenges beyond the primitive capabilities of single agent always leads to stringent performance degradation on visual localisation, such as sensor heterogeneity, team dynamism, bandwidth threshold and communication latency [7]. In a multi-robot system, these influential factors are intrinsic with robot members joining or leaving due to task reassignment, energy constraints or network disruptions. The frequent use of heterogeneous hardware may even cause inconsistent perceptual feedback and algorithmic compatibility failure, whereas rigorous scalability urges efficient message passing under limited bandwidth threshold and communication latency, avoiding the quadratic growth in overhead typical of naive all-to-all data exchange as the team expands. Furthermore, severe perceptual aliasing arises from viewpoint changes, appearance variations (e.g., illumination, weather, season) and dynamic occlusions, which can mislead conventional feature matching and degrade global pose refinement. Consequently, an elastic solution that can adaptively handle team dynamics, prioritise informative data under communication constraints, and maintain robust alignment amidst perceptual ambiguities is paramount to real-world deployment of multi-robot collaboration.

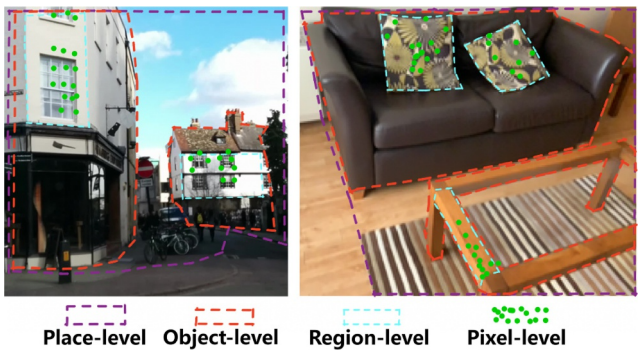


FIGURE 1 | The multi-granular semantic scheme ranges from macro to micro levels: places, objects, regions and pixels.

To address these issues, a principled multi-robot system must balance two complementary objectives: maximising perception efficiency within each robot and ensuring estimation robustness across the team. This motivates the design of a dual consensus strategy [8] to realise synergistic collaboration between individuals and swarms. Internally, each robot must achieve minimum redundancy of the reference database to distil its continuous visual stream into the most informative sparse representations, optimising terminal processing and communication efficiency. Externally, the team must achieve maximum correlation of interactive messages by establishing high-confidence correspondences based on the most reliable and distinctive observations shared among robots, enabling accurate collaborative pose estimation despite bandwidth limits and perceptual noise.

In this paper, a novel Semantic and Synergic Consensus framework is proposed for Multi-Robot embodied visual localisation as SSC-MR, which strives to cohesively learn global and local semantic consistency constraints as well as intra-robot and inter-robot consensus mechanisms. Specifically, it comprises four core modules to achieve reliable and scalable visual localisation in diverse and dynamic multi-robot scenarios. First, the Global Topo-Semantic Image Retrieval (GTSIR) module is to extract place-level and object-level semantics to generate a topological scene graph attentively and then retrieve candidate reference images efficiently. Second, the Local Sync-Semantic Pose Estimation (LSSPE) module leverages region-level and pixel-level semantic cues to screen and prioritise scale-invariant feature transform (SIFT) keypoints [9], significantly improving the accuracy and robustness of pose estimation. Third, the Intra-Robot Minimum Redundancy Consensus (iRmRC) module selects informative keyframes and compresses feature representations to reduce overall environmental perception redundancy and optimise individual perception efficiency. Finally, the Inter-Robot

Maximum Correlation Consensus (IRMCC) module facilitates efficient collaboration across multiple robots by building high-confidence feature associations and performing incremental pose updating, which further enhances system-wide localisation reliability under different bandwidth conditions. These components are integrated into a cascaded pipeline that seamlessly couples image retrieval with pose estimation, enabling efficient capture of scene semantics across multiple levels of granularity. Meanwhile, it also incorporates multi-robot coordination strategies to enhance robustness against occlusions and viewpoint variations, along with a hierarchical consensus mechanism to further accelerate large-scale retrieval.

Extensive experiments conducted on five public benchmark datasets(i.e., 7Scenes [10], Cambridge Landmarks [11], InLoc [12], Aachen Day-Night v1.1 [13] and Extended CMU Seasons [13]) demonstrate that SSC-MR not only achieves superior indoor and outdoor performance within RPE paradigm but also outperforms other representative APE and SCR methods. Specifically, quantitative results and ablation counterparts clearly confirm that the semantic-guided matching strategy significantly improves accuracy and robustness compared to previous baselines, validating the effectiveness of leveraging multi-level semantic information for visual localisation. Moreover, it's also proven that the proposed framework obtains exceptional scalability and collaborative efficiency in multi-robot settings and maintains state-of-the-art (SoTA) performance under dynamic team changes and communication constraints.

The main contributions of this work are as follows:

- A multi-granular semantic representation is introduced to unify global image retrieval with local pose estimation. Place-level and object-level topology are captured by global semantic graph for hierarchical image retrieval, whereas region-level and pixel-level cues filter local keypoints to enhance the geometric robustness of pose estimation with open-vocabulary semantic alignment.
- Two complementary consensus mechanisms are designed for multi-robot collaboration. Intra-robot minimum redundancy consensus (iRmRC) optimises the high-confidence perception efficiency through adaptive keyframe selection and feature code compression. Inter-robot maximum correlation consensus (IRMCC) establishes the distributed semantic synchronisation for collaborative pose smoothing under discrepant bandwidth constraints.
- Comprehensive evaluations on five public datasets demonstrate superior performance in localisation accuracy, retrieval quality and system efficiency. The framework exhibits high scalability and elasticity, supporting teams of 2–30 robots with graceful degradation and fast recovery from dynamic team changes.

2 | Related Work

2.1 | Monocular Visual Localisation

2.1.1 | *Absolute Pose Estimation*

APE is regarded as a classical structure-driven paradigm that relies on explicit 2D–3D keypoint correspondences or implicit neural map representations. The former reduces the task to a Perspective-n-Point (PnP) problem, which can be solved properly using the Random Sample Consensus (RANSAC) algorithm [14]. The latter, pioneered by PoseNet [11], enables a simple learning-based pattern to regress pose parameters in an end-to-end manner. Although effective in constrained conditions, these methods often struggle with perspective ambiguity or environmental changes due to their reliance on low-level feature matching. Chen et al. [15] incorporated a differentiable neural rendering module to PoseNet for direct image matching with photometric refinement and semisupervised training. Sarlin et al. [16] performed pixel-level feature alignment with the geometric precision of coordinate regression. Polizzi et al. [17] leveraged a globally sparse yet locally dense voxel representation to render view-adaptive 2D feature descriptors for lightweight relocalisation.

2.1.2 | *Scene Coordinate Regression*

SCR approaches are supposed to predict per-pixel 3D world coordinates as an intermediate geometric representation, circumventing the need for epipolar geometry constraints. Brachmann et al. [18–20] introduced a differentiable PnP–RANSAC solver (DSAC) and later enhanced its training stability with a probabilistic formulation. Huang et al. [21] utilised a regional voting scheme to learn compact scene-specific landmarks of a 3D scene. Meanwhile, Li et al. [22, 23] proposed a hierarchical architecture to capture multiscale details of feature-wise linear modulation under coarse-to-fine conditioning. Nie et al. [24] introduced a semantic–structural partition learning scheme, which executes semantic-guided feature decomposition and ambiguous point suppression to reduce intrasubscene variances.

2.1.3 | *Relative Pose Estimation*

RPE generally follows a two-stage process for map-free visual localisation: (i) retrieving one or more visually similar images with known poses from a reference database, and (ii) computing the relative transformation between the query and the retrieved image(s) to derive the final camera pose. Sarlin et al. [25] proposed HF-Net, a unified hierarchical pipeline that concurrently predicts global descriptors for image retrieval and local features for pose estimation. Xue et al. [26] employed graph neural networks (GNNs) to aggregate neighbourhood information from multiple retrieved candidates for more robust pose estimation. Zhou et al. [27] presented NC-EssNet to directly estimate the essential matrix relating image pairs, explicitly focussing on the relative geometry. Khatib et al. [28] leveraged a pretrained attention-based image matching network to extract semidense features for end-to-end relative pose regression, decoupling the bipartite loss scale to achieve superior performance and generalisation. Giang et al. [29] designed a point inference network, DeViLoc, to regress more structural semidense correspondences particularly in low-texture or ambiguous regions. More recently, Jaenal et al. [30] trained a compact multilayer perceptron (MLP) to map heterogeneous patch-level local features into a shared embedding space, enabling interoperability between different visual localisation and mapping systems.

2.2 | Semantic Visual Localisation

The integration of semantics offers a promising pathway to enhance the robustness of visual localisation. Kobyshev et al. [31]

augmented traditional local feature descriptors with semantic labels, promoting matches that are consistent in both geometry and semantic category. Toft et al. [32] proposed SemRank that reweights putative 2D–3D correspondences based on the semantic consistency of the surrounding image regions, effectively filtering mismatches. Wang et al. [33] unified object detection, semantic segmentation and spatial reasoning within a holistic 3D embodied perception suites. Scene graph is adopted and extended to represent the object-level topological relationships of a real scene [34]. Evolving from planar associations in 2D images, 3D scene graphs are further developed to comprehensively reflect the topological, semantic and geometric structure of physical environments [35–37]. Most pertinent to visual localisation is the line of work focussing on constructing such representations directly from images. Zhang et al. [38] proposed the Multiview Scene Graph (MSG) to build a semantic–topological map from an unordered image collection without relying on precise 3D reconstruction. It is especially advantageous under challenging conditions where precise geometry is difficult to obtain, as the consistency of semantic relations provides a powerful prior for cross-view matching and pose estimation.

2.3 | Multi-Robot Visual Localisation

Multi-robot visual localisation is typically organised around centralised and distributed paradigms [1]. Centralised systems collect all sensory data at a single server to provide globally consistent state estimates, although at the cost of high communication overhead and vulnerability to single-point failures. In contrast, distributed systems enable agents to perform local computations and exchange information only with immediate neighbours, often employing distributed pose graph optimisation or consensus algorithms to achieve reliable and scalable collaboration. Edge-assisted multi-robot SLAM [39] offers an alternative paradigm where a central server aggregates partial observations to reduce communication overhead. Consensus mechanisms serve as an essential factor to the distributed optimisation problem, which supports collaborative decision-making among multiple robots under perceptual heterogeneity and communication constraints [40]. Recent advances introduce learning-based strategies to model inter-robot dependencies and prioritise high-confidence observations. Zhou et al. [41] employed GNNs to model inter-robot dependencies and learn more efficient consensus protocols, whereas Wang et al. [42] formulated consensus as a matching problem over semantic or geometric graphs to enhance robustness against outliers and dynamic changes. Nevertheless, existing methods usually assume homogeneous robots or static team compositions, lacking the elasticity to handle real-world dynamism and heterogeneous perceptions [43]. The proposed SSC-MR addresses these gaps by integrating semantic consistency constraints with duplex consensus mechanisms for a scalable multi-robot system.

3 | Proposed Method

3.1 | Overview

As illustrated in Figure 2, the proposed SSC-MR is designed for multi-robot embodied visual localiation by elastically integrating multi-granular semantic information across four cohesive modules, which further ensures that semantic cues are consistently leveraged from coarse to fine granularities, enhancing both accuracy and robustness of visual localisation even in a long-term, large-scale and multiagent system.

3.2 | Global Topo-Semantic Image Retrieval

Given a scene image $\boldsymbol{I} \in \mathbb{R}^{h\times w\times 3}$, the feature extraction phase first employs DINOv2 [44] as the visual encoder $f_{enc}(\cdot)$ to transform it into an intermediate feature map $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ with $\boldsymbol{x} = f_{enc}(\boldsymbol{I})$.

Then the semantic encoding phase projects $\boldsymbol{x}$ to the coarse-grained semantic space, which yield initial place and object vectors as $\boldsymbol{v}_p = \phi_p(\boldsymbol{x})$ and $\boldsymbol{v}_o = \phi_o(\boldsymbol{x})$, respectively.

The key relational inference phase employs a Transformer decoder $f_{dec}(\cdot)$ to deeply fuse contextual information and semantic embeddings and produce an Attention-driven Semantic Scene Graph (AdSSG) [38], which is maintained as:

$$\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}) = \left(\boldsymbol{v}_p, \boldsymbol{v}_o,\ f_{dec}\left(\boldsymbol{x}, \boldsymbol{v}_p, \boldsymbol{v}_o\right)\right), \tag{1}$$

where $\boldsymbol{V}$ represents the embedding nodes and $\boldsymbol{E}$ denotes their edges recorded by an adjacency matrix.

On the offline stage, each reference image $\boldsymbol{I}_{r_i}$ is preprocessed to produce the primary, fine-grained feature database as:

$$\boldsymbol{F}_r = \left\{\boldsymbol{v}_{r_1}, \boldsymbol{v}_{r_2}, ..., \boldsymbol{v}_{r_m}\right\}, \tag{2}$$

where $\boldsymbol{v}_{r_i} = \left[\boldsymbol{v}_{p_{r_i}}, \boldsymbol{v}_{o_{r_i}}\right]$ is the concatenated embedding of $\boldsymbol{I}_{r_i}$ and m is the total number of reference images.

To accelerate the search process, a coarse-grained keyframe subset $\boldsymbol{F}_c \subset \boldsymbol{F}_r$ is constructed by traversing $\boldsymbol{F}_r$ and retaining $\boldsymbol{v}_{r_i}$ as a new keyframe only if its cosine similarity with the previously selected keyframe $\boldsymbol{v}_{c_last}$ satisfies:

$$\frac{\boldsymbol{v}_{r_i} \cdot \boldsymbol{v}_{c_last}}{\|\boldsymbol{v}_{r_i}\| \, \|\boldsymbol{v}_{c_last}\|} < \tau_{\text{frame}}, \tag{3}$$

where τ_{frame} is a predefined similarity threshold for keyframe selection. This ensures that only sufficiently distinct keyframes are preserved to reduce index redundancy.

During online inference, a hierarchical multi-robot search strategy is used for efficient distributed collaboration. The composite query consists of N simultaneous views $\boldsymbol{V}_q = \left\{\boldsymbol{I}_q^1, ..., \boldsymbol{I}_q^N\right\}$, with one designated as the main query and the remaining $N-1$ as auxiliary. Coarse-to-fine retrieval first matches the main query's embedding against the keyframe subset $\boldsymbol{F}_c$ to obtain candidate windows $\boldsymbol{C}_w \subset \boldsymbol{F}_r$; it then performs fine-grained search within $\boldsymbol{C}_w$. For each candidate embedding $\boldsymbol{v}_k \in \boldsymbol{C}_w$, a fused confidence score is computed as the average cosine similarity over all query views:

$$S_{\text{fused}}(\boldsymbol{v}_k) = \frac{1}{N}\sum_{j=1}^{N} \frac{\boldsymbol{v}_k \cdot \boldsymbol{v}_q^j}{\|\boldsymbol{v}_k\| \, \|\boldsymbol{v}_q^j\|}, \tag{4}$$

where each $\boldsymbol{v}_q^j$ is the embedding of query view $\boldsymbol{I}_q^j$.

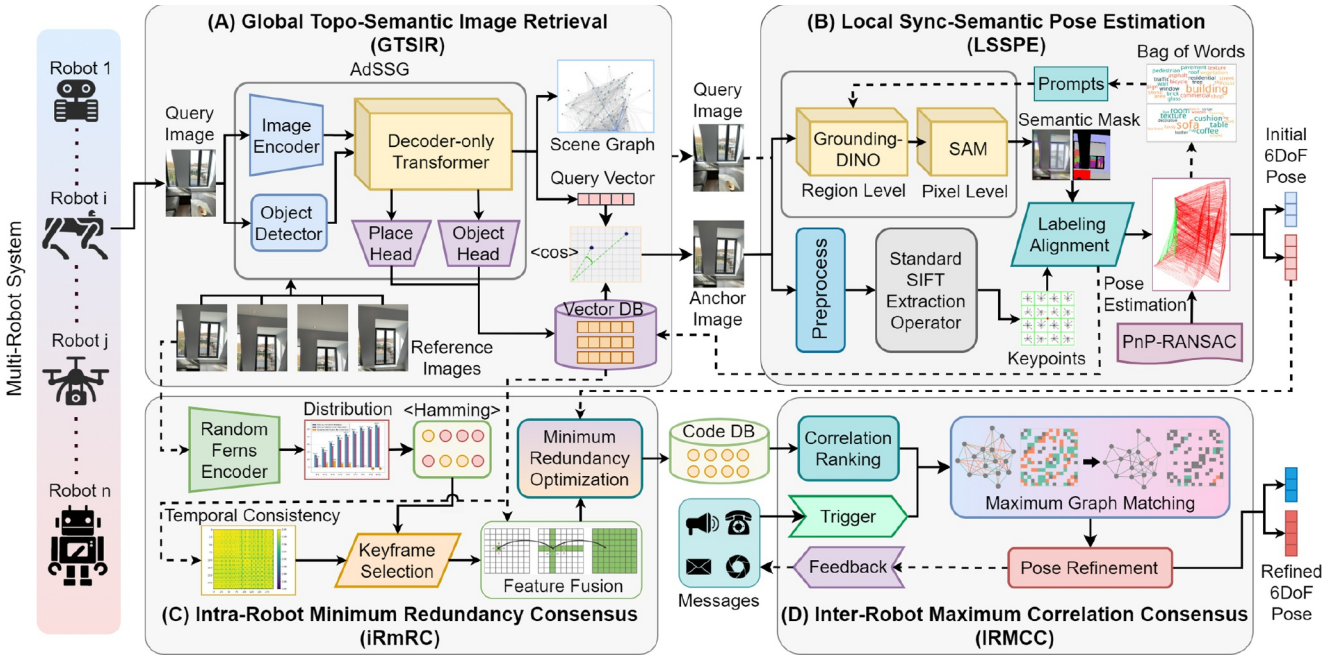


FIGURE 2 | Overview of the proposed SSC-MR framework, comprising four core modules: the Global Topo-Semantic Image Retrieval (GTSIR) module, the Local Sync-Semantic Pose Estimation (LSSPE) module, the Intra-Robot Minimum Redundancy Consensus (iRmRC) module, and the Inter-Robot Maximum Correlation Consensus (IRMCC) module.

The optimal index k^* is determined by finding the candidate that maximises the fused score:

$$k^* = \underset{k\text{s.t.}\boldsymbol{v}_k \in \boldsymbol{C}_w}{\operatorname{argmax}} S_{\text{fused}}(\boldsymbol{v}_k), \tag{5}$$

and the anchor image is then selected as $\boldsymbol{I}_{r_{k^*}}$.

3.3 | Local Sync-Semantic Pose Estimation

Based on the open-vocabulary textual prompts set $\boldsymbol{T}$, the Grounding-DINO [45] model $f_{gd}(\cdot,\cdot)$ is chosen to identify the salient regions in each scene image $\boldsymbol{I}$ as:

$$\boldsymbol{B}_n = f_{\text{gd}}(\boldsymbol{I}, \boldsymbol{T}, \tau_{\text{box}}, \tau_{\text{text}}), \tag{6}$$

where n is the number of bounding boxes. τ_{box} and τ_{text} are the confidence thresholds for bounding box and text predictions, respectively.

In particular, an automatic selection mechanism is added for $\boldsymbol{T}$ to improve the geometric–semantic consistency of candidate elements.

Subsequently, $\boldsymbol{B}_n$ is aligned to $\boldsymbol{I}$ by using the SAM [46] model $f_{\text{sam}}(\cdot,\cdot)$, thus a pixel-wise binary semantic mask $\boldsymbol{M} \in \mathbb{R}^{h\times w\times 3}$ can be obtained after morphological dilation [47] $f_{mor}(\cdot)$ and channel aggregation $\cup$ as:

$$\boldsymbol{M} = \bigcup_{j=1}^{n} f_{\text{mor}}\left(f_{\text{sam}}\left(\boldsymbol{I}, \boldsymbol{B}_{n_j}\right)\right) = \{x, y, b, \boldsymbol{l}\}, \tag{7}$$

where (x, y) is the pixel coordinate, whereas $b \in \{0, 1\}$ denotes the binary mask and $\boldsymbol{l}$ denotes the semantic label.

Besides, the standard keypoints $\boldsymbol{P}$ of $\boldsymbol{I}$ can be obtained by the original SIFT algorithm $f_{\text{sift}}(\cdot)$ as:

$$\boldsymbol{P} = f_{\text{sift}}(\boldsymbol{I}) = \left\{x_p, y_p, \boldsymbol{d}\right\}, \tag{8}$$

where $\left(x_p, y_p\right)$ is the keypoint coordinate and $\boldsymbol{d}$ denotes the feature descriptor.

Labelling alignment is further performed between $\boldsymbol{P}$ and $\boldsymbol{M}$ to get the semantic SIFT keypoints $\boldsymbol{P}'$ as:

$$\boldsymbol{P}' = \boldsymbol{P} \cap \boldsymbol{M} = \left\{x_p, y_p, b, \boldsymbol{l}, \boldsymbol{d}\right\}. \tag{9}$$

For a matched image pair $\{\boldsymbol{I}_q, \boldsymbol{I}_{r_k}\}$, a semantically guided matching strategy is devised to dynamically adjust the Euclidean distance between their semantic keypoints $\left(\boldsymbol{P}'_q,\ \boldsymbol{P}'_{r_k}\right)$. Using the threshold τ_{lowe} of Lowe's ratio test [48], the adjusted matching criterion is:

$$\|\boldsymbol{d}_q - \boldsymbol{d}_{r_k}\|_2 \leq \begin{cases} \dfrac{1+\lambda}{1-\lambda}\tau_{\text{lowe}}, & \text{if } b_q = 1 \wedge b_{r_k} = 1, \\ \dfrac{1+\mu}{1-\mu}\tau_{\text{lowe}}, & \text{if } b_q = 1 \wedge b_{r_k} = 0, \\ \tau_{\text{lowe}}, & \text{otherwise}, \end{cases} \tag{10}$$

where λ and μ are scaling factors.

On this basis, this strategy relaxes matching criteria to form a high-quality 2D–2D feature correspondence subset for relative 6DoF pose estimation:

$$\boldsymbol{C} = \left\{\left(\boldsymbol{P}'_q,\ \boldsymbol{P}'_{r_k}\right) \mid \boldsymbol{l}_q = \boldsymbol{l}_{r_k}\right\}. \tag{11}$$

Furthermore, the essential matrix $\boldsymbol{\Omega}$ of the PnP-RANSAC algorithm is robustly estimated under pairwise epipolar constraints, that is, $\boldsymbol{P}_{r_k}'^{T}\boldsymbol{\Omega}\boldsymbol{P}_q' = 0$. As a result, the relative rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and the normalised translation vector $\boldsymbol{t} \in \mathbb{R}^{1\times3}$ are decomposed from $\boldsymbol{\Omega} = [\boldsymbol{t}]_\times \boldsymbol{R}$, where $[\cdot]_\times$ denotes the antisymmetric matrix.

3.4 | Intra-Robot Minimum Redundancy Consensus

Each robot employs a random fern [49] classifier $\mathcal{F}$ for autonomous keyframe selection. For a frame sequence $\{\boldsymbol{I}_t\}_{t=1}^{T}$, each frame is encoded into a binary code $\boldsymbol{c}_t = \mathcal{F}(\boldsymbol{I}_t) \in \{0,1\}^L$, where L is the code length. The redundancy between consecutive frames is computed as:

$$\mathcal{R}_t = \mathrm{ham}(\boldsymbol{c}_t,\ \boldsymbol{c}_{t-1}) + \beta \cdot \mathcal{D}(\boldsymbol{I}_t), \tag{12}$$

where $\mathrm{ham}(\cdot,\cdot)$ is the Hamming distance, $\mathcal{D}(\boldsymbol{I}_t)$ measures semantic divergence and β is a balancing factor.

A frame $\boldsymbol{I}_t$ is selected for further processing only if $\mathcal{R}_t > \rho$, where ρ is an adaptive threshold tuned based on available computational resources.

For each selected keyframe $\boldsymbol{I}_k$, multi-granularity semantic features are extracted:

$$\boldsymbol{F}_k = \left\{\boldsymbol{f}_k^{\text{scene}},\ \boldsymbol{f}_k^{\text{object}},\ \boldsymbol{f}_k^{\text{part}}\right\}, \tag{13}$$

corresponding to scene-level, object-level and part-level descriptors.

Given a set of such features $\mathcal{K} = \{\boldsymbol{F}_1, ..., \boldsymbol{F}_K\}$, the intra-robot redundancy minimisation is formalised as selecting a subset $\mathcal{S}\subseteq\mathcal{K}$ that minimises feature redundancy under a storage budget B_{intra}:

$$\min_{\mathcal{S}\subseteq\mathcal{K}} \sum_{\boldsymbol{F}_i,\ \boldsymbol{F}_j\in\mathcal{S}} \mathcal{R}\left(\boldsymbol{F}_i,\ \boldsymbol{F}_j\right) \quad \text{s.t.} \quad |\mathcal{S}| \le B_{\text{intra}}, \tag{14}$$

where $\mathcal{R}(\cdot,\cdot)$ quantifies the redundancy between two feature sets.

A greedy algorithm selects features with the lowest redundancy to the current cache. The resulting compact feature set is used for local pose optimisation and is ready for inter-robot exchange. $\mathcal{F}$ is fine-tuned online via contrastive learning to adaptively distinguish between redundant and informative frames. The pseudocode for iRmRC module is shown in Algorithm 1.

ALGORITHM 1 | Pseudocode for iRmRC module.

Require: Image sequence $\{I_t\}$, storage budget B_{intra}, threshold ρ, classifier $\mathcal{F}$

Ensure: Compact feature subset $\mathcal{S}$

$\mathcal{K} \leftarrow \emptyset$, $c_{\text{prev}} \leftarrow$ None
for each I_t **do**
  $c_t \leftarrow \mathcal{F}(I_t)$
  $\mathcal{R}_t \leftarrow \mathrm{Hamming}(c_t, c_{\text{prev}}) + \beta \cdot \mathcal{D}(I_t)$
  if $\mathcal{R}_t > \rho$ **then**
    $\mathbf{F}_t \leftarrow$ extract_features(I_t)
    $\mathcal{K} \leftarrow \mathcal{K} \cup \{\mathbf{F}_t\}$
  end if
  $c_{\text{prev}} \leftarrow c_t$
end for
$\mathcal{S} \leftarrow$ select_least_redundant($\mathcal{K}, B_{\text{intra}}$) ▷ Greedy selection
return $\mathcal{S}$

3.5 | Inter-Robot Maximum Correlation Consensus

Consider a team of n robots. For robots i and j, the correlation between their semantic observations $\boldsymbol{e}_a$ and $\boldsymbol{e}_b$ is defined as:

$$\mathcal{C}_{\boldsymbol{e}_a\boldsymbol{e}_b}^{ij} = \mathrm{sim}\left(\boldsymbol{f}_{\boldsymbol{e}_a}^{i},\ \boldsymbol{f}_{\boldsymbol{e}_b}^{j}\right) \cdot \mathrm{geo}\left(\boldsymbol{p}_{\boldsymbol{e}_a}^{i},\ \boldsymbol{p}_{\boldsymbol{e}_b}^{j}\right), \tag{15}$$

where $\mathrm{sim}(\cdot,\cdot)$ measures semantic feature similarity and $\mathrm{geo}(\cdot,\cdot)$ evaluates geometric consistency given the robots' pose estimates.

Therefore, a correlation graph can be constructed as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where vertices $\mathcal{V}$ are observations and edges $\mathcal{E}$ are weighted by $\mathcal{C}_{ij}$. To establish reliable inter-robot correspondences, we formulate a maximum correlation matching problem, which selects a subset of edges $\mathcal{M} \subseteq \mathcal{E}$ that maximises the total correlation while ensuring that each observation is matched at most once, defined as an integer linear programme:

$$\max_{x_{ij}\in0,1} \sum_{(\boldsymbol{v}_i,\boldsymbol{v}_j)\in\mathcal{E}} \mathcal{C}_{ij}x_{ij} \quad \text{s.t.} \quad \sum_{j:(\boldsymbol{v}_i,\boldsymbol{v}_j)\in\mathcal{E}} x_{ij} \le 1,\ \forall \boldsymbol{v}_i \in \mathcal{V}, \tag{16}$$

where the binary variable x_{ij} indicates whether edge $(\boldsymbol{v}_i, \boldsymbol{v}_j)$ is included in the matching $\mathcal{M}$ (i.e., $x_{ij} = 1$ if observations $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are matched), and the constraint enforces that each vertex is incident to at most one selected edge, thereby guaranteeing a valid matching.

Based on the matches in $\mathcal{M}$, we formulate a consensus optimisation to refine robot poses:

$$\min_{\{\boldsymbol{T}_i\}} \sum_{(\boldsymbol{v}_i,\boldsymbol{v}_j)\in\mathcal{M}} \|\boldsymbol{T}_i(\boldsymbol{p}_i) - \boldsymbol{T}_j(\boldsymbol{p}_j)\|^2, \tag{17}$$

where $\boldsymbol{T}_i$ is the pose transformation for robot i and $\boldsymbol{p}_i$ is the observed position of vertex v_i. This optimises global consistency based on high-correlation observations.

An incremental update mechanism is employed for efficiency: when a robot obtains new observations, it broadcasts compact descriptors. Receivers compute correlations, update $\mathcal{G}$ and trigger consensus optimisation when sufficient high-confidence edges $(\mathcal{C}_{ij} > \tau_{\text{corr}})$ accumulate. To enhance robustness against perceptual noise, a correlation metric is formulated as:

$$\widetilde{\mathcal{C}}_{ij} = \frac{\mathcal{C}_{ij}}{1 + \alpha \cdot \mathrm{var}(\{\mathcal{C}_{ik} : k \in \mathcal{N}_i\})}, \tag{18}$$

where α controls the sensitivity to observation variance, whereas var(·) denotes the sample variance of the correlation values between robot i and all its neighbouring observations $\mathcal{N}_i$.

Finally, the consensus results are fed back to individual robots to refine their local maps and poses, creating a synergistic loop: reduced intra-robot redundancy yields higher-quality data for inter-robot correlation, which in turn produces more accurate pose constraints to guide future intra-robot perception. The pseudocode for IRMCC module is shown in Algorithm 2.

ALGORITHM 2 | Pseudocode for IRMCC module.

Require: Pose $\boldsymbol{p}_i$, feature stream $\boldsymbol{F}_i$ for robot i
Ensure: Refined pose transformation $\boldsymbol{T}_i$
$\mathcal{G}_i \leftarrow (\mathcal{V}, \mathcal{E})$
for each $(\boldsymbol{F}_i, \boldsymbol{p}_i)$ at robot i **do**
  broadcast$(\boldsymbol{F}_i, \boldsymbol{p}_i)$
  for each received $(\boldsymbol{F}_j, \boldsymbol{p}_j)$ from robot j **do**
    $\mathcal{C}_{ij} \leftarrow \text{sim}(\boldsymbol{F}_i, \boldsymbol{F}_j) \cdot \text{geo}(\boldsymbol{p}_i, \boldsymbol{p}_j)$
    $\mathcal{G}_i \leftarrow \boldsymbol{v}_i, \boldsymbol{e}_{ij}, \mathcal{C}_{ij}$
  end for
  if high-confidence edges exist **then**
    $\mathcal{M} \leftarrow$ max_weight_matching(G_i)
    $\boldsymbol{T}_i \leftarrow \arg\min \sum_{(\boldsymbol{v}_a, \boldsymbol{v}_b) \in \mathcal{M}} \|\boldsymbol{T}_a(p_a) - \boldsymbol{T}_b(p_b)\|^2$
    broadcast$(\boldsymbol{T}_i)$
  end if
end for
return $\boldsymbol{T}_i$

4 | Experiments and Discussions

4.1 | Datasets and Evaluation Metrics

4.1.1 | *Datasets*

Five public benchmark datasets are selected for experimental evaluations: (1) 7Scenes [10] comprises seven indoor scenes captured by a handheld Kinect RGB-D camera and may contain viewpoint variations, motion blurs or repetitive textures; (2) Cambridge Landmarks [11] includes six outdoor scenes taken around the University of Cambridge with scale and occlusion challenges, five of which are commonly used except for the Street scene; (3) InLoc [12] is a challenging indoor visual localisation dataset that features large-scale scene retrieval and strong viewpoint variations; (4) Aachen Day-Night v1.1 [13] and (5) Extended CMU-Seasons [13] provide challenging outdoor scenarios with significant illumination and seasonal variations, respectively. Notably, ARKitScenes [50] is employed for the pretraining of AdSSG to learn semantic topology, leveraging its high-resolution depth maps and rich structural details as precise ground truth.

4.1.2 | *Evaluation Metrics*

The performance of visual localisation is evaluated using two complementary protocols. First, the median translation error (in metres) and rotation error (in degrees) of relative pose estimation are reported on holistic scenes, following standard practice on 7Scenes and Cambridge Landmarks. Second, for challenging benchmarks such as InLoc, Aachen Day-Night, and Extended CMU Seasons, localisation accuracy is measured by the percentage of query frames that fall within predefined thresholds. Besides, Image retrieval performance is measured by Recall@K and Precision@K.

The performance of multi-robot collaboration is further quantified by: (1) Collaborative Accuracy Gain: the relative reduction in localisation error compared to a single-robot baseline, defined as $G(n) = (1 - E_n/E_1) \times 100\%$, where E_n is the average pose error for an n–robot team; (2) Consensus Efficiency Ratio: $\eta = G(n)/C(n)$, relating accuracy gain $G(n)$ to the per-robot communication cost $C(n)$ (in KB/s); (3) Elasticity Index: measures the recovery speed and final accuracy after a 20% random subset of robots dynamically drops out and rejoins the team.

4.2 | Implementation Details

4.2.1 | *Perception Settings*

For GTSIR, 200 query images are randomly selected per dataset, and the system retrieves the top 20 candidate images to quantify their performance. Temporal sensitivity is also analysed by varying the thresholds of time intervals between query and reference images. For LSSPE, the textual prompt set $\boldsymbol{T}$ is initialised with several representative categories, for example, buildings, corners, structures and landmarks. $\tau_{\text{box}} = 0.35$ and $\tau_{\text{text}} = 0.25$ in Equation (6) are applied for Grounding-DINO to filter low-quality detections. The maximum number of SIFT keypoints is set to 1000, and Lowe's ratio threshold τ_{lowe} is set to 0.5 to ensure its capability of scene generalisation. The scaling factors in Equation (10) are set to $\lambda = 0.2$ and $\mu = 0.1$.

4.2.2 | *Interaction Settings*

For iRmRC, each robot employs a random fern classifier $\mathcal{F}$ with a binary code length $L = 256$, which is sufficient for adopted moderate-scale databases. The keyframe redundancy threshold ρ is initialised to 0.7 and adjusted adaptively based on computational resources. The balance factor β in Equation (12) is set to 0.3, and the intra-robot storage budget $\boldsymbol{B}_{intra}$ is 500 feature vectors. $\mathcal{F}$ is fine-tuned online via contrastive learning for 100 iterations with a learning rate of 1×10^{-4}. For IRMCC, the correlation threshold $\tau corr$ for high-confidence edges is set to 0.8. The robust correlation metric in Equation (18) uses $\alpha = 0.5$ to down-weight inconsistencies. Maximum correlation matching is solved as an integer linear programming (ILP) problem with Gurobi optimiser [51] under a 100 ms per-round time limit. Consensus optimisation is triggered once at least 5 high-confidence edges $(C_{ij} > \tau_{corr})$ accumulate. Geometric consistency in Equation (15) is evaluated using a Gaussian kernel with $\sigma = 1.0$.

4.2.3 | *System Platform*

All experiments are conducted on an Ubuntu 20.04 server equipped with an NVIDIA RTX 3090 GPU and configured with Python 3.8, PyTorch 1.12 and CUDA 11.3 for fair and reproducible comparisons.

4.3 | Performance Comparisons

4.3.1 | *Comparisons on Visual Localisation*

4.3.1.1 | *Indoor Performance.* As detailed in Table 1, SSC-MR attains an average localisation error of 0.018 m/1.12° on 7Scenes dataset, surpassing all RPE-based competitors including DeViLoc (0.050 m/1.68°) and CrossFeat (0.054 m/1.65°). Notably, it also outperforms the leading methods from other paradigms: it exhibits lower translation error than the best APE method FaVoR (0.019 m) and the best SCR method VS-Net (0.024 m). Per-scene analysis confirms that SSC-MR achieves the best translation accuracy in six of the seven scenes (except for Heads) with competitive rotational errors, demonstrating robust performance under diverse indoor conditions.

4.3.1.2 | *Outdoor Performance.* As listed in Table 2, SSC-MR maintains exceptional robustness on the Cambridge Landmarks dataset, achieving an average error of 0.063 m/0.16° on the four core scenes. This significantly outperforms other RPE-based methods such as DeViLoc (0.090 m/0.23°) and CrossFeat (0.173 m/0.36°). Even when including the large-scale, repetitive GreatCourt sequence, SSC-MR maintains a leading average of 0.098 m/0.14°, surpassing all RPE competitors. Furthermore, it exceeds the performance of the best APE method PixLoc (0.150 m/0.25°) and the best SCR method (0.136 m/0.24°), validating that semantic-consistent matching provides a more reliable prior than pose regression and direct matching in unstructured outdoor environments.

4.3.1.3 | *Challenging Performance.* The robustness of SSC-MR is further validated on three challenging benchmarks (i.e., InLoc, Aachen v1.1, and Extended CMU Seasons) against competitive RPE baselines. As reported in Table 3, SSC-MR consistently achieves the highest accuracy among all compared RPE methods across every subset and evaluation thresholds. For example, on InLoc DUC1 at the strictest criterion (0.25 m, 10°), SSC-MR outperforms the second-best CrossFeat by a clear margin. On Aachen v1.1 night, SSC-MR attains the top results (88.6/94.8/100%), surpassing other competitive RPE methods such as DeViLoc and CrossFeat. On the Extended CMU Seasons dataset, SSC-MR maintains superior accuracy in urban, suburban and park environments, demonstrating strong generalisation under severe appearance changes, dynamic lighting and large-scale scene retrieval conditions.

4.3.2 | *Comparisons on Multi-robot Collaboration*

The collaborative efficiency of SSC-MR is evaluated with robot teams ranging from 2 to 30 members. As summarised in Table 4, SSC-MR exhibits progressive improvement in localisation accuracy as the robot count increases from 2 to 10, achieving a maximum collaborative gain ($G(n)$) of 62.1% in translation error reduction relative to the single-robot baseline. The consensus efficiency ratio (η) remains high (above 1.27) in this regime, indicating the efficient usage of collaborative resources while maintaining real-time interactions. These results confirm that the proposed semantic and synergic consensus mechanism is highly effective for small-to-medium-sized teams (2–10 robots), delivering significant accuracy improvement while maintaining real-time feasibility.

4.4 | Visualisation Results

4.4.1 | *Visualisation of Scene Graph*

Figure 3 illustrates an example of global scene graph generated from the GTSIR module. The graph encodes an input scene by representing place-level (e.g., 'kitchen', 'corridor') and object-level entities (e.g., 'table', 'chair') as nodes, with edges capturing their semantic and spatial relationships via a decoder-only Transformer. This representation abstracts away low-level visual variations, resulting in a semantic topology that enables fast space retrieval with semantically coherent results.

TABLE 1 | Median errors (m/°) of competitive methods for visual localisation on 7Scenes dataset.

	Method	Chess	Fire	Heads	Office	Pumpkin	RedKitchen	Stairs	Average
APE	PoseNet [11]	0.320/8.12	0.470/14.4	0.290/12.0	0.480/7.68	0.470/8.42	0.590/8.64	0.470/13.8	0.441/10.4
	Direct-PN [15]	0.100/3.52	0.270/8.66	0.170/13.1	0.160/5.96	0.190/3.85	0.220/5.13	0.320/10.6	0.200/7.26
	PixLoc [16]	0.020/0.80	0.020/0.73	0.010/0.82	0.030/0.82	0.040/1.21	0.030/1.20	0.050/1.30	0.028/0.98
	FaVoR [17]	0.010/0.20	0.010/0.40	0.010/0.60	0.020/0.40	0.010/0.30	0.010/0.30	0.060/1.60	0.019/0.50
SCR	DSAC* [20]	0.018/0.60	0.019/0.90	0.012/0.70	0.026/0.80	0.041/1.10	0.039/1.30	0.043/1.30	0.028/1.00
	VS-Net [21]	0.015/0.50	0.019/0.80	0.012/0.70	0.021/0.60	0.047/1.00	0.036/1.10	0.028/0.80	0.024/0.80
	HSC-Net++ [23]	0.020/0.70	0.020/0.72	0.010/0.80	0.020/0.69	0.040/1.00	0.040/1.15	0.030/1.02	0.026/0.87
RPE	HF-Net [25]	0.026/0.90	0.027/1.00	0.014/0.90	0.043/1.20	0.058/1.60	0.053/1.60	0.072/1.90	0.042/1.30
	MV-GNN [26]	0.080/2.80	0.260/8.94	0.170/11.4	0.180/5.08	0.150/2.77	0.250/4.48	0.230/8.87	0.190/6.33
	NC-EssNet [27]	0.120/5.60	0.260/9.60	0.140/10.7	0.200/6.70	0.220/5.70	0.220/6.30	0.310/7.90	0.210/7.50
	IM-RPR [28]	0.091/3.84	0.136/5.09	0.091/5.77	0.112/3.79	0.139/3.66	0.127/3.62	0.265/6.05	0.137/4.55
	DeViLoc [29]	0.020/**0.78**	0.020/**0.74**	**0.010**/**0.65**	0.030/**0.82**	0.040/1.02	0.030/1.19	0.040/1.12	0.050/1.68
	CrossFeat [30]	0.029/0.99	0.030/1.15	0.019/1.27	0.047/1.32	0.072/1.94	0.055/1.65	0.123/3.22	0.054/1.65
	SSC-MR (Ours)	**0.013**/0.84	**0.018**/1.09	0.015/1.33	**0.029**/1.26	**0.015**/**0.59**	**0.014**/**0.84**	**0.020**/**1.09**	**0.018**/**1.12**

Note: Best RPE results are highlighted in bold and second best RPE results are underlined.

TABLE 2 | Median errors (m/°) of competitive methods for visual localisation on Cambridge Landmarks dataset.

	Method	Kings college	Old hospital	Shopfacade	St. marys church	Average (4)	Great court	Average (5)
APE	PoseNet [11]	1.92/5.40	2.31/5.40	1.46/8.10	2.66/8.50	2.086/6.80	–/–	–/–
	PixLoc [16]	0.14/0.24	0.16/0.32	0.05/0.23	0.10/0.34	0.113/0.28	0.30/0.14	0.150/0.25
	FaVoR [17]	0.18/0.30	0.27/0.50	0.05/0.30	0.11/0.40	0.153/0.38	0.29/0.20	0.180/0.34
SCR	DSAC* [20]	0.15/0.30	0.21/0.40	0.05/0.30	0.13/0.40	0.135/0.35	0.49/0.30	0.206/0.34
	VS-Net [21]	0.16/0.20	0.16/0.30	0.06/0.30	0.08/0.30	0.115/0.28	0.22/0.10	0.136/0.24
	HSC-Net++ [23]	0.19/0.34	0.20/0.31	0.06/0.24	0.09/0.30	0.135/0.30	0.39/0.23	0.186/0.28
	SSD [24]	0.26/0.30	0.27/0.50	0.05/0.30	0.20/0.60	0.195/0.43	0.27/0.10	0.210/0.36
RPE	HF-Net [25]	0.34/0.40	0.43/0.60	0.09/0.40	0.16/0.50	0.255/0.48	0.76/0.30	0.356/0.31
	MV-GNN [26]	0.59/0.65	0.50/2.87	1.90/3.29	1.88/2.78	1.216/2.40	6.67/2.79	2.308/2.48
	NC-EssNet [27]	0.61/1.60	0.95/2.70	0.71/3.40	1.12/3.60	0.848/2.83	–/–	–/–
	IM-RPR [28]	0.27/4.39	0.28/5.51	0.17/7.29	0.22/6.25	0.235/5.86	–/–	–/–
	DeViLoc [29]	0.12/0.21	0.10/0.18	0.04/**0.18**	0.07/0.23	0.090/0.23	**0.18**/0.11	0.108/0.20
	CrossFeat [30]	0.16/0.27	0.38/0.59	0.05/0.26	0.10/0.32	0.173/0.36	0.56/0.33	0.250/0.35
	SSC-MR (Ours)	**0.10/0.15**	**0.07/0.12**	**0.03/0.18**	**0.05/0.19**	**0.063/0.16**	0.24/**0.08**	**0.098/0.14**

Note: Best RPE results are highlighted in bold and second best RPE results are underlined. '–' means the value is unconverged.

TABLE 3 | Average accuracies (%) of competitive RPE methods on challenging visual localisation datasets.

	InLoc (indoor)		Aachen v1.1 (outdoor)		Extended CMU seasons (outdoor)		
	Acc.@(0.25/0.5/1.0; m, 10°)		Acc.@(0.25/0.5/5.0; m, 2/5/10°)		Acc.@(0.25/0.5/5.0; m, 2/5/10°)		
Method	**DUC1**	**DUC2**	**Day**	**Night**	**Urban**	**Suburban**	**Park**
HFNet [25]	49.0/68.7/80.8	53.4/77.1/82.4	80.5/87.4/94.2	68.4/77.6/88.8	89.5/94.2/97.9	76.5/82.7/92.7	57.4/64.4/80.4
DeViLoc [29]	55.8/63.8/88.9	61.0/72.5/89.4	87.4/94.8/98.2	87.8/93.9/**100.**	95.7/98.4/99.2	97.1/98.3/99.4	92.1/95.1/96.3
CrossFeat [30]	56.9/66.3/89.4	62.0/73.6/90.9	89.9/95.6/98.9	87.1/92.6/97.9	95.3/98.2/99.1	96.2/97.8/99.1	91.3/94.2/95.8
SSC-MR (ours)	**58.2/69.4/91.0**	**63.5/78.1/92.3**	**91.5/96.3/99.2**	**88.6/94.8/100.**	**96.3/98.9/99.6**	**97.8/99.1/99.7**	**93.8/96.4/97.6**

Note: results are highlighted in bold and second best results are underlined.

TABLE 4 | Visual localisation performance of SSC-MR with robot counts from 2 to 30 (on Office scene of 7Scenes dataset).

Robot count (n)	Translation error (m) ↓	Rotation error (°) ↓	Communication cost (KB/s) ↓	Consensus time (ms) ↓	Collaborative gain $G(n)$ ↑	Efficiency ratio η ↑
1 (baseline)	0.029	1.26	0.0	—	0.0%	—
2	0.021 ± 0.003	1.12 ± 0.14	**15.2 ± 2.1**	**45 ± 6**	27.6%	1.82
3	0.018 ± 0.002	1.02 ± 0.12	20.5 ± 2.8	53 ± 7	37.9%	**1.85**
4	0.016 ± 0.002	0.94 ± 0.11	25.1 ± 3.2	62 ± 8	44.8%	1.78
5	0.015 ± **0.001**	0.88 ± 0.10	29.8 ± 3.5	70 ± 9	48.3%	1.62
6	0.014 ± **0.001**	0.83 ± 0.09	34.0 ± 4.0	80 ± 10	51.7%	1.52
7	0.013 ± **0.001**	0.79 ± 0.09	38.3 ± 4.3	92 ± 11	55.2%	1.44
8	0.012 ± **0.001**	0.76 ± 0.08	42.1 ± 4.7	105 ± 13	58.6%	1.39
9	0.012 ± **0.001**	0.74 ± 0.08	45.5 ± 5.0	120 ± 15	60.2%	1.32
10	**0.011 ± 0.001**	0.73 ± 0.08	48.9 ± 5.3	138 ± 17	62.1%	1.27
15	**0.011** ± 0.001	**0.71 ± 0.07**	64.7 ± 6.5	235 ± 25	**62.5%**	0.96
20	**0.011** ± 0.002	0.72 ± 0.09	78.5 ± 7.5	380 ± 40	62.4%	0.79
30	0.012 ± 0.002	0.75 ± 0.10	105.2 ± 9.0	720 ± 70	59.8%	0.56

Note: Median errors of collaborative localisation are calculated with standard deviations, whereas per-robot communication cost, consensus time, collaborative gain and efficiency ratio are compared to the single-robot baseline. Best and second-best results are highlighted in bold and underlined, respectively.

4.4.2 | *Visualisation of Semantic Parsing*

Figure 4 demonstrates the process of fine-grained semantic parsing employed in the LSSPE module. Grounding-DINO performs open-vocabulary detection to identify semantically salient objects (e.g., 'refrigerator', 'wall') and produces coarse bounding boxes. These detections are then refined by SAM into precise, pixel-wise segmentation masks, resulting in region-level semantic segments that preserve object boundaries and structural continuity.

4.4.3 | *Instances of Image Retrieval*

Figure 5 presents the examples of the image retrieval. The top-3 retrieved reference images for each query demonstrate high semantic and structural consistency, despite significant variations in viewpoint, illumination and partial occlusions. This is evident in both indoor (e.g., consistent matching of furniture) and outdoor scenes (e.g., correct identification of building facades), which validates the capability of the AdSSG to encode scene representations that are invariant to low-level visual changes and anchored in high-level semantic topology.

4.4.4 | *Instances of Random Ferns*

Figure 6 visually presents four clustering instances formed by the random fern classifier during keyframe selection. Specifically, each sub-figure displays a set of images associated with a single fern, grouped together after being identified as highly similar in both visual appearance and semantic content within the binary code space. Reliable retrieval of such geometrically and semantically coherent candidates is a critical prerequisite for accurate downstream pose estimation.

4.5 | Qualitative Analysis

4.5.1 | *Efficiency of Semantic Retrieval*

As shown in Figure 7, SSC-MR achieves accurate retrieval across indoor (7Scenes: Stairs, Office) and outdoor (Cambridge: Shop, Court) scenes. Recall@1 exceeds 0.97 (reaching 1.00 on Shop and Court), and Recall@5 reaches 1.00. Precision@K remains above 0.9 for $K \leq 3$. Elastic retrieval time grows sublinearly (e.g., from 2.5 s at 20 k images to 5.5 s at 500 k images), whereas brute-force time grows faster (4.2–29.1 s), with elastic accuracy above 88% at 1,000 k images. These results indicate

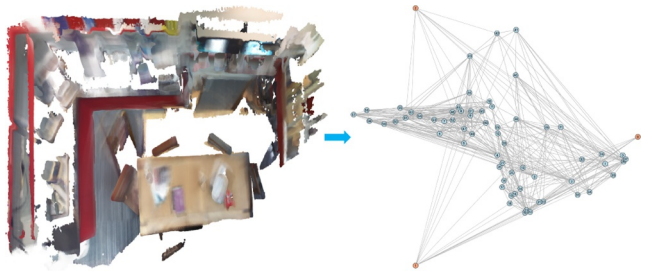


FIGURE 3 | An example of the global scene graph generated by AdSSG for a scene image. The graph nodes represent place-level and object-level semantic concepts, whereas edges capture their contextual and spatial relationships, forming a topological representation for hierarchical image retrieval.

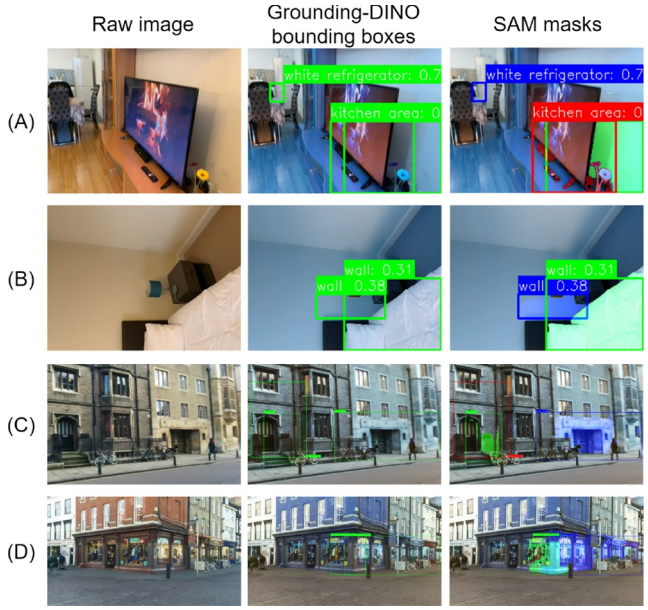


FIGURE 4 | Examples of local semantic parsing using Grounding-DINO and SAM.

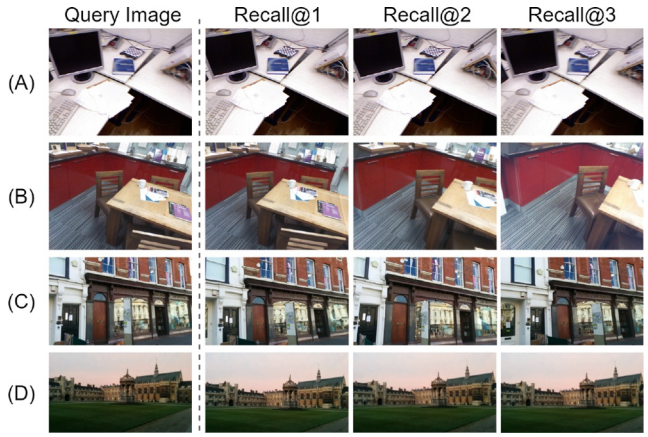


FIGURE 5 | Query instances (left) and their top-3 retrieved reference candidates (right) ranked by topological semantics across indoor (A, B) and outdoor (C, D) scenes.



FIGURE 6 | Cluster instances of random ferns from intra-robot keyframe selection. Each cluster temporally groups similar leaf frames (right) identified by the same root (left) across indoor (A, B) and outdoor (C, D) scenes.

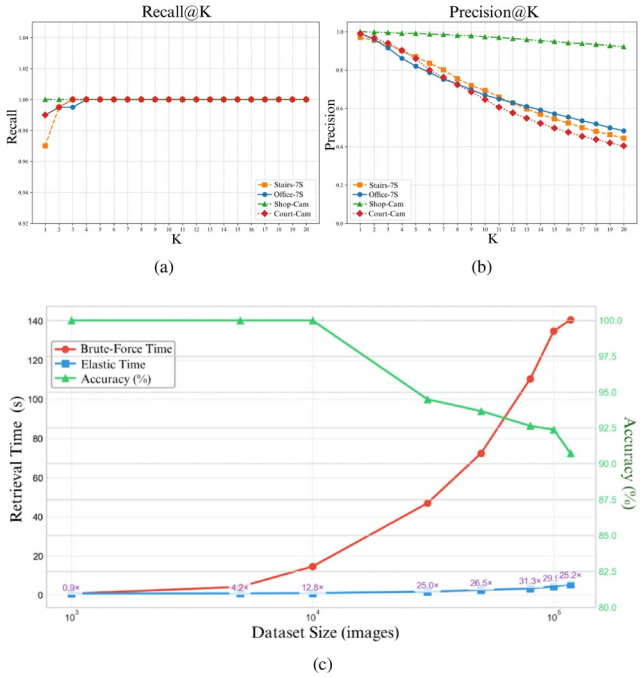


FIGURE 7 | Semantic-assisted image retrieval with the GTSIR module. (a) Recall@K and (b) Precision@K on four representative scenes. (c) Comparison of brute-force and elastic retrieval time (s) as a function of database size, with elastic accuracy (%) reported.

that SSC-MR handles short-term appearance changes effectively and scales efficiently to large databases.

4.5.2 | *Reliableness of Semantic Guidance*

Figure 8 illustrates the semantic-guided feature matching process performed by LSSPE module. In Figure 8a, SIFT keypoints are concentrated within semantically stable regions (e.g., the refrigerator door, the structured wall) identified by the segmentation masks, filtering out unstable points from ambiguous or dynamic backgrounds. In Figure 8b, matched point pairs are colour-coded by semantic consistency: green refers to bilateral semantic, orange refers to unilateral semantic and red refers to neither semantic. The predominance of green and orange connections demonstrates that the strategy effectively prioritises geometrically reliable matches anchored in semantically consistent areas. This visual evidence confirms that semantic guidance suppresses outliers and repetitive patterns, leading to more robust and accurate pose estimation.

4.5.3 | *Scalablity of Elastic Consensus*

Figure 9 compares the scalability of SSC-MR against two baseline strategies: a Distributed Pose-Graph (DPG) method and a Centralised Matching (CM) server approach. SSC-MR achieves the lowest translation error across 2–10 robots, consistently decreasing from 0.021 to 0.011 m, whereas CM errors start to increase after 5–6 robots (from 0.0165 to 0.019 m) and DPG errors rise slightly after 8 robots (from 0.015 to 0.016 m), validating the superior scalability of our approach. Its consensus time remains significantly lower than the iterative DPG method. Although the CM approach shows low latency for $N \leq 8$, its central server becomes a bottleneck for larger teams, causing a sharp rise in latency. Consequently, under strict real-time constraints (≈100 ms), our dual-consensus mechanism supports up to 10–12 robots; beyond this scale it gracefully transitions to map refinement mode (e.g., 0.012 m at 30 robots, a 59.8% improvement over a single robot), and recovers within 95% of prior accuracy after dynamic robot dropouts, demonstrating flexible, robust and efficient collaboration.

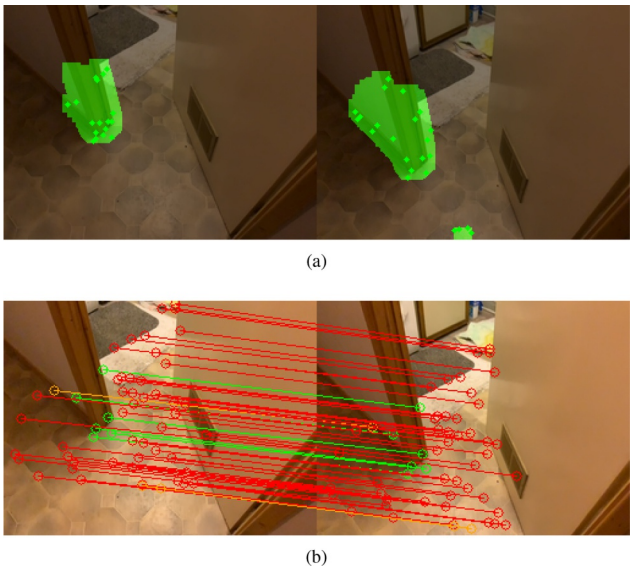


FIGURE 8 | Illustration of semantic-guided feature matching. (a) SIFT keypoints overlaid on semantic masks, showing the concentration of features within semantically stable regions. (b) Matched point pairs coloured by semantic consistency: green refers to both points lie in semantic regions, orange refers to one point is semantic and red refers to neither point is semantic.

4.6 | Ablation Study

4.6.1 | *Ablation on Semantic Components*

4.6.1.1 | *Multi-Granular Semantic Hierarchy.* To quantify the contribution of each semantic level, main components (AdSSG, G-DINO, and SAM) are incrementally removed from the full SSC-MR configuration. As shown in Table 5, removing SAM (i.e., using only object-level and place-level semantics) increases translation error by 33.9% on 7Scenes and 34.6% on Cambridge Landmarks compared to the full model, whereas rotation error rises from 1.12° to 1.28° (indoor) and from 0.16° to 0.27° (outdoor). Further removal of G-DINO (leaving only region/pixel masks) leads to additional increases of 21.7% (indoor) and 24.2% (outdoor) in translation error, with rotation errors rising to 1.41° and 0.34°, respectively. The baseline without any semantic guidance achieves the highest errors, with translation error 52.2% (indoor) and 56.8% (outdoor) higher than the full configuration and rotation errors 47.6% and 51.3% higher, respectively. These results confirm that global scene graph retrieval provides a strong topological prior that complements local geometric matching.

4.6.1.2 | *Geometric Feature Extractor Comparison.* Alternative geometric feature extractors (ORB [52], SuperPoint [53], DISK [54] and LoFTR [55]) are evaluated under the same semantic guidance pipeline on the ARKitScenes dataset, with

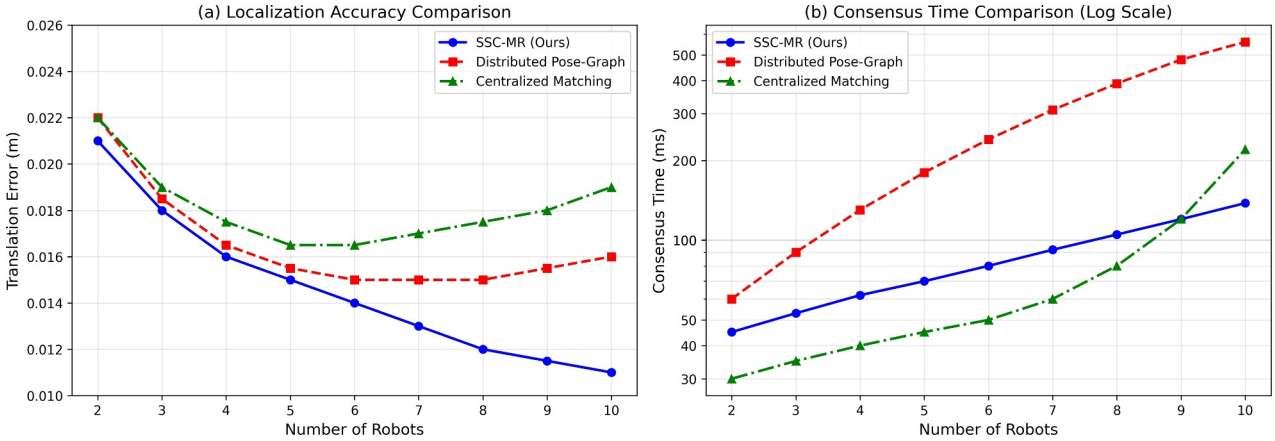


FIGURE 9 | System scalability of SSC-MR compared to centralised matching and distributed pose-graph optimisation across 2–10 robots: (a) localisation accuracy with translation errors; (b) consensus time in logarithmic form.

TABLE 5 | Ablation study on multi-granular semantic hierarchy with average error (m/°) on 7Scenes and Cambridge Landmarks datasets.

AdSSG	G-DINO	SAM	7Scenes ↓	Cambridge ↓
✓	✓	✓	**0.018/1.12**	**0.063/0.16**
	✓	✓	0.021/1.28	0.077/0.27
		✓	0.027/1.41	0.102/0.34
			0.041/1.53	0.157/0.41

Note: Best results are highlighted in bold and second best are underlined.

SIFT without semantic guidance as a baseline. Table 6 reveals that semantic guidance dramatically improves all feature types: comparing guided vs. unguided SIFT, translation error is reduced by 91.7% (0.036–0.003 m), rotation error by 72.3% (1.41°–0.39°) and inlier count is increased by 19.1%. Among guided methods, SIFT achieves the lowest translation (0.003 m) and rotation (0.39°) errors, attributed to its scale and rotation invariance. ORB offers the fastest extraction (9.3 ms) with moderate accuracy (0.008 m/0.52°). Learnt features produce more inliers (SuperPoint: 61, DISK: 65, LoFTR: 72) but at higher computational cost (28.4–58.6 ms). Consequently, SIFT is retained as the default geometric backbone, whereas ORB or SuperPoint are noted as alternatives for latency-sensitive or inlier-critical applications.

4.6.2 | *Ablations on Consensus Strategies*

4.6.2.1 | *Mechanism Contribution.* As shown in Table 7, the two consensus mechanisms (iRmRC and IRMCC) are evaluated on the Office scene (7Scenes) with a team of 5 robots. Compared to the no-consensus baseline (0.024 m, 1.32°), iRmRC alone reduces translation error by 20.8% with a 46.1% increase in communication cost (18.7 KB/s). IRMCC alone achieves a stronger reduction (33.3%) but more than doubles communication cost (29.8 KB/s). The full duplex consensus yields the best accuracy (37.5% reduction, 0.015 m/0.83°) while keeping communication cost at the IRMCC-only level (29.8 KB/s), because iRmRC compresses features before transmission. Thus, the two mechanisms are complementary: iRmRC reduces intra-robot redundancy without sacrificing the data quality needed for IRMCC to establish reliable cross-robot correspondences. Consensus latency remains 70 ms for the full configuration, which is manageable for real-time operation with 5 robots within typical control loop requirements (e.g., 10–20 Hz).

4.6.2.2 | *Scalable Behaviour.* The behaviour of the consensus strategies under varying team sizes (2–30 robots) is further assessed on the Office scene, comparing the full SSC-MR framework against configurations lacking either iRmRC or IRMCC. As shown in Table 8, for small teams (2-5 robots), iRmRC alone achieves the highest efficiency ratio η due to its low communication footprint, albeit with moderately higher rotation errors (e.g., 1.28° vs. 1.12° at 2 robots). When team size grows beyond 10 robots, the full duplex consensus maintains the lowest translation and rotation errors (e.g., 0.011 m/0.73° at 10 robots) while offering superior efficiency compared to IRMCC-only, which suffers from quadratic communication growth. The reason is that iRmRC's compression becomes increasingly beneficial as more robots share data, mitigating the communication bottleneck. These findings confirm that the proposed dual consensus strategy provides scalable performance across a wide range of team sizes, gracefully balancing accuracy and communication cost as the team expands.

4.7 | Latency and Memory

4.7.1 | *Per-Module Latency Breakdown*

The end-to-end processing pipeline of a single robot is decomposed into six functional stages, with latency measured over four datasets as detailed in Table 9. Offline preprocessing (feature database construction) takes 183.5 ms per scene. Online per-frame stages include online retrieval (24.7 ms), semantic parsing (68.3 ms), feature extraction (12.9 ms), pose estimation (9.8 ms) and consensus messaging (4.2 ms), totaling 119.9 ms. Semantic parsing dominates the online latency but remains within typical real-time constraints (10–15 Hz), primarily due to the forward pass of the large visual foundation models (Grounding-DINO and SAM). For applications demanding higher update rates (e.g., 30 Hz), the semantic parsing stage can be executed at a reduced frequency (e.g., every 3–5 frames) while maintaining full-rate geometric tracking, as semantic guidance serves as a robust prior rather than a per-frame necessity. This flexible design allows SSC-MR to adapt to diverse

TABLE 6 | Comparison of geometric feature extractors with semantic guidance.

Method	Semantic guidance	Trans. Error (m) ↓	Rotation error (°) ↓	Inliers pointers ↑	Extraction latency (ms) ↓
SIFT [9]	✗	0.036	1.41	47	<u>11.5</u>
SIFT [9]	✓	**0.003**	**0.39**	56	12.7
ORB [52]	✓	0.008	0.52	42	**9.3**
SuperPoint [53]	✓	0.005	0.41	<u>61</u>	28.4
DISK [54]	✓	<u>0.004</u>	0.40	65	35.2
LoFTR [55]	✓	**0.003**	0.38	**72**	58.6

Note: Best results are highlighted in bold and second best are underlined.

TABLE 7 | Ablation study on consensus strategies.

iRmRC	IRMCC	Trans. error (m) ↓	Rotation error (°) ↓	Comm. cost (KB/s) ↓	Consensus latency (ms) ↓
✗	✗	0.024 ± 0.004	1.32 ± 0.17	**12.8 ± 1.5**	**28 ± 4**
✓	✗	0.019 ± <u>0.002</u>	1.05 ± 0.12	<u>18.7</u> ± <u>2.1</u>	<u>52</u> ± <u>7</u>
✗	✓	<u>0.016</u> ± <u>0.002</u>	<u>0.88</u> ± <u>0.10</u>	29.8 ± 3.5	70 ± 9
✓	✓	**0.015 ± 0.001**	**0.83 ± 0.09**	29.8 ± 3.5	70 ± 9

Note: Results for a team of 5 robots on the Office scene. Best and second-best results are highlighted in bold and underlined, respectively.

TABLE 8 | Scalability analysis of consensus strategies.

Robot count	iRmRC + IRMCC			iRmRC only			IRMCC only		
	Trans.	Rot.	η	Trans.	Rot.	η	Trans.	Rot.	η
2	0.021	1.12	1.82	0.022	1.28	2.15	0.023	1.20	1.46
5	0.015	0.83	1.62	0.019	1.05	1.78	0.016	0.88	1.12
10	0.011	0.73	1.27	0.014	0.96	1.09	0.012	0.77	0.83
20	0.011	0.72	0.79	0.016	1.02	0.64	0.014	0.80	0.51

Note: Translation error (m), rotation error (°), and efficiency ratio η (accuracy gain per communication cost) for different team sizes on the Office scene of 7Scenes dataset.

real-time requirements, enhancing its practicality for real-world deployment.

4.7.2 | *Consensus Messaging Overhead*

The communication overhead of the consensus mechanisms is analysed for multi-robot operation on a simulated 100 Mbps Wi-Fi network with 5 ms baseline latency, as listed in Table 10. Messages contain compressed feature descriptors (iRmRC output) and pose estimates. The per-robot communication cost scales approximately linearly with team size, from 16.3 KB/s (2 robots) to 78.5 KB/s (20 robots). Consensus latency, which includes ILP solving (Gurobi) under a 100 ms per-round time limit, remains below 100 ms for up to 10 robots (e.g., 92 ms at 10 robots), confirming real-time feasibility at 10 Hz. For larger teams (e.g., 20 robots), latency increases to 241 ms, which may violate stricter real-time constraints (e.g., 50–100 ms).

4.7.3 | *Memory Footprint Analysis*

The memory footprint of SSC-MR is analysed in Figure 10. During image retrieval (Figure 10a), the CPU resident set size (RSS) increases during the initialisation and feature extraction steps, then stabilises at approximately 1.50 GB after the hierarchical feature database is fully loaded. This constant overhead is modest for large-scale scene indexing and remains unchanged during online querying. In pose estimation (Figure 10b), the active GPU memory allocated for inference is only 8.75 MB, despite the use of heavy pre-trained models

TABLE 9 | Per-module latency breakdown for a single robot.

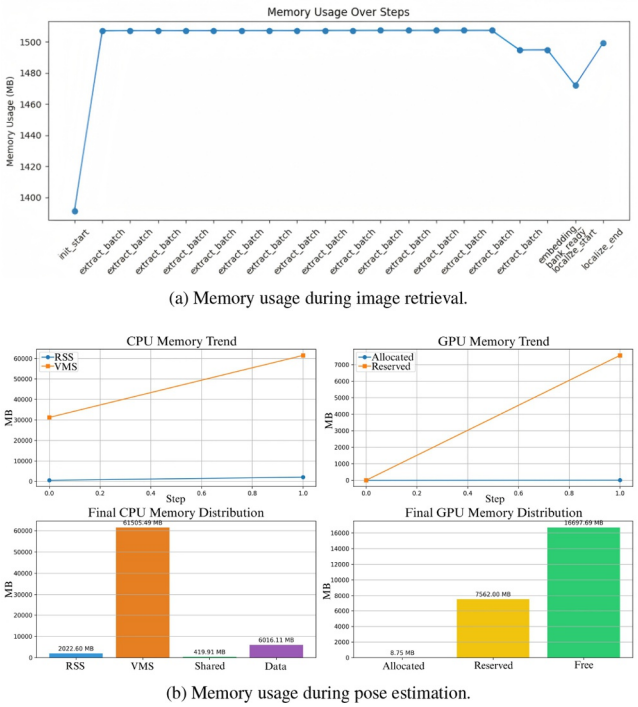
Stage	Module	Latency (ms)
Offline	Feature database construction	183.5 ± 12.3
Online	Image retrieval (GTSIR)	24.7 ± 3.1
	Semantic parsing (Grounding-DINO + SAM)	68.3 ± 5.2
	Feature extraction (SIFT within masks)	12.9 ± 1.4
	Pose estimation (matching + PnP)	9.8 ± 1.2
	Consensus messaging (iRmRC + IRMCC)	4.2 ± 0.6
	Total online latency	119.9 ± 9.5

Note: Measurements are averaged over four datasets. Offline preprocessing is performed once per scene and excluded from online runtime.

TABLE 10 | Consensus messaging overhead for different team sizes.

Robot count	Communication cost (KB/s)	Consensus latency (ms)	Total latency (ms)
2	16.3 ± 2.1	28 ± 4	44.3
5	29.8 ± 3.5	52 ± 7	81.8
10	48.9 ± 5.3	92 ± 12	140.9
15	64.7 ± 6.5	158 ± 21	222.7
20	78.5 ± 7.5	241 ± 32	319.5

Note: Communication cost is measured per robot per consensus round, whereas total latency includes message propagation and ILP solving.



(a) Memory usage during image retrieval.

(b) Memory usage during pose estimation.

FIGURE 10 | Memory usage profile of SSC-MR across the visual localisation pipeline: (a) CPU memory during hierarchical retrieval stabilises at ~1.50 GB after feature database loading; (b) GPU memory remains extremely low (8.75 MB allocated) even with large segmentation models loaded, highlighting the system's efficiency and suitability for resource-constrained platforms.

(Grounding-DINO and SAM). The driver reserves a larger pool of 7.56 GB for caching, which is a standard behaviour of deep learning frameworks (e.g., PyTorch) and does not reflect actual memory demand. The continuous generation of semantic labels adds a peak of 2.1 GB for intermediate activations, but this memory is released after each frame, preventing unbounded growth. For most robotics platforms equipped with 8–16 GB GPU memory, this peak is acceptable. Lightweight variants (e.g., MobileSAM, EfficientDINO) can further reduce the memory footprint on resource-constrained platforms at the cost of a minor accuracy loss, offering a practical trade-off for embedded deployment.

4.7.4 | *Efficiency Trade-Off Statistics*

The overall efficiency of SSC-MR is contextualised in Table 11, which compares end-to-end inference time and memory usage with representative RPE-based methods (HF-Net, DeViLoc, CrossFeat) on the 7Scenes dataset. HF-Net achieves the lowest inference latency (65.2 ms) and CPU memory (892 MB). CrossFeat offers a balanced trade-off (82.3 ms, 386 MB GPU, 1012 MB CPU). DeViLoc has the highest latency (192.0 ms) due to its computationally expensive matching pipeline. SSC-MR, while having a latency of 119.7 ms and CPU memory of 1183 MB, achieves the lowest GPU memory footprint (320 MB). More importantly, SSC-MR delivers the best localisation accuracy (0.018 m/1.12°). Therefore, a favourable accuracy–efficiency trade-off is demonstrated: the moderate increase in inference time (1.5× relative to CrossFeat) is outweighed by a 3.0× reduction in translation error and a 1.5× reduction in rotation error.

TABLE 11 | Efficiency comparison with prior RPE methods.

Method	Inference Latency (ms)	GPU Memory (MB)	CPU Memory (MB)
HF-Net [25]	**65.2 ± 4.8**	412	**892**
DeViLoc [29]	192.0 ± 15.1	568	1245
CrossFeat [30]	<u>82.3</u> ± <u>6.2</u>	<u>386</u>	<u>1012</u>
SSC-MR (ours)	119.7 ± 9.5	**320**	1183

Note: All measurements on RTX 3090 with 7Scenes dataset. Best and second-best results are highlighted in bold and underlined, respectively.

5 | Conclusions

This paper presented SSC-MR, a unified framework for scalable multi-robot visual localisation that integrates multi-granular semantics with a synergic consensus mechanism. Instead of relying solely on low-level cues, the framework leverages semantic coherence across multiple granularities to overcome variations in viewpoint, appearance and platform. The key contribution lies in its dual-consensus design: intra-robot redundancy compression (by iRmRC) improves individual efficiency, whereas inter-robot correlation matching (by IRMCC) enables robust cross-robot alignment. Beyond the reported performance gains, this work offers a broader insight: semantic stability can systematically compensate for environmental dynamics where traditional visual matching fails.

Despite its strengths, SSC-MR remains sensitive to semantic parsing errors, which may propagate through the consensus pipeline. Its correlation-graph optimisation has quadratic complexity, limiting real-time scalability beyond 10–12 robots and the ILP solver can be replaced by approximate heuristics (e.g., greedy or Hungarian).

Future work will focus on tighter integration of semantic and geometric learning, adaptive consensus for larger-scale deployments and lifelong updating of semantic–topological models in nonstationary environments.

Acknowledgements

This work was jointly supported by the grants from the National Key Research and Development Program of China (No. 2024YFB4709802),

the National Natural Science Foundation of China (No. 62373009), the Guangdong S&T Program (No. 2024B0101050002), the Major Key Project of Pengcheng Laboratory (No. PCL2024A01), and the Mobile Information Networks-National Science and Technology Major Project (No. 2025ZD1302900).

Conflicts of Interest

Hong Liu is an Executive Editor-in-Chief for the journal, and was not involved in peer review process or the decision to publish this article. The authors declare that they have no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. Code is available at: https://github.com/wwsbot/SSC-MR.

References

1. L. Chen, C. Liang, S. Yuan, M. Cao, and L. Xie, "Relative Localizability and Localization for Multirobot Systems," *IEEE Transactions on Robotics* 41, no. 1 (2025): 2931–2949, https://doi.org/10.1109/tro.2025.3544103.

2. D. S. Drew, "Multi-Agent Systems for Search and Rescue Applications," *Current Robotics Reports* 2, no. 2 (2021): 189–200, https://doi.org/10.1007/s43154-021-00048-3.

3. M. Xu, Y. Wang, B. Xu, et al., "A Critical Analysis of Image-Based Camera Pose Estimation Techniques," *Neurocomputing* 570 (2024): 127125, https://doi.org/10.1016/j.neucom.2023.127125.

4. C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, "Deep Learning for Visual Localization and Mapping: A Survey," *IEEE Transactions on Neural Networks and Learning Systems* 35, no. 12 (2023): 17000–17020, https://doi.org/10.1109/tnnls.2023.3309809.

5. J. Miao, K. Jiang, T. Wen, et al., "A Survey on Monocular Re-Localization: From the Perspective of Scene Map Representation," *IEEE Transactions on Intelligent Vehicles* 10, no. 4 (2024): 2519–2550, https://doi.org/10.1109/tiv.2024.3378716.

6. W. Chen, W. Chi, S. Ji, et al., "A Survey of Autonomous Robots and multi-robot Navigation: Perception, Planning and Collaboration," *Biomimetic Intelligence and Robotics* 5, no. 2 (2025): 100203, https://doi.org/10.1016/j.birob.2024.100203.

7. H. Wan, Y. Chen, Y. Deng, et al., "Embodiedagent: A Scalable Hierarchical Approach to Overcome Practical Challenge in Multi-Robot Control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2025), 12140–12146.

8. S. Wan, W. Zou, J. Dong, and Y. Gao, "Dual Strategies Consensus Reaching Process for Ranking Consensus Based Probabilistic Linguistic Multi-Criteria Group Decision-Making Method," *Expert Systems with Applications* 262, no. 1 (2025): 125342, https://doi.org/10.1016/j.eswa.2024.125342.

9. W. Burger and M. J. Burge, "Scale-Invariant Feature Transform (SIFT)," in *Digital Image Processing: An Algorithmic Introduction* (Springer, 2022), 709–763.

10. J. Shotton, B. Glocker, C. Zach, et al., "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2013), 2930–2937.

11. A. Kendall, M. Grimes, and R. Cipolla, "POSENET: A Convolutional Network for Real-Time 6-DoF Camera Relocalization," in *Proceedings of the IEEE International Conference on Computer Vision* (2015), 2938–2946.

12. H. Taira, M. Okutomi, T. Sattler, et al., "INLOC: Indoor Visual Localization With Dense Matching and View Synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7199–7209.

13. C. Toft, W. Maddern, A. Torii, et al., "Long-Term Visual Localization Revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 4 (2020): 2074–2088, https://doi.org/10.1109/tpami.2020.3032010.

14. M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography," *Communications of the ACM* 24, no. 6 (1981): 381–395, https://doi.org/10.1145/358669.358692.

15. S. Chen, Z. Wang, and V. Prisacariu, "Direct-Posenet: Absolute Pose Regression With Photometric Consistency," in *2021 International Conference on 3D Vision* (2021), 1175–1185.

16. P. E. Sarlin, A. Unagar, M. Larsson, et al., "Back to the Feature: Learning Robust Camera Localization From Pixels to Pose," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 3247–3257.

17. V. Polizzi, M. Cannici, D. Scaramuzza, et al., "FAVOR: Features via Voxel Rendering for Camera Relocalization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2025), 44–53.

18. E. Brachmann, A. Krull, S. Nowozin, et al., "DSAC-Differentiable Ransac for Camera Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 6684–6692.

19. E. Brachmann and C. Rother, "Learning Less is More-6D Camera Localization via 3D Surface Regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 4654–4662.

20. E. Brachmann and C. Rother, "Visual Camera Re-Localization From RGB and RGB-D Images Using DSAC," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 9 (2021): 5847–5865, https://doi.org/10.1109/TPAMI.2021.3070754.

21. Z. Huang, H. Zhou, Y. Li, et al., "VS-NET: Voting With Segmentation for Visual Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 6101–6111.

22. X. Li, S. Wang, Y. Zhao, et al., "Hierarchical Scene Coordinate Classification and Regression for Visual Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 11983–11992.

23. S. Wang, Z. Laskar, I. Melekhov, et al., "HSCNet++: Hierarchical Scene Coordinate Classification and Regression for Visual Localization With Transformer," *International Journal of Computer Vision* 132, no. 7 (2024): 2530–2550, https://doi.org/10.1007/s11263-023-01982-9.

24. J. Nie, L. Cai, Q. Geng, et al., "Understanding Matters: Semantic-Structural Determined Visual Relocalization for Large Scenes," in *Proceedings of the International Joint Conference on Artificial Intelligence* (2025), 8759–8767.

25. P. E. Sarlin, C. Cadena, R. Siegwart, et al., "From Coarse to Fine: Robust Hierarchical Localization at Large Scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 12716–12725.

26. F. Xue, X. Wu, S. Cai, et al., "Learning Multi-View Camera Relocalization With Graph Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 11372–11381.

27. Q. Zhou, T. Sattler, M. Pollefeys, et al., "To Learn or Not to Learn: Visual Localization From Essential Matrices," in *Proceedings of the IEEE International Conference on Robotics and Automation* (2020), 3319–3326.

28. F. Khatib, Y. Margalit, M. Galun, et al., "Leveraging Image Matching Toward end-to-end Relative Camera Pose Regression," in *DAGM German Conference on Pattern Recognition* (Springer, 2024), 185–201.

29. K. T. Giang, S. Song, and S. Jo, "Learning to Produce Semi-Dense Correspondences for Visual Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), 19468–19478.

30. A. Jaenal, P. C. Cubero, J. Araújo, et al., "Towards Visual Localization Interoperability: Cross-Feature for Collaborative Visual Localization and Mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2025), 26783–26792.

31. N. Kobyshev, H. Riemenschneider, and L. Van Gool, "Matching Features Correctly Through Semantic Understanding," in *Proceedings of the IEEE International Conference on 3D Vision* (2014), 472–479.

32. C. Toft, E. Stenborg, L. Hammarstrand, et al., "Semantic Match Consistency for Long-Term Visual Localization," in *Proceedings of the European Conference on Computer Vision* (Springer, 2018), 383–399.

33. T. Wang, X. Mao, C. Zhu, et al., "Embodiedscan: A Holistic Multi-Modal 3d Perception Suite Towards Embodied AI," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), 19757–19767.

34. S. Wu, H. Fei, and T. S. Chua, "Universal Scene Graph Generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), 14158–14168.

35. I. Armeni, Z. Y. He, J. Gwak, et al., "3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 5664–5673.

36. N. Hughes, Y. Chang, and L. Carlone, "Hydra: A Real-Time Spatial Perception System for 3D Scene Graph Construction and Optimization," in *Proceedings of the Robotics: Science and Systems Conference*, vol. 50 (2022), 1–13.

37. N. Hughes, Y. Chang, S. Hu, et al., "Foundations of Spatial Perception for Robotics: Hierarchical Representations and Real-Time Systems," *International Journal of Robotics Research* 43, no. 10 (2024): 1457–1505, https://doi.org/10.1177/02783649241229725.

38. J. Zhang, G. Zhu, S. Li, et al., "Multiview Scene Graph," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 37 (2024): 17761–17788.

39. X. Liu, S. Wen, J. Zhao, T. Z. Qiu, and H. Zhang, "Edge-Assisted Multi-Robot Visual-Inertial SLAM With Efficient Communication," *IEEE Transactions on Automation Science and Engineering* 22 (2024): 2186–2198, https://doi.org/10.1109/tase.2024.3376427.

40. T. K. Tasooji and R. Parasuraman, "Distributed Fault-Tolerant Multi-Robot Cooperative Localization in Adversarial Environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (2025), 4626–4633.

41. Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno, "Multi-Robot Collaborative Perception With Graph Neural Networks," *IEEE Robotics and Automation Letters* 7, no. 2 (2022): 2289–2296, https://doi.org/10.1109/lra.2022.3141661.

42. F. Wang, C. Zhang, W. Zhang, et al., "Semantic-Geometric Triple Constraint-Based Graph Matching for Robust Multi-Robot Global Localization in Large-Scale Environments," *IEEE Robotics and Automation Letters* (2025).

43. G. G. de Castro, T. M. Santos, F. A. Andrade, et al., "Heterogeneous Multi-Robot Collaboration for Coverage Path Planning in Partially Known Dynamic Environments," *Machines* 12, no. 3 (2024): 200, https://doi.org/10.3390/machines12030200.

44. M. Oquab, T. Darcet, T. Moutakanni, et al., "DINOv2: Learning Robust Visual Features Without Supervision," *Transactions on Machine Learning Research* (2024): 1–32.

45. S. Liu, Z. Zeng, T. Ren, et al., "Grounding DINO: Marrying Dino With Grounded Pre-Training for Open-Set Object Detection," in *Proceedings of the European Conference on Computer Vision* (Springer, 2024), 38–55.

46. A. Kirillov, E. Mintun, N. Ravi, et al., "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 4015–4026.

47. F. Distel, I. Bloch, and J. Atif, "Concept Dissimilarity Based on Tree Edit Distance and Morphological Dilation," in *Proceedings of the European Conference on Artificial Intelligence* (IOS Press, 2014), 249–254.

48. A. Kaplan, T. Avraham, and M. Lindenbaum, "Interpreting the Ratio Criterion for Matching SIFT Descriptors," in *Proceedings of the European Conference on Computer Vision* (Springer, 2016), 697–712.

49. R. Lu, Z. Zhu, S. Fu, et al., "Self-Supervised Camera Relocalization With Hierarchical Fern Encoding," *IEEE Transactions on Instrumentation and Measurement* 73, no. 1 (2023): 1–12, https://doi.org/10.1109/tim.2023.3347802.

50. G. Baruch, Z. Chen, A. Dehghan, et al., "ARKitScenes: A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data," in *Proceedings of the Advances in Neural Information Processing Systems* (2021), 1–12.

51. L. Gurobi Optimization, Gurobi Optimizer Reference Manual (2024), https://www.gurobi.com.

52. E. Rublee, V. Rabaud, K. Konolige, et al., "ORB: An Efficient Alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision* (2011), 2564–2571.

53. D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-Supervised Interest Point Detection and Description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2018), 224–236.

54. M. Tyszkiewicz, P. Fua, and E. Trulls, "DISK: Learning Local Features With Policy Gradient," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33 (2020): 14254–14265.

55. J. Sun, Z. Shen, Y. Wang, et al., "LoFTR: Detector-Free Local Feature Matching With Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 8922–8931.