SceneGraphLoc: Cross-Modal Coarse Visual Localization on 3D Scene Graphs

Anonymous ECCV 2024 Submission

Paper ID #1255

Abstract. This paper introduces a novel problem, *i.e.*, the coarse localization of an input image within a multi-modal reference map represented by a 3D scene graph. This graph comprises multiple modalities, including object-level point clouds, images, attributes, and relationships between objects, offering a lightweight and efficient alternative to conventional methods that rely on extensive image databases. Given the available modalities, the proposed method SceneGraphLoc learns a fixed-sized embedding for each node (*i.e.*, representing an object instance) in the scene graph, enabling effective matching with the objects visible in the input query image. This strategy significantly outperforms other cross-modal methods, even without incorporating images into the map embeddings. When images are leveraged, SceneGraphLoc achieves performance close to that of state-of-the-art techniques depending on large image databases, while requiring *three* orders-of-magnitude less storage and operating orders-of-magnitude faster. The source code will be made publicly available.

1 Introduction

Coarse visual localization, or place recognition, is a fundamental component in computer vision and robotics applications, defined as the task of identifying the approximate location where a query image was taken, given a certain tolerance level [2, 9, 18–20, 24, 29, 40, 41, 51, 58, 59, 69, 113–115, 124, 129]. This capability is crucial for estimating the state of robots and is widely utilized in autonomous, unmanned aerial, terrestrial, and underwater vehicles, as well as AR/VR devices. The task is typically approached as an image retrieval problem, where the image to be localized is compared against a large database of posed images and, optionally, a 3D reconstruction of the scene. The most similar images retrieved from the database are used to estimate the precise location of the query image.

The challenge with current state-of-the-art image-based localization methods, such as [55], is their dependency on extensive image databases, which are not only *storage-heavy* but also *slow* to query, despite optimizations through hashing and other tricks. Moreover, these methods typically necessitate that the query and database share the same modality, limiting the scope of their application. Cross-modal approaches, such as [97, 136], which attempt to bridge different types of data, often restrict their scope to connecting two modalities at a time

(*e.g.*, image-to-point cloud or image-to-bird's eye view map), one for the query and one for the database, thus narrowing their potential applications.

This paper addresses the *novel* challenge of localizing a query image within a database that is represented not by conventional images but by a 3D scene graph [3, 121] that integrates a diverse set of modalities, including point clouds, images, semantics, object attributes, and relationships. We tackle this problem by learning to map these modalities into a unified embedding space, thus allowing us to represent indoor scenes compactly through their objects (*e.g.*, table and wall). This method enables the creation of small, efficient databases and significantly accelerates the coarse localization process.

Contributions. The primary contributions of this paper are as follows:

1. Introducing a novel problem: cross-modal localization of a query image within a 3D scene graph incorporating a mixture of modalities.
2. SceneGraphLoc, a new method for the coarse localization of an input image given a reference map represented by a 3D scene graph. Even without incorporating images into the map, SceneGraphLoc largely outperforms other cross-modal methods on two large-scale, real-world indoor datasets. With images utilized, SceneGraphLoc achieves performance close to that of state-of-the-art image-based methods while requiring *three* orders-of-magnitude less storage and operating orders-of-magnitude faster.

2 Related Work

Localization, the process of determining the position and orientation of an agent within a pre-built map, is pivotal across various domains such as autonomous driving [43], drones [65], and augmented reality [16, 71]. The differentiation in localization techniques arises from their scene representation methods – be it through explicit 3D models [32, 33, 52, 64, 68, 72, 95, 96, 98, 99, 102, 109, 130], sets of posed images [10, 85, 133, 135], or implicitly via neural network weights [6, 11–13, 17, 56, 57, 76, 116, 119] – and their approach to camera pose estimation, whether by 2D-3D [32, 33, 95, 96, 99, 109, 130] or 2D-2D [10, 135] matches, or through a composite of base poses [56, 57, 76, 85, 100, 119]. In practice, localization comprises two main steps: a coarse and precise stage. Here, we focus on the coarse step, finding potential locations of a query image.

Coarse Localization (or place recognition) is often cast as an image retrieval problem [2, 8, 9, 28, 41, 51, 55, 59, 82, 83, 124] that consists of two phases. In the offline indexing phase, a reference map (image or point cloud database) is gathered. In the online retrieval phase, a query image – captured during a future traverse of the environment – is localized coarsely by retrieving the closest match to this image in the reference map. Recent methods perform the retrieval using learned embeddings that are produced by a feature extraction backbone equipped with a head that implements some form of aggregation or pooling, the most notable being NetVLAD [2]. While these methods achieve impressive results, they are limited to a single modality (*e.g.*, images) and require large databases.

Localization using multi-modal data. While dense mesh models are not as widely adopted as sparse Structure-from-Motion-based approaches, they have nonetheless been the focus of considerable research efforts [4,5,14,15,36,80,86,90, 104,106,107,112,134,136]. The body of prior work can be broadly segmented into two main strategies: The first entails the precise alignment of actual images with three-dimensional models (which may be coarse) through applying specialized techniques such as ones using contours [86] or skylines [90]. The second strategy emphasizes the identification and matching of local image features [4, 5, 14, 36, 104,106,107,112,123,134,136], a method that has gained traction for its ability to match real-world images with non-photorealistic renderings of colored meshes, or even meshes without color [14,80,112]. CAD and other models are also commonly used by object pose estimation [4, 31, 35, 37, 45–47, 62, 87]. Image to LiDAR localization [7, 27, 44, 105] is also relevant, especially in robotics applications.

Another variant of multi-modal data localization involves cross-view matching. This technique determines the camera position by finding correspondences between a ground-level query image and a two-dimensional bird's eye view map, such as a satellite image or a semantic landscape map [48,49,66,97,104,118,126]. Other cross-modal techniques were also proposed to involve semantics [29, 30] and event cameras [54] in image-based localization.

These approaches, while demonstrating promising localization results, are limited to interactions between two modalities – one for the query and one for the reference database. Our approach, in contrast, seeks the cross-modal coarse localization of a query image in a database composed of multiple modalities, *e.g.*, 3D point clouds, semantics, object attributes, and relationships.

Scene representation, encapsulating various scene attributes has evolved significantly, yielding diverse surface representations from explicit forms (3D point clouds [88], meshes [38], surfels [108]) to implicit ones (occupancy [61], signed distance functions [22,53]). The advent of neural representations has introduced novel means of encoding geometry [73, 81, 84, 125], appearance [74, 78], and semantics [77]. A comprehensive review is provided by Tewari et al. [111]. The integration of directions/rays [75] and visibility encoding in surface reconstruction [101] has further enriched this domain. Armeni et al. [3] introduced the 3D scene graph structure as a multi-layer representation of a scene that captures geometry, semantics, objects, and camera poses in a unified manner. Subsequent efforts [93, 121] have further advanced 3D scene graph learning and structure.

The increasing interest in 3D scene graphs [3, 60, 93, 121] underscores their potential as structured, rich descriptors for real-world scenes. Methods range from online incremental construction [50, 127] to offline generation from RGB-D imagery [3, 92, 121], and approaches for scene graph prediction [131, 132]. Their application spans embodied AI [91,92,103], task completion [1,26], variability estimation [70], and SLAM [50,92]. Recent studies like [128] introduce frameworks for localizing unseen objects by utilizing 3D scene graphs and graph neural networks for relation prediction, showcasing the utility of scene graphs in enhancing spatial understanding. Similarly, [94] offers new perspectives on 3D scene alignment, employing node matching within overlapping scene graphs to facilitate

precise 3D map alignment. Despite these significant advancements underscoring the value of scene graphs, their potential in multi-modal localization remains largely untapped. In this paper, we use a scene graph representation of the map in which we aim to localize a query image.

3 Visual Localization with 3D Scene Graph

Problem Statement. Let us assume that we are provided with a pre-constructed map of the environment, denoted as $\mathcal{G}$, which is represented as a set of $N \in \mathbb{N}^+$ 3D scene graphs $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$, such that $\mathcal{G} = \{\mathcal{G}_i\}_{i\in[0,N)}$. Having separate graphs $\mathcal{G}_i$ is analogous to the hierarchy levels presented in [50, 92], each representing a group of object instances constituting a place like a room or building. Vertices $v \in \mathcal{V}_i$ symbolize instances of objects (*e.g.*, chairs, tables) and large instances of semantic categories (*e.g.*, walls, ground) within the scene, $i \in [0, N)$. Let $\mathcal{V} = \bigcup_{i\in[0,N)} \mathcal{V}_i$ aggregate all objects across the scenes. Edges $\mathcal{E}_i = \{(v_j^i, v_k^i) | v_j^i, v_k^i \in \mathcal{V}_i\}$ delineate the relationships between objects, such as "nearby", "standing on", and "attached to".

For each vertex v, we introduce $M \in \mathbb{N}^+$ modalities $f_j : \mathcal{V} \to \mathcal{M}_j$ for $j \in [0, M)$, where f_j maps vertex v to the j^{th} modality $\mathcal{M}_j$ that may be $\mathcal{M}_j \in \{\texttt{position}, \texttt{orientation}, \texttt{point cloud}, \texttt{semantic category}, \texttt{image}, \texttt{attribute}\}$. While this paper focuses on these modalities, the set is easily extendable by incorporating additional modalities, such as `textual description` or `floor plan`.

Given $\mathcal{G}$ and an input query image I, the objective is to identify the scene graph $\mathcal{G}_i$ that corresponds to the space depicted in image I. Formally, we aim to resolve the following problem:

$$\mathcal{G}_{i^*} = \underset{i\in[0,N)}{\arg\max} [\![\texttt{contains}(\mathcal{G}_i, \mathbf{p}_I)]\!], \tag{1}$$

where $\mathbf{p}_I \in \mathbb{R}^3$ is the unknown 3D position of the image, and $[\![.]\!]$ is the Iverson bracket which equals to 1 if the condition inside holds and 0 otherwise. It is important to note that our objective is *coarse* localization, opting for the selection of $\mathcal{G}_i$ without needing precise estimation of $\mathbf{p}_I$.

To this end, the optimization problem can be reformulated to incorporate the chirality constraint, asserting that if an object is visible in image I, it must be positioned in front of the camera in 3D space, unobstructed by any entities (*e.g.*, walls). Consequently, the problem becomes:

$$\mathcal{G}_{i^*} = \underset{i\in[0,N)}{\arg\max} \sum_{o_I\in\mathcal{O}_I} [\![\texttt{visible}(\mathcal{G}_i, o_I)]\!] \approx \underset{i\in[0,N)}{\arg\max} \sum_{o_I\in\mathcal{O}_I} \log P(o_I \mid \mathcal{G}_i), \tag{2}$$

where $\mathcal{O}_I$ is the set of objects visible in image I, object $o_I \subseteq \{(x, y) \in I\}$ is a set of pixels in the image ($o_I \in \mathcal{O}_I$), function $\texttt{visible} : \mathcal{G}_i \times \mathcal{O}_I \to \{0, 1\}$ implies that an object is visible in a scene graph, $P(o_I \mid \mathcal{G}_i)$ is the probability of o_I stemming from $\mathcal{G}_i$, and $[\![.]\!]$ is the Iverson bracket, which equals to 1 if the condition inside holds, and 0 otherwise.

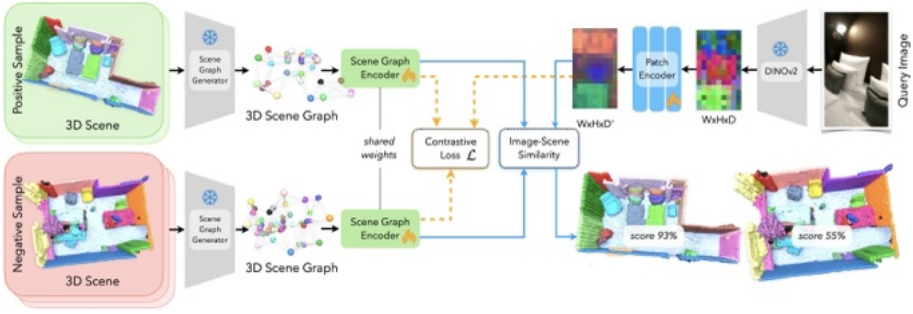

Fig. 1: Overview. The blue arrange arrows indicate training time and the blue show inference time.

Function `visible` can be approached as an indicator of whether object o_I appears in graph G_i. It holds if and only if there exists a $v \in \mathcal{V}_i$ such that v represents the same object as o_I. Therefore, this problem can be approached as matching a set of image pixels (o_I) to a scene graph node (v) that comprises a set of modalities. In the next sections, we will describe a way to learn a unified embedding space for both o_I and v such that they become matchable.

Proposed Pipeline is shown in Fig. 1. It consists of two concurrent stages: the first one generates object embeddings $e_q \in \mathbb{R}^D$ from patches $q \in \mathcal{Q}_I$ within the query image I, and the second derives node embeddings $e_v \in \mathbb{R}^D$ for nodes $v \in \mathcal{V}_i$ in the scene graph $\mathcal{G}_i$. The training objective is to make $\delta(e_q, e_v) = 0$ if and only if the object associated with node v is directly visible (*i.e.*, neither occluded nor outside the camera frustum) through the image patch q. Here, $\delta : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ denotes inverse cosine similarity.

After generating e_q and e_v, the model performs nearest neighbor matching (NN) for each patch in each scene graph $\mathcal{G}_i$, assigning node v to q such that

$$\mathrm{NN}(q, \mathcal{V}_i) = \arg\min_{v \in \mathcal{V}_i} \delta(e_q, e_v). \tag{3}$$

Through this matching procedure, we establish patch-to-node correspondences $\mathcal{C}_i = \{(q, v) \mid v = \mathrm{NN}(q, \mathcal{V}_i) \in \mathcal{V}_i, q \in \mathcal{Q}_I\}$. Based on $\mathcal{C}_i$, we devise an image-to-scene graph similarity score enabling us to deduce whether image I corresponds to the space represented by the scene graph $\mathcal{G}_i$. Finally, potential coarse locations of the image are selected by maximizing the similarity score across the stored scene graphs as depicted in Eq 2.

TODO. Add a figure better depicting the task.

add a figure better depicting the definition of scene graph

3.1 Object Embeddings in the Scene Graph

This section aims to obtain an embedding for each node within the scene graph $\mathcal{G}_i$, encapsulating information from *all* available modalities. Our method builds

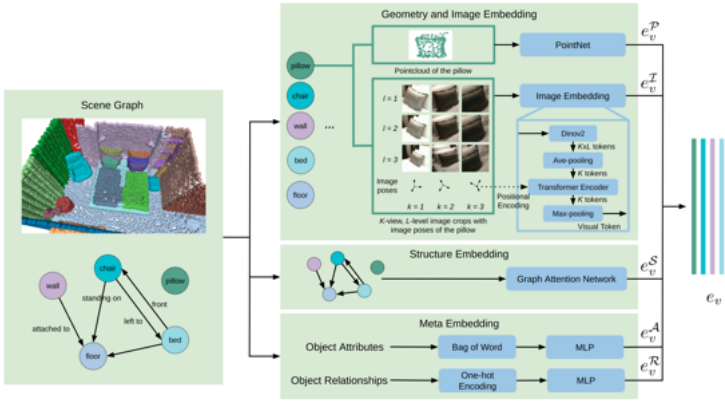

Fig. 2: Scene Graph Encoder.

upon the method of Sarkar et al. [94], with enhancements to include the `image` modality and to distill a unified embedding from all modalities, rather than merely concatenating separate embeddings as in [94].

Scene graphs are conceptualized as multi-modal knowledge graphs, similar to those used in entity alignment, treating semantic and geometric information as distinct modalities. The objective is to learn a joint multi-modal embedding from the individual modal encodings (uni-modal), ensuring nodes corresponding to the same object instance across different graphs are closely positioned. This involves the creation of uni-modal embeddings for the three primary types of 3D scene graph information: *object* embeddings encoding nodes in $\mathcal{V}$, *structure* embeddings $\mathcal{S}$ representing edges in $\mathcal{E}$ as a structured graph, and two *meta* modalities encoding attributes ($\mathcal{A}$) and relationships ($\mathcal{R}$) between objects as one-hot vectors. These uni-modal embeddings are then combined in a weighted manner and optimized jointly through knowledge distillation.

Each of these modalities is processed separately to generate uni-modal embeddings, which are subsequently integrated to model complex inter-modal interactions within the joint embedding space.

Object Embedding. Node $v \in \mathcal{V}$ may contain multiple modalities, such as `point cloud` ($\mathcal{P}$) and `image` ($\mathcal{I}$). Point clouds contain rich geometric information about objects. The point cloud corresponding to each $v \in \mathcal{V}$ is inputted to the object encoder. We employ the PointNet architecture [88] as the object encoder to extract the geometric feature $e_v^{\mathcal{P}}$ for every node.

Furthermore, to enrich the scene graph encoder with a more nuanced understanding of image information, we integrate multi-level and multi-view visual embeddings. The pipeline is visualized in Fig. TODO.

For each node v denoting a 3D object, a selection process identifies a subset of $K \in \mathbb{N}^+$ posed images $\{I^v_{\text{db},k} \mid k = [0, K)\} \subseteq \mathcal{I}_{\text{db}}$ from the database $\mathcal{I}_{\text{db}}$, where v is visible. Such image data is usually available during the construction of the

scene graph [50,127]. This subset is ordered such that $\phi(I^v_{\text{db},0}, v) \geq \phi(I^v_{\text{db},1}, v) \geq ... \geq \phi(I^v_{\text{db},K-1}, v)$, with the visibility function ϕ quantifying the extent of node v observed in each image through pixel count. Visibility determination leverages image poses and the 3D point cloud.

Drawing inspiration from OpenMask3D [110], for any given view I^v_{db} of object v, initial steps include calculating the bounding box $b_{v,0}$ of v within the image, followed by the generation of multi-level bounding boxes $\{b_{v,l} \mid l \in [0, L)\}$ through iterative enlargement of $b_{v,0}$. This enlargement strategy aims to capture contextual information around object v. Subsequently, Dino v2 [79] processes the image crops defined by $b_{v,l}$, extracting multi-level features $\{f_l \mid l \in [0, L)\}$. For each image $I^v_{\text{db},k}$, a max pooling operation aggregates these multi-level crop features into a singular feature vector $f = \max \text{pool}\{f_l \mid l \in [0, L)\}$. The final step of this process involves the application of a Transformer encoder, which incorporates image poses as positional encodings. This step synthesizes multi-view object tokens into a cohesive visual embedding $e^{\mathcal{I}}_v$, effectively integrating the diverse perspectives and levels of contextual information pertaining to each object within the database. Please note that the image database does not necessarily have to be stored after distilling the object embeddings.

Structure Embedding. 3D Scene Graphs encapsulate the object relationships, which we exploit to encode their spatial configuration. This relational data is represented through a *structure* graph, where node features embody the relative translations between object instances, and edges denote these relationships. The relative translation is determined by calculating the distance from the object instance with the maximal number of connections to any other object in the scene. Specifically, this distance is computed using the centroid of the object point cloud, approximated as the centroid of its convex hull. To encapsulate this structural information within $\mathcal{G}_i$, a Graph Attention Network (GAT) [117] is utilized, with the weight matrix constrained to a diagonal form to reduce computational demands and enhance model scalability. Following the method outlined in [94], the structural embedding $e^{\mathcal{S}}_v$ is derived from the final layer of a two-layer GAT model, aggregating neighborhood data across multiple hops.

Meta Embeddings. In addition to the geometry and structure, the object attributes and the inter-object relationships are captured in two distinct embeddings, $e^{\mathcal{R}}_v$ and $e^{\mathcal{A}}_v$. The relationships an object maintains with others are conceptualized as a bag-of-words feature vector, which is input to a feed-forward neural network layer, distilled in the relational embedding $e^{\mathcal{R}}_v$. A similar approach is employed for the attributes associated with the objects, producing the attribute embedding $e^{\mathcal{A}}_v$.

Joint Embedding. Similar to [94], we concatenate each uni-modal feature to a single compact representation for each object v. In contrast to [94], we encapsulate it in a multi-layer perceptron (MLP) in order to learn an embedding whose size is independent of the number of available modalities and fuses information from all. Embedding e_v is calculated as follows:

$$e_v = \text{MLP}\left(\oplus_{k \in \mathcal{K}} \left[\frac{\exp(w_k)}{\sum_{j \in \mathcal{K}} \exp(w_j)} e^k_v\right]\right), \tag{4}$$

where $\oplus$ is the concatenation operator, $\mathcal{K} = \{\mathcal{P}, \mathcal{I}, \mathcal{S}, \mathcal{R}, \mathcal{A}\}$, and w_m is a trainable attention weight for each modality $k \in \mathcal{K}$. A two-layer MLP is applied to map the dimension of the concatenated multi-modal descriptors from $D^{\mathcal{K}}$ to D. We apply L_2 normalization to each uni-modal feature before concatenation.

In practice, these independent modalities are only required and used in the mapping phase of the procedure. This stage involves the construction of an environmental map in the form of a 3D scene graph, during which each node v is distilled into an embedding e_v. During localization, we can ignore independent modalities and use only the fixed-sized embeddings e_v. This approach allows for leveraging pre-computed, compact representations of the environment, thereby significantly reducing storage requirements and enhancing efficiency.

3.2 Object Embeddings in the Query Image

In order to solve the optimization problem described in Eq. 2, we need to find object instances $o_I \in \mathcal{O}_I$ in query image I. A straightforward approach to do so would be to apply a 2D panoptic segmentation algorithm, such as Mask2Former [21]. However, we noticed in our experiments that such an approach is particularly susceptible to inaccuracies and failures (over- and under-segmentation) in the panoptic segmentation, severely affecting the accuracy. Therefore, we approach this problem by a visual Transformer (ViT), breaking up the image into rectangular patches $q \in \mathcal{Q}_I$ and distilling an independent embedding for each patch q based on the object visible throughout q. We use Dino v2 [79] as a backbone to obtain patch-level features. These features are then passed through an additional patch encoder trained to create embeddings from the patch features considering the objects visible from each q. For this encoder, we use a 4-layer convolution neural network (CNN) with residual blocks introduced in [42] on the Dino v2 features and a 3-layer MLP to further map the patch feature to dimension D.

3.3 Contrastive Learning

To learn a joint embedding space for both the scene graph nodes and the image patches, we employ contrastive learning. To do so, we form query image and scene graph pairs $(I, \mathcal{G})$. Real-world scenes are rarely static, *e.g.*, objects move or undergo non-rigid deformations and undergo illumination changes [120]. To ensure that the learned embedding is robust to such temporal changes, we use scene graph $\mathcal{G}_I$ from the same temporal point when I was captured, as positive samples, as well as a scene graph $\mathcal{G}_I^t$ from another scan. Graph $\mathcal{G}_I^t$ represents the same place as $\mathcal{G}_I$ but it undergoes temporal changes. An example is shown in Fig. 3.

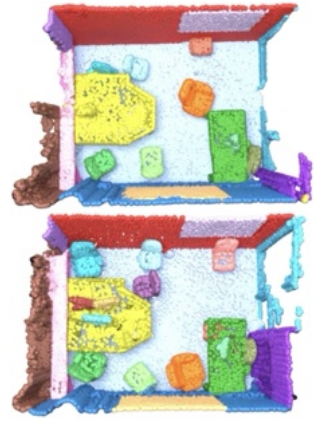

Fig. 3: The same scene at different points in time.

For a query image I, a set of candidate scene graphs $\{\mathcal{G}_I, \mathcal{G}_I^t, \mathcal{G}_1, ..., \mathcal{G}_N\}$ is provided for training, where $\{\mathcal{G}_1, ..., \mathcal{G}_N\}$ act as N negative samples, depicting different scenes than the target scene of the query image.

We train our model by optimizing both a static loss and a temporal loss as:

$$\mathcal{L} = \alpha * \mathcal{L}_{\text{static}} + (1 - \alpha) * \mathcal{L}_{\text{temp}} \tag{5}$$

During training, we assume that image patch to graph node pairs are available [120] such that $P_I = \{(q, v) \mid q \in \mathcal{Q}_I, v \in \mathcal{G}_I\}$ and $\{P_I^t = (q, v^t) \mid q \in \mathcal{Q}_I, v \in \mathcal{G}_I^t\}$. For each pair (q, v) from P_I and each pair from P_I^t, we use the following notation. Set $N_q^{\mathcal{I}} = \{q' \mid q' \in \mathcal{Q}_I, v_{q'} \neq v\}$ contains patches seeing objects other than v. $N_v^{\mathcal{G}} = \{v_n \mid v_n \in \mathcal{V}_I \cup \mathcal{V}_1 \cup ... \cup \mathcal{V}_N\} \setminus \{v\}$ contains the 3D objects of all candidate scene graphs other than v, where $\mathcal{V}_I$ represents the objects of the target graph $\mathcal{G}_I$ and $\mathcal{V}_i$ is the object nodes of other graphs $\mathcal{G}_i$. $N^{\mathcal{G}_v^t} = \{v_n | v_n \in \mathcal{V}_I^t \cup \mathcal{V}_1 \cup ... \cup \mathcal{V}_n\} \setminus \{v\}$, where $\mathcal{V}_I$ represent the object nodes of the target graph $\mathcal{G}_I^t$. The static loss is defined as bi-directional N-pair loss [67] as:

$$\mathcal{L}_{\text{static}} = E_{P_I \in B}\left[E_{(q,v)\in P_I}\left[-\log(\frac{1}{2}(s(q, v, N_q^{\mathcal{I}}, N_v^{\mathcal{G}}) + s(v, q, N_q^{\mathcal{I}}, N_v^{\mathcal{G}}))\right]\right], \tag{6}$$

where $E_{P_I \in B}$ represent loss averaging over a batch of query images and their corresponding candidate scene graphs; $s(q, v, N_q^{\mathcal{I}}, N_v^{\mathcal{G}})$ and $s(v, q, N_q^{\mathcal{I}}, N_v^{\mathcal{G}})$ represent the bi-directional probability distributions of the positive pair as:

$$s(q, v, N_q^{\mathcal{I}}, N_v^{\mathcal{G}}) = \frac{f(e_q, e_v)}{f(e_q, e_v) + \sum_{q_n \in N_q^{\mathcal{I}}} f(e_q, e_{q_n}) + \sum_{v_n \in N_v^{\mathcal{G}}} f(e_q, e_{v_n})},$$
$$s(v, q, N_v^{\mathcal{G}}, N_q^{\mathcal{I}}) = \frac{\delta(e_v, e_q)}{f(e_v, e_q) + \sum_{v_n \in N_v^{\mathcal{G}}} f(e_v, e_{v_n}) + \sum_{q_n \in N_q^{\mathcal{I}}} f(e_v, e_{q_n})},$$

where $f(e_q, e_v) = \exp(\frac{\delta(e_q, e_v)}{\tau})$, $\delta(e_q, e_v)$ represents the cosine similarity between embeddings e_q and e_v, and τ is a temperature parameter. Similarly, the temporal loss term is defined the same as Eq.6 but with $N^{\mathcal{G}_v^t}$ as follows:

$$\mathcal{L}_{\text{temp}} = E_{P_I \in B}\left[E_{(q,v)\in P_I}\left[-log(\frac{1}{2}(s(q, v, N_q^{\mathcal{I}}, N^{\mathcal{G}_v^t}) + s(v, q, N^{\mathcal{G}_v^t}, N_q^{\mathcal{I}}))\right]\right], \tag{7}$$

By minimizing Eq.5, the paired cross-modal embeddings e_q and e_v are pulled together while the embeddings from different objects are pushed apart.

3.4 Scene Graph Retrieval

Given a pre-established map of an environment represented through a collection of scene graphs $\mathcal{G} = \{\mathcal{G}_i \mid i \in [0, N)\}$, where each node embedding has been precomputed, the goal during inference is to identify the top-K scene graphs

$\mathcal{G}_i$ in which image I was likely captured. The method to address this challenge involves calculating the similarity between a graph and the image as follows:

$$s(\mathcal{G}_i, I) = \frac{1}{|\mathcal{Q}_I|} \sum_{q \in \mathcal{Q}_I} \delta(e_q, \mathrm{NN}(q)), \tag{8}$$

where $\mathcal{Q}_I$ denotes the set of image patches in I, and for each patch q, $\mathrm{NN}(q, \mathcal{V}_i) \in \mathcal{V}_i$ represents the nearest node in terms of embedding similarity. The function δ, which is assumed to map values to the interval $[0, 1]$, interprets these distances as probabilities, facilitating the identification of the optimal scene graph $\mathcal{G}_{i^*}$ that maximizes $s(\mathcal{G}_i, I)$. This optimal graph is identified by simply iterating through all potential scene graphs and selecting the one with the highest similarity.

This approach can be accelerated through the use of spatial partitioning techniques, such as kd-trees, for the preprocessing of node embeddings within the map. The process becomes highly efficient by applying a K nearest neighbors search within the embedding space for each image patch. It is important to note that not all scene graphs may be represented among the K nearest nodes for every patch, leading to scenarios where multiple nodes from the same graph dominate the nearest neighbors. In such cases, if a scene graph does not feature within these top candidates for a particular patch, it is assumed that the patch is not visible from that graph. Consequently, the similarity score for that graph is set to zero to reflect this absence of visibility.

4 Experiments

Baselines. Although no existing methods directly address the same challenge as our proposed one, several recent developments in cross-modal coarse localization offer relevant baselines. LidarCLIP [44], tailored for autonomous driving, encodes with Single-stride Sparse Transformer [25] LiDAR point clouds of the observed scene into a global descriptor and aligns it to the embeddings of images with CLIP image encoder [89], subsequently matching these embeddings with a query text. This method, albeit not an exact match for our problem, can be adapted to align point clouds with the CLIP embedding of a query image. Similarly, LIP-Loc [105] utilizes a comparable strategy but projects the LiDAR point cloud into a 2D range image, directly encodes the range image, and matches the embeddings to the query image. For a fair comparison, we fine-tuned both methods on our dataset. Additionally, we formulated a object-retrieval-based baseline with OpenMask3D [110] and OpenSeg [34]. OpenMask3D [110] is implemented to assign CLIP descriptors to 3D object instances by aggregating features from multiple images observing the objects. OpenSeg is applied to extract pixel-level CLIP features from the query image, which are then aggregated in the patch level to patch features by average pooling. Although OpenMask3D and OpenSeg were not initially designed for localization, they can be modified to match the CLIP descriptor of the query image to those of the object instances. To benchmark against leading visual localization methods that rely on extensive image datasets,

Table 1: Retrieval recall (%; the target scene ranked within the top 1, 3, and 5 of the retrieved list) and storage requirements (MB) for methods utilizing point clouds ($\mathcal{P}$), images ($\mathcal{I}$), and other modalities (O = $\{\mathcal{A}, \mathcal{S}, \mathcal{R}\}$)) for map representation in the 3RScan dataset [120]. Additionally, metrics for single-modal methods (CVNet and AnyLoc) reliant on extensive image datasets are presented. The results are reported for scenarios where the target room is chosen from a subset of 10 and 50 candidate scenes.

Method	Map modalities			10 scenes						50 scenes						Storage
	$\mathcal{P}$	$\mathcal{I}$	O	R@1	@3	@5	R^t@1	@3	@5	R@1	@3	@5	R^t@1	@3	@5	(MB)
LidarCLIP [44]	✓	✗	✗	16.3	41.4	60.6	16.3	39.8	61.1	4.7	11.0	16.3	4.1	10.3	15.6	**0.4**
LIP-Loc [105]	✓	✗	✗	14.0	35.8	57.9	10.9	30.0	52.7	2.0	9.1	14.2	2.3	8.6	15.2	1.0
OMask3D [110]	✓	✓	✗	–	–	–	42.3	71.5	85.8	–	–	–	21.1	38.1	48.0	20.1
SceneGraphLoc (Ours)	✓	✗	✓	**53.6**	**81.9**	**92.8**	50.5	76.8	88.4	**30.2**	**50.2**	**61.2**	28.2	46.2	56.4	5.4
SceneGraphLoc (Ours)	✓	✓	✓	–	–	–	**81.5**	**93.9**	**98.0**	–	–	-	**69.3**	**78.6**	**84.4**	5.4
CVNet [63]	✗	✓	✗	–	–	–	79.2	91.0	95.4	–	–	–	66.5	77.0	81.7	239.1
AnyLoc [55]	✗	✓	✗	–	–	–	87.9	94.7	97.5	–	–	–	80.6	87.4	90.0	5720.3

we included CVNet [63] and AnyLoc [55] in our experiments. These methods, while advanced in their performance, require significant storage for global image descriptors and are noted for their slower inference times.

Map Generation. For visual localization approaches, the mapping stage is executed offline as a preprocessing step, necessitating specific mapping operations for each method before proceeding to localization. In the case of our proposed method, this entails passing the point cloud, images, metadata, and relationships through the 3D scene graph encoder outlined in Section 3.1. For LIP-Loc and LidarCLIP, this preparatory step involves converting point clouds into range images and then computing the CLIP embeddings for these images or directly encoding the point cloud of the scenes into global descriptors. In OpenMask3D, the CLIP embeddings corresponding to each object instance are calculated and subsequently stored. For image-based methods like CVNet and AnyLoc, we ensure that embeddings for all images in the database are precomputed.

Metrics. To evaluate the efficacy of a given method, we focus on the recall of scene selection. This entails analyzing the scenario where, given a query image and corresponding scene pair $(I, \mathcal{G}_i)$, alongside $N-1$ alternative scenes from the database, an ordering is established for these scenes according to their computed similarity to I as determined by the tested method. The metric Recall@K is employed to ascertain whether the target scene $\mathcal{G}_i$ is ranked among the top K scenes in terms of similarity as identified by the method being evaluated. Additionally, we will report the inference time and storage requirements.

Experiments on the 3RScan Dataset. The 3RScan dataset [120] comprises 1335 annotated indoor scenes, representing 385 distinct spaces (rooms), with 1178 scenes (338 rooms) allocated for training and 157 (47 rooms) designated for validation. Both the training and validation datasets include semantically annotated 3D point clouds for each scene, with some scenes captured over extended periods (e.g., several months) showcasing environmental changes. Annotations for 3D scene graphs within the 3RScan dataset are provided in [122]. Due to the absence of such annotations in the test set, it was excluded from our exper-

iments. Consequently, we reorganized the original validation set, allocating 34 scenes (17 rooms) for validation purposes and 123 scenes (30 rooms) for testing. For full reproducibility, we will publish this split.

The training procedure involves selecting a query image from the training set alongside its corresponding 3D scene graph. Rectangular patches are formed within the image by a 2D grid, and the directly visible object instance (represented by a graph node) from each patch is determined. In patches where multiple objects are visible, the one with largest visibility is chosen. These object-to-patch matches serve as positive pairs. For negative sample generation, objects from randomly selected alternate scenes were paired with a given patch.

During testing, we examine all 30 rooms within the test set, selecting query images from each room. The next step involves matching this image against all rooms to ascertain whether the correct one could be identified by a method. This procedure is repeated for every image in each room. In total, all methods are tested on 30462 query images. Furthermore, we evaluate scene selection through two approaches: first, by considering all candidate rooms, and second, from a subset of 10 randomly chosen rooms (including the target one). This latter approach emulates a scenario where a preselection strategy is employed, for example, utilizing a global scene descriptor. In image-based methods, we use all images from the database and determine the scene based on the retrieved image.

Additionally, the methods are evaluated under both static and temporal conditions. In the static scenario, the target scene graph for a given query image originates from the same scan, albeit from a different sequence to ensure no image overlap. Conversely, in the temporal scenario, the scene graph is derived from a sequence captured at a different temporal stage than the query, introducing potential environmental changes. We do not show results for methods exploiting images in their maps in the static stage, given that it implies the unrealistic scenario of mapping and localizing small spaces simultaneously.

The results are reported in Table 1. Despite being retrained, LidarCLIP and LIP-Loc display inaccurate results, particularly in scenarios involving the selection of the target room from the entire scene set. LIP-Loc barely surpasses random selection. Although LidarCLIP exhibits marginally better accuracy, it remains substantially inferior to alternative methods. The temporal case further decreases the performance of both methods. OpenMask3D, while achieving better results than LidarCLIP and LIP-Loc, is less accurate than the proposed SceneGraphLoc. SceneGraphLoc, even when excluding the `image` modality ($\mathcal{I}$), outperforms other cross-modal strategies by a significant margin. Incorporating images significantly enhances its performance, positioning it close to that of image-based approaches but with *three orders of magnitude* smaller storage requirements. Also, the storage of SceneGraphLoc with and without images is the same due to its design of distilling knowledge into fixed-sized embeddings.

Experiments on the ScanNet Dataset. In order to evaluate generalization ability of our methods in real-world applications when scene graph annotations are not available, we conduct further experiments in the ScanNet dataset [23]. ScanNet encompasses 1613 monocular sequences of room-scale 3D scenes, offer-

Table 2: Retrieval recall on ScanNet(%; the target scene ranked within the top 1, 3, and 5 of the retrieved list) and storage requirements (MB) for methods utilizing point clouds ($\mathcal{P}$), images ($\mathcal{I}$), and other modalities (O = $\{\mathcal{S}, \mathcal{R}\}$) for map representation in the 3RScan dataset [120]. The modality $\mathcal{A}$ is not available in the scene graphs predicted from [127]. Additionally, metrics for single-modal methods (CVNet and AnyLoc) reliant on extensive image datasets are presented. The results are reported for scenarios where the target room is chosen from a subset of 10, 50 and the complete set of all (210) scenes.

Method	Map modalities			10 scenes			50 scenes			All scenes			Storage
	$\mathcal{P}$	$\mathcal{I}$	O	R^t@1	@3	@5	R^t@1	@3	@5	R^t@1	@3	@5	(MB)
LidarCLIP [44]	✓	✗	✗	19.4	47.5	67.6	4.7	14.8	22.2	5.9	15.0	21.9	**0.7**
LIP-Loc [105]	✓	✗	✗	10.3	27.0	43.6	1.9	6.0	8.1	1.8	3.1	4.0	1.7
OpenMask3D [110]	✓	✓	✗	54.9	84.8	94.0	31.3	51.3	63.2	16.5	27.2	34.5	17.8
SceneGraphLoc (Ours)	✓	✗	✓	54.1	81.4	91.9	29.0	47.4	58.0	13.5	26.4	34.2	9.3
SceneGraphLoc (Ours)	✓	✓	✓	**78.5**	**92.7**	**98.3**	**61.6**	**83.2**	**91.6**	**53.4**	**69.8**	**78.7**	9.3
CVNet [63]	✗	✓	✗	96.5	98.9	99.6	92.6	96.0	97.0	89.9	93.4	94.6	239.1
AnyLoc [55]	✗	✓	✗	98.4	99.4	99.8	96.5	98.1	98.6	95.1	96.9	97.4	5720.3

ing 3D mesh reconstructions alongside the RGBD frame sequences utilized for the reconstructions. Given the absence of scene graph annotations in ScanNet, we run the SceneGraphFusion [127] on the RGBD sequences of scans for 3D reconstruction and scene graph prediction with 3D instance segmentation and object relationships (*i.e.*, graph edges) within these graphs. As the process of scene graph prediction uses the RGBD frames of each scan, we avoid using those RGB images to match to the scene graph predicted of the scan. Thus, we only measure recall in the temporal scenario. Additionally, unlike 3RScan, the frame rate of RGBD sequences in Scannet is high, and motion between consecutive frames is small, for image-based methods [55, 63], each database image is selected from every 25 consecutive frames in the sequence. For a fair comparison, all the methods only use the same selected images for training and evaluation.

For training, we adhere to the use of the official training set, similar to the 3RScan dataset. We divide the official validation set, which includes 312 scenes, into two distinct subsets: the first 100 scenes form our validation set, while the remaining 212 ones are allocated for testing. To ensure full reproducibility, we will make this split publicly available.

The findings on the test set are detailed in Table 2, where results are presented for scenarios selecting the target room from subsets of 10, 50, and the entire set of 210 scenes. The performance of LIP-Loc exhibits a similar pattern to that observed in the 3RScan dataset, performing only slightly better than random selection. LidarCLIP shows a modest improvement in accuracy. OpenMask3D, which leverages images, attains an accuracy comparable to our proposed method without incorporating the image modality. Our proposed SceneGraphLoc, when including the image modality in its map, significantly outperforms all cross-modal approaches. Although there remains a gap in accuracy compared to methods that use extensive image collections as maps (such as CVNet and AnyLoc), SceneGraphLoc benefits from a database size *three* orders-of-magnitude smaller,

Table 3: Average time (ms) of obtaining the query image embedding (t_{e_q}) and of the retrieval from 10, 50, and all scenes from the 3RScan [120] and ScanNet [23] datasets.

Method	3DRScan [120]			ScanNet [23]			
	t_{e_q}	t^{10}_{retr}	t^{50}_{retr}	t_{e_q}	t^{10}_{retr}	t^{50}_{retr}	$t^{\text{all}}_{\text{retr}}$
LidarCLIP [44]	4.1	0.1	0.3	4.9	0.1	0.2	0.6
LIP-Loc [105]	2.7	0.1	0.2	4.1	0.1	0.2	0.5
OpenMask3D [110]	41.5	4.8	7.4	20.1	55.4	1.1	4.5
SceneGraphLoc (Ours)	28.0	0.3	1.5	16.6	1.3	3.7	17.0
CVNet [63]	14.3	9.0	60.0	54.0	10.6	74.1	311.3
AnyLoc [55]	658.40	354.64	1826.39	242.99	68.22	329.01	1451.10

Table 4: Ablation study performed on the val. split of 3RScan [120], analysing map modalities ($\mathcal{P}$ – point cloud, $\mathcal{I}$ – image, $\mathcal{A}$ – attributes, $\mathcal{S}$ – structure, $\mathcal{R}$ – relationships) and the method (Dino v2 or GCVit) to obtain the image embeddings.

Map modalities					Dino v2 [79]						GCVit [39]					
$\mathcal{P}$	$\mathcal{I}$	$\mathcal{A}$	$\mathcal{S}$	$\mathcal{R}$	R@1	@3	@5	R^t@1	@3	@5	R@1	@3	@5	R^t@1	@3	@5
✓					45.2	81.9	93.7	43.9	79.5	91.4	24.6	56.0	76.9	23.2	54.7	77.3
✓		✓			56.3	85.6	95.0	54.8	84.0	95.0	44.2	**76.9**	91.2	43.4	75.1	89.4
✓		✓	✓		58.4	**87.3**	**95.9**	56.5	85.7	93.6	43.3	75.8	**91.3**	41.5	72.8	89.3
✓		✓	✓	✓	**63.7**	86.8	95.8	62.7	87.4	96.3	**45.3**	75.5	90.5	46.6	76.2	90.2
	✓				–	–	–	80.2	96.0	99.0	–	–	–	69.4	87.4	93.7
✓	✓				–	–	–	*	*	*	–	–	–	**73.2**	**89.7**	**95.9**
✓	✓	✓	✓	✓	–	–	–	**88.5**	**97.7**	**99.6**	–	–	–	72.1	88.8	95.7

highlighting its efficiency and effectiveness. We partly attribute this performance gap to the lack of object attributes in the dataset.

Processing Time, measured in milliseconds for various methods applied to the 3DRScan and ScanNet datasets, are detailed in Table 3. The computation time required to generate an embedding for the query image (t_{e_q}) is notably small across all methods, typically not exceeding a few tens of milliseconds, with the exception of AnyLoc, which runs for nearly a second.

The retrieval phase for cross-modal approaches is generally limited to a few tens of milliseconds. However, methods such as CVNet and AnyLoc exhibit slower performance, due to searching through extensive image collections. When tasked with selecting from a large number of images, the processing times of these methods can extend into the range of several hundred milliseconds or even reach upwards of a second.

Ablation Study. Generally, the proposed pipeline consists of query image embedding module, 3D scene graph encoding embedding module and coarse localization module based on both embeddings. In order to better understand the proposed method, we provide ablation studies on (i) the impact of multiple modalities and different 2D backbone choices on the performance of coarse localization; (ii) the correlation between the object information in the scene and the localization performance; (iii) and the impact of different settings of the image

Table 5: Statistics Analysis on the val. split of 3RScan [120], analysing the correlation between multiple factors ($|\mathcal{V}_0^t|$, $\mathcal{H}_I$ and s_I) and the performance of coarse localization (R^t@1 abbreviated as R_1^t and Acc_q^t) under multiple modalities.

Map modalities					R^t@1	$\rho_{(\lvert\mathcal{V}_0^t\rvert, R_1^t)}$	$\rho_{(\mathcal{H}_I, R_1^t)}$	$\rho_{(s_I, R_1^t)}$	Acc_q^t	$\rho_{(\lvert\mathcal{V}_0^t\rvert, Acc_q^t)}$	$\rho_{(\mathcal{H}_I, Acc_q^t)}$
$\mathcal{P}$	$\mathcal{I}$	$\mathcal{A}$	$\mathcal{S}$	$\mathcal{R}$							
✓					43.9	0.20	0.15	0.02	49.2	-0.10	-0.01
✓		✓			54.8	0.21	0.16	0.07	53.8	-0.06	-0.03
✓		✓	✓		56.5	0.29	0.17	0.11	55.9	-0.04	-0.04
✓		✓	✓	✓	62.7	0.38	0.20	0.19	54.8	-0.07	-0.01
	✓				80.2	0.15	0.06	0.19	55.6	-0.06	-0.12
✓	✓										
✓	✓	✓	✓	✓	88.5	0.28	0.15	0.17	64.2	-0.07	-0.03

Table 6: Ablation study on the methods generating image embeddings for the map.

K	R^t@1	@3	@5
1	85.7	96.6	99.4
3	86.1	96.7	99.5
5	87.0	97.4	99.5
7	87.7	97.0	99.4
10	**88.5**	**97.7**	**99.6**

(a) The recall values w.r.t. the number (K) of views used to create an image embedding for a particular object.

Configuration				R^t@1	@3	@5
TE	PE	Max	Mean			
✗	✗	✗	✓	85.5	96.8	99.4
✗	✗	✓	✗	86.0	97.1	99.4
✓	✗	✓	✗	86.6	97.2	99.4
✓	✓	✓	✗	88.5	97.7	99.6

(b) Multi-view image fusion. "Max" and "Mean" indicate max- and average-pooling over the K views, respectively. "TE" indicates using the transformer encoder. "PE" means using camera poses for positional encoding in "TE".

modality encoding on the performance. Additional ablation studies can be found in supplementary materials.

In Table 4, the coarse localization performance of the proposed method is shown with different modalities ($\mathcal{P}$, $\mathcal{I}$, $\mathcal{A}$, $\mathcal{S}$, $\mathcal{R}$) and with different 2D image backbones (Dino V2 [79] and GCVit [39]). From the table, we can see that with Dino V2 as the query image encoding backbone, the localization performance is significantly better than the one with GCVit. Meanwhile, with Dino V2, all introduced modalities of 3D scene graph have contributions to the localization performance to some degree. Among them, integration of modalities of attribute $\mathcal{A}$ and image $\mathcal{A}$ significantly improve the performance, as $\mathcal{I}$ provide distinctive information in texts (e.g. objects' color, category and states) and $\mathcal{I}$ can provide distinctive visual information.

In Table 5, we report the correlation between multiple factors and the localization performance R^t@1 and Acc_q^t with multiple settings of modalities. The following notations are defined:

- $|\mathcal{V}_0^t|$ represents the number of objects of the target scene with temporal changes $\mathcal{G}_I^t$.
- $\mathcal{H}_I$ represents the Shannon entropy of object information observed in image patches $q \in \mathcal{Q}_I$, defined in Eq. 9.

- $s_I = s(\mathcal{G}_I^t, I)$ represents the similarity score between $\mathcal{G}_I^t$ and the query image, defined in Eq. 8.
- Acc_q^t represents the percentage of image patches q_I that are correctly associated to the objects in the scene graph given Eq. 3.
- $\rho(*,*) \in [-1,1]$ represents the Pearson Correlation coefficient between two variable.

$$\begin{aligned} \mathcal{H}_I &= -\sum_{o \in \mathcal{O}_I^{gt}} p_I(o) \log p_I(o), \\ p_I(o) &= \frac{|\{q_I | q_I \in \mathcal{Q}_I, o_I(q_I) = o\}|}{|\mathcal{Q}_I|}, \end{aligned} \tag{9}$$

$\mathcal{O}_I^{gt}$ is the ground truth set of objects observed in query image I and $p_I(o)$ is the frequency of patches observing the object o. $\mathcal{H}_I$ denotes the diversity of objects observed in I, as illustrated in Fig. figure for image object entropy

From the table, we can see that:

- $\rho_{(|\mathcal{V}_0^t|, R_1^t)}$ and $\rho_{(|\mathcal{H}_I|, R_1^t)}$ are greater than 0 by a not negligible amount, denoting positive correlation between $|\mathcal{V}_0^t|$ and R_1^t, and the positive correlation between $\mathcal{H}_I$ and R_1^t. The intuition is that the more objects observed in the query image and located in the target 3D scene graph, the easier the query image can be localized.
- Noticeably, with integration of modalities $\{\mathcal{S}, \mathcal{R}\}$, the correlation $\rho_{(|\mathcal{V}_0^t|, R_1^t)}$, $\rho_{(\mathcal{H}_I, R_1^t)}$. The intuition is that by incorporating $\{\mathcal{S}, \mathcal{R}\}$, the proposed modules learns to leverage scene-context information, e.g. relationship between objects, for object embedding and coarse localization. Thus, the localization accuracy R_1^t benefits from more context information (larger $\rho_{(|\mathcal{V}_0^t|, R_1^t)}$ and $\rho_{(\mathcal{H}_I, R_1^t)}$).
- For patch-object association accuracy Acc_q^t, all modalities except $\mathcal{R}$ have contributions to improving Acc_q^t. On the other hand, there is slightly negative correlation between $\mathcal{H}_I$ or $|V_0^t|$ and Acc_q^t, denoting that the more diversity of the objects, the slightly harder for the image patches to be correctly assigned to certain objects.

In Table 6, we explore the impact on the localization performance of number (K) of views and the multi-view fusion methods in the module of scene graph embedding of $\mathcal{I}$. Table 6a shows that by using more views for modality $\mathcal{I}$, the localization performance improves. Furthermore, Table 6b shows that the localization performance benefits from the transformer encoder with positional encoding followed by max-pooling. The intuition behind positional encoding with image poses is to integrate spatial context with the multi-view visual information for more-informed visual embedding of objects within the scene graph.

References

1. Agia, C., Jatavallabhula, K.M., Khodeir, M., Miksik, O., Vineet, V., Mukadam, M., Paull, L., Shkurti, F.: Taskography: Evaluating robot task planning over large 3d scene graphs. In: Conference on Robot Learning. pp. 46–58. PMLR (2022) 3

2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016) 1, 2
3. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5664–5673 (2019) 2, 3
4. Aubry, M., Russell, B.C., Sivic, J.: Painting-to-3d model alignment via discriminative visual elements. ACM Transactions on Graphics (TOG) **33**(2), 14 (2014) 3
5. Aubry, M., Russell, B.C., Sivic, J.: Visual geo-localization of non-photographic depictions via 2d-3d alignment. In: Large-Scale Visual Geo-Localization (2016) 3
6. Balntas, V., Li, S., Prisacariu, V.: Relocnet: Continuous metric learning relocalisation using neural nets. In: The European Conference on Computer Vision (ECCV) (September 2018) 2
7. Bernreiter, L., Ott, L., Nieto, J., Siegwart, R., Cadena, C.: Spherical multi-modal place recognition for heterogeneous sensor systems. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 1743–1750. IEEE (2021) 3
8. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888 (2022) 2
9. Berton, G., Paolicelli, V., Masone, C., Caputo, B.: Adaptive-attentive geolocalization from few queries: A hybrid approach. In: IEEE Winter Conference on Applications of Computer Vision. pp. 2918–2927 (January 2021) 1, 2
10. Bhayani, S., Sattler, T., Barath, D., Beliansky, P., Heikkilä, J., Kukelova, Z.: Calibrated and partially calibrated semi-generalized homographies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5936–5945 (2021) 2
11. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: Dsac - differentiable ransac for camera localization. In: CVPR (2017) 2
12. Brachmann, E., Rother, C.: Learning less is more - 6d camera localization via 3d surface regression. In: CVPR (2018) 2
13. Brachmann, E., Rother, C.: Visual camera re-localization from rgb and rgb-d images using dsac. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5847–5865 (2021) 2
14. Brejcha, J., Lukač, M., Hold-Geoffroy, Y., Wang, O., Cadik, M.: Landscapear: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In: European Conference on Computer Vision. pp. 295–312. Springer (2020) 3
15. Cadik, M., Sykora, D., Lee, S.: Automated outdoor depth-map generation and alignment. Comput. Graph. **74**, 109–118 (2018) 3
16. Castle, R., Klein, G., Murray, D.W.: Video-rate localization in multiple maps for wearable augmented reality. In: 2008 12th IEEE International Symposium on Wearable Computers. pp. 15–22. IEEE (2008) 2
17. Cavallari, T., Bertinetto, L., Mukhoti, J., Torr, P., Golodetz, S.: Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation. In: 3DV (2019) 2

18. Chen, Z., Jacobson, A., Sunderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., Milford, M.: Deep learning features at scale for visual place recognition. In: 2017 IEEE International Conference on Robotics and Automation. pp. 3223–3230 (2017) 1
19. Chen, Z., Liu, L., Sa, I., Ge, Z., Chli, M.: Learning context flexible attention model for long-term visual place recognition. IEEE Robotics and Automation Letters **3**(4), 4015–4022 (2018) 1
20. Chen, Z., Maffra, F., Sa, I., Chli, M.: Only look once, mining distinctive landmarks from convnet for visual place recognition. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 9–16 (2017) 1
21. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022) 8
22. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996) 3
23. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 12, 14
24. Doan, A.D., Latif, Y., Chin, T.J., Liu, Y., Do, T.T., Reid, I.: Scalable place recognition under appearance change for autonomous driving. In: IEEE International Conference on Computer Vision. pp. 9319–9328 (October 2019) 1
25. Fan, L., Pang, Z., Zhang, T., Wang, Y.X., Zhao, H., Wang, F., Wang, N., Zhang, Z.: Embracing single stride 3d object detector with sparse transformer (2022) 10
26. Gadre, S.Y., Ehsani, K., Song, S., Mottaghi, R.: Continuous scene representations for embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14849–14859 (2022) 3
27. Gao, P., Liang, J., Shen, Y., Son, S., Lin, M.C.: Visual, spatial, geometric-preserved place recognition for cross-view and cross-modal collaborative perception. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 11079–11086. IEEE (2023) 3
28. Garg, S., Fischer, T., Milford, M.: Where is your place, visual place recognition? (2021) 2
29. Garg, S., Suenderhauf, N., Milford, M.: Semantic–geometric visual place recognition: a new perspective for reconciling opposing views. The International Journal of Robotics Research (2019) 1, 3
30. Garg, S., Sünderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., Wu, Q., Chin, T.J., Reid, I., Gould, S., et al.: Semantics for robotic mapping, perception and interaction: A survey. Foundations and Trends® in Robotics **8**(1–2), 1–224 (2020) 3
31. Georgakis, G., Karanam, S., Wu, Z., Kosecka, J.: Learning local rgb-to-cad correspondences for object pose estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8966–8975 (2019) 3
32. Germain, H., Bourmaud, G., Lepetit, V.: Sparse-to-dense hypercolumn matching for long-term visual localization. In: International Conference on 3D Vision (3DV) (2019) 2
33. Germain, H., Bourmaud, G., Lepetit, V.: S2dnet: Learning image features for accurate sparse-to-dense matching. In: European Conference on Computer Vision (ECCV) (2020) 2
34. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels (2022) 10

35. Grabner, A., Roth, P.M., Lepetit, V.: 3d pose estimation and 3d model retrieval for objects in the wild. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3022–3031 (2018) 3
36. Grelsson, B., Robinson, A., Felsberg, M., Khan, F.S.: Gps-level accurate camera localization with horizonnet. Journal of Field Robotics **37**, 951–971 (2020) 3
37. Gumeli, C., Dai, A., Nießner, M.: Roca: Robust cad model retrieval and alignment from a single image. ArXiv **abs/2112.01988** (2021) 3
38. Hanocka, R., Metzer, G., Giryes, R., Cohen-Or, D.: Point2mesh: A self-prior for deformable meshes. arXiv preprint arXiv:2005.11084 (2020) 3
39. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. In: International Conference on Machine Learning. pp. 12633–12646. PMLR (2023) 14, 15
40. Hausler, S., Jacobson, A., Milford, M.: Multi-process fusion: Visual place recognition using multiple image processing methods. IEEE Robotics and Automation Letters **4**(2), 1924–1931 (2019) 1
41. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 14141–14152 (2021) 1, 2
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 8
43. Heng, L., Choi, B., Cui, Z., Geppert, M., Hu, S., Kuan, B., Liu, P., Nguyen, R., Yeo, Y.C., Geiger, A., et al.: Project autovision: Localization and 3d scene perception for an autonomous vehicle with a multi-camera system. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4695–4702. IEEE (2019) 2
44. Hess, G., Tonderski, A., Petersson, C., Åström, K., Svensson, L.: Lidarclip or: How i learned to talk to point clouds. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7438–7447 (2024) 3, 10, 11, 13, 14
45. Hodan, T.: Pose Estimation of Specific Rigid Objects. Ph.D. thesis, Czech Technical University in Prague (2021) 3
46. Hodan, T., Barath, D., Matas, J.: Epos: Estimating 6d pose of objects with symmetries. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11700–11709 (2020) 3
47. Hodan, T., Zabulis, X., Lourakis, M.I.A., Obdrzalek, S., Matas, J.: Detection and fine 3d pose estimation of texture-less objects in rgb-d images. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4421–4428 (2015) 3
48. Hu, S., Feng, M., Nguyen, R.H.M., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7258–7267 (2018) 3
49. Hu, S., Lee, G.H.: Image-based geolocalization using satellite imagery. International Journal of Computer Vision **128**, 1205–1219 (2019) 3
50. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. arXiv preprint arXiv:2201.13360 (2022) 3, 4, 7
51. Ibrahimi, S., van Noord, N., Alpherts, T., Worring, M.: Inside out visual place recognition. In: British Machine Vision Conference (2021) 1, 2
52. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: CVPR (2009) 2

53. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568 (2011) 3
54. Ji, X., Wei, J., Wang, Y., Shang, H., Kneip, L.: Cross-modal place recognition in image databases using event-based sensors. arXiv preprint arXiv:2307.01047 (2023) 3
55. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. IEEE Robotics and Automation Letters (2023) 1, 2, 11, 13, 14
56. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: CVPR (2017) 2
57. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: ICCV (2015) 2
58. Khaliq, A., Ehsan, S., Chen, Z., Milford, M., McDonald-Maier, K.: A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes. IEEE Transactions on Robotics **36**(2), 561–569 (2020) 1
59. Kim, H.J., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3251–3260 (2017) 1, 2
60. Kim, U.H., Park, J.M., Song, T.J., Kim, J.H.: 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. IEEE transactions on cybernetics **50**(12), 4921–4933 (2019) 3
61. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International journal of computer vision **38**, 199–218 (2000) 3
62. Labbe, Y., Carpentier, J., Aubry, M., Sivic, J.: Cosypose: Consistent multi-view multi-object 6d pose estimation. In: ECCV (2020) 3
63. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5374–5384 (2022) 11, 13, 14
64. Li, Y., Snavely, N., Huttenlocher, D.P., Fua, P.: Worldwide pose estimation using 3d point clouds. In: ECCV (2012) 2
65. Lim, H., Sinha, S.N., Cohen, M.F., Uyttendaele, M.: Real-time image-based 6-dof localization in large-scale environments. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1043–1050. IEEE (2012) 2
66. Lin, T.Y., Cui, Y., Belongie, S.J., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5007–5015 (2015) 3
67. Lin, Z., Zhang, Z., Wang, M., Shi, Y., Wu, X., Zheng, Y.: Multi-modal contrastive representation learning for entity alignment. arXiv preprint arXiv:2209.00891 (2022) 9
68. Liu, L., Li, H., Dai, Y.: Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In: ICCV (2017) 2
69. Liu, L., Li, H., Dai, Y.: Stochastic attraction-repulsion embedding for large scale image localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2570–2579 (2019) 1
70. Looper, S., Rodriguez-Puigvert, J., Siegwart, R., Cadena, C., Schmid, L.: 3d vsg: Long-term semantic scene change prediction through 3d variable scene graphs.

In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8179–8186. IEEE (May 2023) 3
71. Lynen, S., Zeisl, B., Aiger, D., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R., Sattler, T.: Large-scale, real-time visual–inertial localization revisited. The International Journal of Robotics Research **39**(9), 1061–1084 (2020) 2
72. Lynen, S., Zeisl, B., Aiger, D., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R., Sattler, T.: Large-scale, real-time visual–inertial localization revisited. The International Journal of Robotics Research (IJRR) **39**(9), 1061–1084 (2020) 2
73. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019) 3
74. Mihajlovic, M., Weder, S., Pollefeys, M., Oswald, M.R.: Deepsurfels: Learning online appearance fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14524–14535 (2021) 3
75. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 3
76. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Lens: Localization enhanced by nerf synthesis. In: CoRL (2021) 2
77. Murez, Z., Van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 414–431. Springer (2020) 3
78. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4531–4540 (2019) 3
79. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 7, 8, 14, 15
80. Panek, V., Kukelova, Z., Sattler, T.: Meshloc: Mesh-based visual localization. In: European Conference on Computer Vision. pp. 589–609. Springer (2022) 3
81. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019) 3
82. Peng, G., Yue, Y., Zhang, J., Wu, Z., Tang, X., Wang, D.: Semantic reinforced attention learning for visual place recognition. In: IEEE International Conference on Robotics and Automation. pp. 13415–13422. IEEE (2021) 2
83. Peng, G., Zhang, J., Li, H., Wang, D.: Attentional pyramid pooling of salient visual residuals for place recognition. In: IEEE International Conference on Computer Vision. pp. 885–894 (October 2021) 2
84. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 523–540. Springer (2020) 3
85. Pion, N., Humenberger, M., Csurka, G., Cabon, Y., Sattler, T.: Benchmarking image retrieval for visual localization. In: 2020 International Conference on 3D Vision (3DV). pp. 483–494. IEEE (2020) 2

86. Plotz, T., Roth, S.: Automatic registration of images to untextured geometry using average shading gradients. International Journal of Computer Vision **125**, 65–81 (2017) 3
87. Ponimatkin, G., Labbe, Y., Russell, B., Aubry, M., Sivic, J.: Focal length and object pose estimation via render and compare. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3825–3834 (June 2022) 3
88. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 3, 6
89. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision pp. 8748–8763 (2021) 10
90. Ramalingam, S., Bouaziz, S., Sturm, P.F., Brand, M.: Skyline2gps: Localization in urban canyons using omni-skylines. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3816–3823 (2010) 3
91. Ravichandran, Z., Peng, L., Hughes, N., Griffith, J., Carlone, L.: Hierarchical representations and explicit memory: Learning effective navigation policies on 3d scene graphs using graph neural networks. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 9272–9279. IEEE (May 2022) 3
92. Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., Gupta, A., Carlone, L.: Kimera: From slam to spatial perception with 3d dynamic scene graphs. The International Journal of Robotics Research **40**(12-14), 1510–1546 (2021) 3, 4
93. Rosinol, A., Gupta, A., Abate, M., Shi, J., Carlone, L.: 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. arXiv preprint arXiv:2002.06289 (2020) 3
94. Sarkar, S.D., Miksik, O., Pollefeys, M., Barath, D., Armeni, I.: Sgaligner: 3d scene alignment with scene graphs. International Conference on Computer Vision (2023) 3, 6, 7
95. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR (2019) 2
96. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks 2
97. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulò, S.R., Newcombe, R., Kontschieder, P., Balntas, V.: Orienternet: Visual localization in 2d public maps with neural matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21632–21642 (2023) 1, 3
98. Sarlin, P.E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., 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. pp. 3247–3257 (2021) 2
99. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. PAMI (2017) 2
100. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixe, L.: Understanding the limitations of cnn-based absolute camera pose regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3302–3312 (2019) 2

101. Savinov, N., Hane, C., Ladicky, L., Pollefeys, M.: Semantic 3d reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5460–5469 (2016) 3
102. Schönberger, J.L., Pollefeys, M., Geiger, A., Sattler, T.: Semantic visual localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6896–6906 (2018) 2
103. Sepulveda, G., Niebles, J., Soto, A.: A deep learning based behavioral approach to indoor autonomous navigation. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 4646–4653. IEEE (May 2018) 3
104. Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Accurate geo-registration by ground-to-aerial image matching. In: 3DV (2014) 3
105. Shubodh, S., Omama, M., Zaidi, H., Parihar, U.S., Krishna, M.: Lip-loc: Lidar image pretraining for cross-modal localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 948–957 (2024) 3, 10, 11, 13, 14
106. Sibbing, D., Sattler, T., Leibe, B., Kobbelt, L.: Sift-realistic rendering. In: 3DV (2013) 3
107. Steiger Mueller, M., Sattler, T., Pollefeys, M., Jutzi, B.: Image-to-image translation for enhanced feature matching, image retrieval and visual localization. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2019) 3
108. Stückler, J., Behnke, S.: Multi-resolution surfel maps for efficient dense 3d modeling and tracking. Journal of Visual Communication and Image Representation **25**(1), 137–147 (2014) 3
109. Svarm, L., Enqvist, O., Kahl, F., Oskarsson, M.: City-scale localization for cameras with known vertical direction. PAMI **39**(7), 1455–1461 (2017) 2
110. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation (2023) 7, 10, 11, 13, 14
111. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., et al.: Advances in neural rendering. In: Computer Graphics Forum. vol. 41, pp. 703–735. Wiley Online Library (2022) 3
112. Tomesek, J., Cadik, M., Brejcha, J.: Crosslocate: Cross-modal large-scale visual geo-localization in natural environments using rendered modalities. In: WACV (2022) 3
113. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 257–271 (2018) 1
114. Torii, A., Sivic, J., Okutomi, M., Pajdla, T.: Visual place recognition with repetitive structures. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(11), 2346–2359 (2015) 1
115. Torii, A., Taira, H., Sivic, J., Pollefeys, M., Okutomi, M., Pajdla, T., Sattler, T.: Are large-scale 3d models really necessary for accurate visual localization? IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 814–829 (2021) 1
116. Valentin, J., Nießner, M., Shotton, J., Fitzgibbon, A., Izadi, S., Torr, P.: Exploiting uncertainty in regression forests for accurate camera relocalization. In: CVPR (2015) 2

117. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018) 7
118. Viswanathan, A., Rodrigues Pires, B., Huber, D.F.: Vision based robot localization by ground to satellite matching in gps-denied situations. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 192–198 (2014) 3
119. Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using lstms for structured feature correlation. In: ICCV (2017) 2
120. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7658–7667 (2019) 8, 9, 11, 13, 14, 15
121. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020) 2, 3
122. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020) 11
123. Wang, S., Kannala, J., Barath, D.: DGC-GNN: Descriptor-free geometric-color graph neural network for 2d-3d matching. arXiv preprint arXiv:2306.12547 (2023) 3
124. Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (June 2020) 1, 2
125. Weder, S., Schonberger, J.L., Pollefeys, M., Oswald, M.R.: Neuralfusion: Online depth fusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3162–3172 (2021) 3
126. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3961–3969 (2015) 3
127. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7515–7525 (2021) 3, 7, 13
128. Ying, Z., Yuan, X., Yang, B., Song, Y., Xu, Q., Zhou, F., Sheng, W.: Rp-sg: Relation prediction in 3d scene graphs for unobserved objects localization. IEEE Robotics and Automation Letters (2023) 3
129. Zaffar, M., Garg, S., Milford, M., et al.: Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. IJCV pp. 1–39 (2021) 1
130. Zeisl, B., Sattler, T., Pollefeys, M.: Camera pose voting for large-scale image-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2704–2712 (2015) 2
131. Zhang, C., Yu, J., Song, Y., Cai, W.: Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9705–9715 (2021) 3
132. Zhang, S., Hao, A., Qin, H.: Knowledge-inspired 3d scene graph prediction in point cloud. Advances in Neural Information Processing Systems **34**, 18620–18632 (2021) 3

133. Zhang, W., Kosecka, J.: Image based localization in urban environments. In: International Symposium on 3D Data Processing, Visualization, and Transmission (2006) 2
134. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference pose generation for long-term visual localization via learned features and view synthesis. IJCV (2020) 3
135. Zheng, E., Wu, C.: Structure from motion using structure-less resection. In: The IEEE International Conference on Computer Vision (ICCV) (2015) 2
136. Zhou, Q., Agostinho, S., Ošep, A., Leal-Taixé, L.: Is geometry enough for matching in visual localization? In: European Conference on Computer Vision. pp. 407–425. Springer (2022) 1, 3