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

Yang Miao[1], Francis Engelmann[1,2], Olga Vysotska[1], Federico Tombari[2,3], Marc Pollefeys[1,4], and Dániel Béla Baráth[1]

[1]ETH Zurich [2]Google [3]TU Munich [4]Microsoft

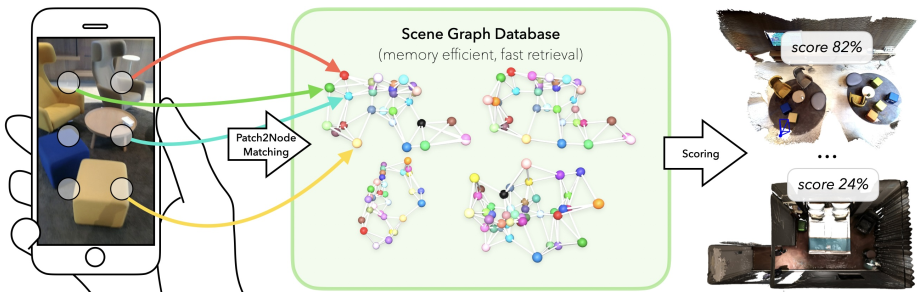

Fig. 1: SceneGraphLoc addresses the novel problem of localizing a query image in a database of 3D scenes represented as compact multi-modal 3D scene graphs.

Abstract. We introduce the task of localizing an input image within a multi-modal reference map represented by a collection of 3D scene graphs. These scene graphs comprise 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 these modalities, the proposed method SceneGraphLoc learns a fixed-sized embedding for each node (*i.e.*, representing object instances) 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 representation. With images, 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. Code and models are available at https://scenegraphloc.github.io.

Keywords: Coarse Localization · 3D Scene Graph · Multi-modality

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

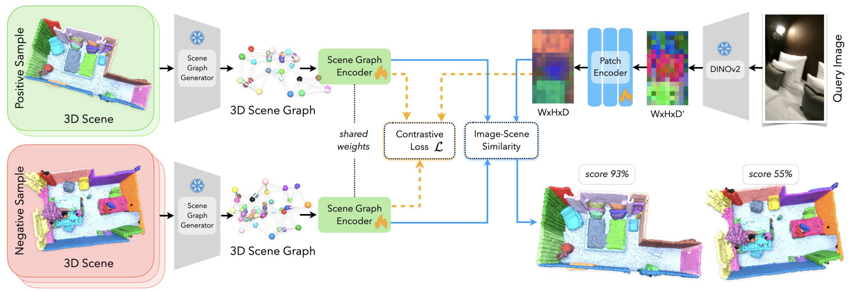

Fig. 2: Overview. The training phase is represented by orange arrows, while blue arrows denote the inference phase. During training, a query image and its associated 3D scene graph form a positive sample within a contrastive learning framework, where negative samples are generated by associating scene graphs of different scenes with the same query image. The objective is to learn the embeddings of both the graph and the image so that embeddings of the positive pair are drawn closer, whereas those of the negative pair are pushed apart. In the inference phase, the task involves assigning the correct scene graph to a given query image from a selection of multiple graphs, achieved by optimizing the cosine similarity between their embeddings.

level [2,9,22,28,39,40,50,58,69,116,117,127,132,143]. 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 coarse localization methods, such as AnyLoc [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 [100,142], 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 the 3D scene graph [3,124] 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 3D scene graphs 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 database of 3D scene graphs.

Even without incorporating images into the map, SceneGraphLoc largely outperforms other cross-modal methods on two large-scale, real-world indoor datasets. With images, 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. The method is visualized in Fig. 1.

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 mobile robots [42, 65, 139, 140], and augmented reality [16, 71]. The differentiation in localization techniques arises from their scene representation methods – be it through explicit 3D models [31, 32, 51, 64, 68, 72, 98, 99, 101, 102, 105, 112, 133, 135], sets of posed images [10, 88, 137, 141], or implicitly via neural network weights [6, 11–13, 17, 56, 57, 78, 118, 122] – and their approach to camera pose estimation, whether by 2D-3D [31,32,98,99,102,112,120,126,133] or 2D-2D [10, 141] matches, or through a composite of base poses [56, 57, 78, 88, 103, 122]. 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 [2,8,9,27,40,50,55,59,84,85,127] 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 [4,5,14,15,35,82,89,93, 107,109,110,115,138,142]. 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 [89] or skylines [93]. The second strategy emphasizes the identification and matching of local image features [4, 5, 14, 35, 107,109,110,115,126,138,142], 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 [14,82,115]. CAD and other models are also commonly used by object pose estimation [4, 30, 34, 36, 44–46, 62, 90]. Image to LiDAR localization [7, 26, 43, 108] 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 [47,48,66,100,107,121,129]. Other cross-modal techniques were also proposed to involve semantics [28, 29] and event cameras [53] 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, images, semantics, object attributes, and relationships.

Scene Representation, encapsulating various scene attributes has evolved significantly, yielding diverse surface representations from explicit forms (3D point clouds [75,91], meshes [37], surfels [111]) to implicit ones (occupancy [61], signed distance functions [20, 52]). The advent of neural representations has introduced novel means of encoding geometry [73, 83, 87, 128], appearance [76, 80], and semantics [74, 79]. A comprehensive review is provided by Tewari et al. [114]. The integration of directions/rays [77] and visibility encoding in surface reconstruction [104] 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 [96, 124] have further advanced 3D scene graph learning and structure.

The increasing interest in 3D scene graphs [3, 54, 60, 96, 124] underscores their potential as structured, rich descriptors for real-world scenes. Methods range from online incremental construction [49, 130] to offline generation from RGB-D imagery [3, 95, 124], and approaches for scene graph prediction [134, 136]. Their application spans embodied AI [94, 95, 106], task completion [1, 25], variability estimation [70], and SLAM [49,95]. Recent studies like [131] 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, [97] offers new perspectives on 3D scene alignment, employing node matching within overlapping scene graphs to facilitate precise 3D map alignment. Kabalar *et al.* [54] assume that the query image has been coarsely localized and leverages a scene graph to identify dynamic objects and for precise localization with image features. 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 [49, 95], each representing a group of object instances constituting a place like a room or building. Moreover, this task can also be straightforwardly redefined as subgraph selection. 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 $\mathcal{E}_i = \{(v^i_j, v^i_k) | v^i_j, v^i_k \in \mathcal{V}_i\}$ delineate the relationships between objects, such as "nearby", "standing on", and "attached to".

For each vertex v, we introduce $M \in \mathbb{N}^+$ map 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 $\mathcal{M}_j \in \{\texttt{position}, \texttt{orientation}, \texttt{pointcloud}, \texttt{semantic category}, \texttt{image}, \texttt{attribute}\}$. While this paper focuses on these modalities, the set is easily extendable by incorporating additional ones, such as `textual description` or `floor plan`.

Given $\mathcal{G}$ and an input query image I, the objective is to identify the scene graph $\mathcal{G}_i$ corresponding to the space depicted in image I. Note that while we focus on having an input image in this paper, the method can be modified to other modalities as well, *e.g.*, depth image. Formally, we aim to resolve 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 object set visible in image I, object $o_I \subseteq \{(x, y) \in I\}$ is a set of pixels in the image ($o_I \in \mathcal{O}_I$), $\texttt{visible} : \mathcal{G}_i \times \mathcal{O}_I \to \{0, 1\}$ indicates if an object is visible in a scene graph, $P(o_I \mid \mathcal{G}_i)$ is the probability of o_I stemming from $\mathcal{G}_i$.

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. 2. 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 normalized to $[0, 1]$.

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

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

Through matching, 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-graph similarity score enabling us to deduce whether image I corresponds to the space represented by $\mathcal{G}_i$. Finally, potential coarse locations of I are selected by maximizing the similarity score across the stored scene graphs as depicted in Eq 2.

3.1 Object Embeddings in the Scene Graph

This section aims to obtain an embedding for each node within graph $\mathcal{G}_i$, encapsulating information from *all* available modalities. Our method builds upon the method of Sarkar et al. [97], with enhancements to include the `image` modality and to distill a unified embedding from all modalities rather than merely concatenating separate embeddings as in [97].

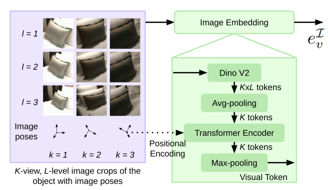

Fig. 3: The embedding of image modality $\mathcal{I}$ for each object. The image crops of a pillow are shown as an example.

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 [91] as the object encoder to extract the geometric feature $e_v^{\mathcal{P}}$ for every node. Furthermore, we integrate multi-level and multi-view visual embeddings to enrich the graph encoder with a more nuanced understanding of image information. The visual embedding pipeline is visualized in Fig. 3. For each node v denoting a 3D object, the top $K_{view} \in \mathbb{N}^+$ images with largest visibility of v is selected: $\{I_{\text{db},k}^v \mid k = [0, K)\} \subseteq \mathcal{I}_{\text{db}}$ from the database $\mathcal{I}_{\text{db}}$. We can define an ordering over these images such that $\phi(I_{\text{db},0}^v, v) \geq \phi(I_{\text{db},1}^v, v) \geq ... \geq \phi(I_{\text{db},K_{view}-1}^v, v)$, with the visibility function ϕ quantifying the extent of node v observed in each image through pixel count. Visibility check is implemented by projecting the 3D model of v to each posed image in $\mathcal{I}_{\text{db}}$, which are usually available when constructing the scene graph [49, 130]. Parameter $K_{view} = 10$ in all our experiments.

Drawing inspiration from OpenMask3D [113], for any given view I_{db}^v 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 [81] processes the image crops defined by $b_{v,l}$, extracting multi-level features $\{f_l \mid l \in [0, L)\}$. For each image $I_{\text{db},k}^v$, an average pooling operation aggregates these multi-level crop features into a singular feature vector $f = \text{avg pool}\{f_l \mid l \in [0, L)\}$. 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 $e_v^{\mathcal{I}}$, 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 to any other object in the scene. To encapsulate this structural information within $\mathcal{G}_i$, a Graph Attention Network (GAT) [119] is utilized, with the weight matrix constrained to a diagonal form to reduce computational demands and enhance model scalability. Following the method in [97], the structural embedding $e_v^{\mathcal{S}}$ is derived from the final layer of a two-layer GAT model.

Meta Embeddings. In addition to geometry and structure, the object attributes and the inter-object relationships are captured in two distinct embeddings, $e_v^{\mathcal{R}}$ and $e_v^{\mathcal{A}}$. The relationships an object maintains with others are conceptualized as a bag-of-words [19] feature vector, which is input to a feed-forward neural network, distilled in relational embedding $e_v^{\mathcal{R}}$. A similar approach is employed for the attributes associated with the objects, producing embedding $e_v^{\mathcal{A}}$.

Joint Embedding. Similar to [97], we concatenate uni-modal features to a compact representation for each object v. Contrary to [97], we encapsulate it in a multi-layer perceptron (MLP) to learn an embedding with size independent of the available modalities, fusing information from all. Embedding e_v is as:

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

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.

3.2 Object Embeddings in the Query Image

To solve the optimization problem 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, *e.g.*, Mask2Former [18]. However, we noticed in our experiments that such an approach is susceptible to inaccuracies and failures (over- and under-segmentation) in the segmentation, severely affecting the accuracy. Thus, we approach the 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 in q. We use Dino V2 [81] as a backbone to obtain patch-level features. These features are passed through a 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 [41] on the Dino V2 features and a 3-layer MLP to further map the patch feature to dimension D.

3.3 Contrastive Learning

We use contrastive learning to learn a joint embedding space for the scene graph nodes and image patches. To do so, we form query image and graph pairs $(I, \mathcal{G}_I)$. Real-world scenes are rarely static, *e.g.*, objects move or undergo non-rigid deformations and illumination changes [123]. 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. 4.

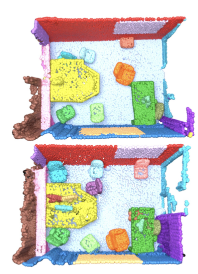

Fig. 4: The same scene at different time steps t.

For a query image I, a set of candidate 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 [123] 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 nodes of the target graph $\mathcal{G}_I$ and $\mathcal{V}_i$ is the nodes of other graphs $\mathcal{G}_i$. $N_v^{\mathcal{G}^t} = \{v_n | v_n \in \mathcal{V}_I^t \cup \mathcal{V}_1 \cup ... \cup \mathcal{V}_n\} \setminus \{v\}$, where $\mathcal{V}_I^t$ represent the nodes of the target graph $\mathcal{G}_I^t$. The static loss is defined as bi-directional N-pair loss [67] as follows:

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

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

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

where $f(e_q, e_v) = \exp(\frac{-\delta(e_q, e_v)}{\tau})$, $\delta(e_q, e_v)$ represents the inverse cosine similarity between embeddings e_q and e_v, and τ is a temperature parameter. Probability distribution $p(v, q, N_q^{\mathcal{I}}, N_v^{\mathcal{G}})$ is written similarly. The temporal loss term is defined the same as Eq.6 but with P_I^t and $N_v^{\mathcal{G}^t}$ as follows:

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

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:

$$s(\mathcal{G}_i, I) = \frac{1}{|\mathcal{Q}_I|} \sum_{q \in \mathcal{Q}_I} \left[1 - \delta(e_q, \text{NN}(q))\right], \tag{8}$$

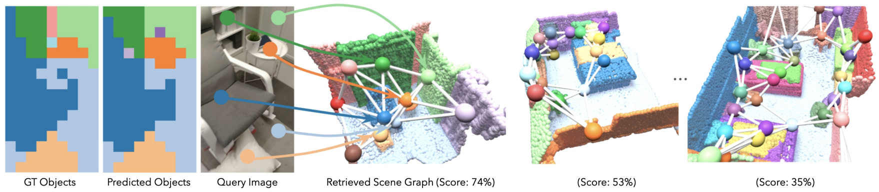

Fig. 5: Qualitative Result of object-association-based scene retrieval from the 3RScan dataset [123]. The two left images show the ground truth (left) and predicted (right) patch-to-node associations of the query image. The right part of the figure illustrates the candidate scene graphs sorted by the image-graph similarity.

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 s, which is assumed to map values to interval $[0, 1]$, facilitating the identification of the optimal scene graph $\mathcal{G}_{i^*}$ that maximizes $s(\mathcal{G}_i, I)$. Graph $\mathcal{G}_{i^*}$ is identified by iterating through all scene graphs and selecting the one with the highest similarity. This approach can be accelerated by using spatial partitioning techniques, such as kd-trees, to preprocess node embeddings within the map.

4 Experiments

Baselines. No existing methods directly tackle our task, but several recent advancements provide relevant baselines. LidarCLIP [43], designed for autonomous driving, transforms LiDAR clouds into global descriptors using the Single-stride Sparse Transformer [24] and matches them with CLIP image encoder embeddings [92]. Though not a perfect fit, it can be adapted for matching point clouds with CLIP embeddings of query images. LIP-Loc [108] employs a similar approach by converting LiDAR clouds into 2D range images for direct encoding and matching. Both methods were fine-tuned on our dataset for accurate comparison. We also explored open-vocabulary [23,33,86,113] object-retrieval-based baselines using OpenMask3D [113] and OpenSeg [33], with OpenMask3D assigning CLIP descriptors to 3D object instances from multiple observing images, and OpenSeg extracting pixel-level CLIP features from the query image. Despite their original purposes, both can be adapted for localization by matching query image descriptors with object instances. For comparison with state-of-the-art visual localization methods requiring large image databases, we included CVNet [63] and AnyLoc [55]. These methods offer advanced performance but demand significant storage for image descriptors and exhibit slower inference times.
Map Generation. The mapping stage is executed offline as a preprocessing step for visual localization approaches, requiring specific mapping operations for each method before proceeding to localization. For 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, this step involves converting point clouds into range images and computing the embeddings of the range images and for LidarCLIP, point clouds are directly encoded into global

Table 1: Retrieval recall on the test set of 3RScan dataset [123] (%; 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. R and R^t represent the recall in the static and temporal scenarios respectively. 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 [43]	✓	✗	✗	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 [108]	✓	✗	✗	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
OpenMask3D [113]	✓	✓	✗	–	–	–	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

descriptors. In OpenMask3D, the CLIP embeddings corresponding to each object are calculated and stored. For image-based methods like CVNet and AnyLoc, embeddings for all images in the database are precomputed.

Metrics. To evaluate the accuracy of a 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 evaluated method. Additionally, we will report the inference time and storage requirements.

Experiments on 3RScan. The 3RScan dataset [123] comprises 1335 annotated indoor scenes, representing 432 distinct rooms, with 1178 scenes (385 rooms) allocated for training and 157 (47 rooms) designated for validation. The training and validation sets include semantically annotated 3D point clouds for each scene, with some captured over extended periods (*e.g.*, several months) showcasing environmental changes. Annotations for graphs within the 3RScan dataset are provided in [125]. Due to the absence of such annotations in the test set, it was excluded from our experiments. Thus, we reorganized the original validation set, allocating 34 scenes (17 rooms) for validation and 123 scenes (30 rooms) for testing. For full reproducibility, we will publish this split.

During testing, we examine all 123 scenes of 30 rooms within the test set, selecting query images from each scene. The next step involves matching this image against N scenes (including the target) 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 settings $N = 50$ and $N = 10$. The latter setting emulates a scenario where a pre-selection 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.

Table 2: Retrieval recall in the temporal scenario on the test set of ScanNet dataset [21] (%; target scene ranked within the top 1, 3, and 5 of the retrieved list) and storage requirements (MB) for methods using point clouds ($\mathcal{P}$), images ($\mathcal{I}$), and other modalities ($\mathcal{S}$, $\mathcal{R}$) for map representation. The modality $\mathcal{A}$ is not available in the scene graphs predicted from [130]. 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 [43]	✓	✗	✗	19.4	47.5	67.6	4.7	14.8	22.2	5.9	15.0	21.9	**0.7**
LIP-Loc [108]	✓	✗	✗	10.3	27.0	43.6	1.9	6.0	8.1	1.8	3.1	4.0	1.7
OpenMask3D [113]	✓	✓	✗	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

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.

The results are in Table 1. Despite being retrained, LidarCLIP and LIP-Loc are inaccurate, 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 inferior to alternative methods. The inferiority is because both methods learn shared embedding between entire scenes and query images of totally different objects due to viewpoint differences. 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 significantly. 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. An example scene is shown in Fig. 5.

Experiments on ScanNet. To evaluate the generalization ability of our methods in real-world applications when scene graph annotations are not available, we conduct further experiments in the ScanNet dataset [21]. ScanNet encompasses 1613 monocular sequences of room-scale 3D scenes, offering 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 [130] 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

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 [123] and ScanNet [21] datasets.

Method	3DRScan [123]			ScanNet [21]			
	t_{e_q}	t^{10}_{retr}	t^{50}_{retr}	t_{e_q}	t^{10}_{retr}	t^{50}_{retr}	$t^{\text{all}}_{\text{retr}}$
LidarCLIP [43]	4.1	0.1	0.3	4.9	0.1	0.2	0.6
LIP-Loc [108]	2.7	0.1	0.2	4.1	0.1	0.2	0.5
OpenMask3D [113]	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]	658.4	354.6	1826.4	243.0	68.2	329.0	1451.1

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. Thus, each database image is selected from every 25 consecutive frames in the sequence, for image-based methods [55, 63]. For a fair comparison, all the methods only use the same selected images for training and evaluation.

For training, we use the official training set. 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 are allocated for testing. To ensure full reproducibility, we will make this split publicly available.

The results are in Table 2 for scenarios selecting the target room from subsets of 10, 50, and the entire set of 210 scenes. The performance of LIP-Loc shows a similar pattern to that observed on 3RScan, performing only slightly better than random selection. LidarCLIP shows a small improvement in accuracy. OpenMask3D 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 with methods that use extensive image collections as maps (such as CVNet and AnyLoc), SceneGraphLoc benefits from a database size *three* orders-of-magnitude smaller, highlighting its efficiency and effectiveness. We partly attribute this performance gap to the lack of object attributes in the dataset and the inaccurate instance segmentation predicted by [130]. More details can be found in the supp. mat.

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

Table 4: Ablation study performed on the val. split of 3RScan [123], 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 [81]						GCVit [38]					
$\mathcal{P}$	$\mathcal{I}$	$\mathcal{A}$	$\mathcal{S}$	$\mathcal{R}$	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
✓	✓				–	–	–	84.7	97.5	**99.6**	–	–	–	**73.2**	**89.7**	**95.9**
✓	✓	✓	✓	✓	–	–	–	**88.5**	**97.7**	**99.6**	–	–	–	72.1	88.8	95.7

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 Studies. To understand the impact of the integration of different modalities, we provide an ablation study on the localization performance with multiple combinations of modalities. We use the validation split of the 3DRScan dataset. Additional ablation studies can be found in supplementary material.

Table 4 displays the coarse localization performance of our method under the incorporation of distinct modalities ($\mathcal{P}$, $\mathcal{I}$, $\mathcal{A}$, $\mathcal{S}$, $\mathcal{R}$) within the pipeline. Additionally, the table reports results exploiting various backbones (Dino V2 [81] and GCVit [38]) for the extraction of image features. Similarly to the main experiments, we only show results on the temporal set when the map contains images. Employing Dino V2 for encoding both the query image and images within the map significantly enhances accuracy over using GCVit. The proposed method significantly outperforms both LidarCLIP and LIP-Loc (their results are in Table 1) even when using only point clouds in the map. Each additional modality contributes to the final accuracy, demonstrating that each plays a crucial role in enhancing the final performance when integrated into the pipeline.

5 Conclusion

In conclusion, we introduce SceneGraphLoc, a novel method for solving the novel problem of localizing an input image within a 3D scene graph-based multi-modal reference map. This approach outperforms existing cross-modal methods by a large margin. It achieves comparable accuracy to state-of-the-art image-based techniques with significantly lower storage requirements and faster processing speeds. Our experiments across the 3RScan and ScanNet datasets demonstrate the effectiveness of SceneGraphLoc, with the best performance achieved when integrating all proposed modalities. We believe that SceneGraphLoc is a step towards lightweight and efficient localization.

Acknowledgements.

We would like to thank our colleagues Ganlin Zhang, Sayan Deb Sarkar, and Cathrin Elich for their valuable advice and insightful discussions throughout the course of this research. Their contributions and suggestions greatly enhanced the quality and depth of this work. This work was partially funded by Design++ initiative of ETH Zurich, by the ETH RobotX research grant, the Hasler Stiftung Research Grant via the ETH Zurich Foundation, an ETH AI Center postdoctoral research fellowship and an ETH Zurich Career Seed Award.

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 (CoRL) (2022) 4
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2, 3
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: International Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2, 4
4. Aubry, M., Russell, B.C., Sivic, J.: Painting-to-3d model alignment via discriminative visual elements. ACM Transactions On Graphics (TOG) (2014) 3, 4
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: European Conference on Computer Vision (ECCV) (2018) 3
7. Bernreiter, L., Ott, L., Nieto, J., Siegwart, R., Cadena, C.: Spherical multi-modal place recognition for heterogeneous sensor systems. In: International Conference on Robotics and Automation (ICRA) (2021) 4
8. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3
9. Berton, G., Paolicelli, V., Masone, C., Caputo, B.: Adaptive-attentive geolocalization from few queries: A hybrid approach. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021) 2, 3
10. Bhayani, S., Sattler, T., Barath, D., Beliansky, P., Heikkilä, J., Kukelova, Z.: Calibrated and partially calibrated semi-generalized homographies. In: International Conference on Computer Vision (ICCV) (2021) 3
11. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC - Differentiable RANSAC for Camera Localization. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
12. Brachmann, E., Rother, C.: Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3

13. Brachmann, E., Rother, C.: Visual camera re-localization from RGB and RGB-D images using DSAC. Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2021) 3
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 (ECCV) (2020) 3, 4
15. Cadik, M., Sykora, D., Lee, S.: Automated outdoor depth-map generation and alignment. Computers & Graphics (2018) 3
16. Castle, R., Klein, G., Murray, D.W.: Video-rate localization in multiple maps for wearable augmented reality. In: IEEE International Symposium on Wearable Computers (2008) 3
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: International Conference on 3D Vision (3DV) (2019) 3
18. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 8
19. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: European Conference on Computer Vision (ECCV) Workshops (2004) 7
20. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Annual conference on Computer graphics and interactive techniques (1996) 4
21. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 12, 13
22. 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: International Conference on Computer Vision (ICCV) (2019) 2
23. Engelmann, F., Manhardt, F., Niemeyer, M., Tateno, K., Tombari, F.: OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. In: International Conference on Learning Representations (ICLR) (2024) 10
24. 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. International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 10
25. Gadre, S.Y., Ehsani, K., Song, S., Mottaghi, R.: Continuous scene representations for embodied ai. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4
26. 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: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2023) 4
27. Garg, S., Fischer, T., Milford, M.: Where is your place, visual place recognition? International Joint Conference on Artificial Intelligence (IJCAI) (2021) 3
28. Garg, S., Suenderhauf, N., Milford, M.: Semantic–geometric visual place recognition: a new perspective for reconciling opposing views. International Journal on Robotics Research (IJRR) (2019) 2, 4

29. 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 (2020) 4
30. Georgakis, G., Karanam, S., Wu, Z., Kosecka, J.: Learning local rgb-to-cad correspondences for object pose estimation. In: International Conference on Computer Vision (ICCV) (2019) 4
31. Germain, H., Bourmaud, G., Lepetit, V.: Sparse-to-Dense Hypercolumn Matching for Long-Term Visual Localization. In: International Conference on 3D Vision (3DV) (2019) 3
32. Germain, H., Bourmaud, G., Lepetit, V.: S2DNet: Learning Image Features for Accurate Sparse-to-Dense Matching. In: European Conference on Computer Vision (ECCV) (2020) 3
33. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In: European Conference on Computer Vision (ECCV) (2022) 10
34. Grabner, A., Roth, P.M., Lepetit, V.: 3D Pose Estimation and 3D Model Retrieval for Objects in the Wild. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
35. Grelsson, B., Robinson, A., Felsberg, M., Khan, F.S.: GPS-level accurate camera localization with HorizonNet. Journal of Field Robotics (2020) 3
36. Gumeli, C., Dai, A., Nießner, M.: ROCA: Robust CAD Model Retrieval and Alignment from a Single Image. arXiv preprint arXiv:2112.01988 (2021) 4
37. Hanocka, R., Metzer, G., Giryes, R., Cohen-Or, D.: Point2mesh: A self-prior for deformable meshes. arXiv preprint arXiv:2005.11084 (2020) 4
38. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. In: International Conference on Machine Learning (ICML) (2023) 14
39. Hausler, S., Jacobson, A., Milford, M.: Multi-process fusion: Visual place recognition using multiple image processing methods. IEEE Robotics and Automation Letters (RA-L) (2019) 2
40. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 3
41. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 8
42. 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 Cehicle with a Multi-Camera System. In: International Conference on Robotics and Automation (ICRA) (2019) 3
43. Hess, G., Tonderski, A., Petersson, C., Åström, K., Svensson, L.: Lidarclip or: How i learned to talk to point clouds. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024) 4, 10, 11, 12, 13
44. Hodan, T.: Pose Estimation of Specific Rigid Objects. Ph.D. thesis (2021) 4
45. Hodan, T., Barath, D., Matas, J.: EPOS: Estimating 6D Pose of Objects With Symmetries. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
46. 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: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2015) 4

47. 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: International Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
48. Hu, S., Lee, G.H.: Image-Based GeoLocalization Using Satellite Imagery. International Journal on Computer Vision (IJCV) (2019) 4
49. 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) 4, 5, 7
50. Ibrahimi, S., van Noord, N., Alpherts, T., Worring, M.: Inside out visual place recognition. In: British Machine Vision Conference (2021) 2, 3
51. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From Structure-from-Motion Point Clouds to Fast Location Recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 3
52. 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: ACM symposium on User interface software and technology (2011) 4
53. 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) 4
54. Kabalar, J., Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Towards long-term retrieval-based visual localization in indoor environments with changes. IEEE Robotics and Automation Letters (2023) 4
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 (RA-L) (2023) 2, 3, 10, 11, 12, 13
56. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
57. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: International Conference on Computer Vision (ICCV) (2015) 3
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 (T-RO) (2020) 2
59. Kim, H.J., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geolocalization. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
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 (2019) 4
61. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International Journal on Computer Vision (IJCV) (2000) 4
62. Labbe, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: Consistent multi-view multi-object 6D pose estimation. In: European Conference on Computer Vision (ECCV) (2020) 4
63. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 10, 11, 12, 13

64. Li, Y., Snavely, N., Huttenlocher, D.P., Fua, P.: Worldwide Pose Estimation Using 3D Point Clouds. In: European Conference on Computer Vision (ECCV) (2012) 3
65. Lim, H., Sinha, S.N., Cohen, M.F., Uyttendaele, M.: Real-time Image-based 6-DoF Localization in Large-scale Environments. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 3
66. Lin, T.Y., Cui, Y., Belongie, S.J., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 4
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: International Conference on Computer Vision (ICCV) (2017) 3
69. Liu, L., Li, H., Dai, Y.: Stochastic attraction-repulsion embedding for large scale image localization. In: International Conference on Computer Vision (ICCV) (2019) 2
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: International Conference on Robotics and Automation (ICRA) (2023) 4
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. International Journal on Robotics Research (IJRR) (2020) 3
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. International Journal on Robotics Research (IJRR) (2020) 3
73. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
74. Miao, Y., Armeni, I., Pollefeys, M., Barath, D.: Volumetric Semantically Consistent 3D Panoptic Mapping. arXiv preprint arXiv:2309.14737 (2024) 4
75. Miao, Y., Li, C., Li, Z., Yang, Y., Yu, X.: A novel algorithm of ship structure modeling and target identification based on point cloud for automation in bulk cargo terminals. Measurement and Control (2021) 4
76. Mihajlovic, M., Weder, S., Pollefeys, M., Oswald, M.R.: DeepSurfels: Learning online appearance fusion. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
77. 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 (2021) 4
78. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: LENS: Localization enhanced by neRF synthesis. In: Conference on Robot Learning (CoRL) (2021) 3
79. Murez, Z., Van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images. In: European Conference on Computer Vision (ECCV) (2020) 4
80. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: International Conference on Computer Vision (ICCV) (2019) 4

81. 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
82. Panek, V., Kukelova, Z., Sattler, T.: Meshloc: Mesh-based visual localization. In: European Conference on Computer Vision (ECCV) (2022) 3, 4
83. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
84. Peng, G., Yue, Y., Zhang, J., Wu, Z., Tang, X., Wang, D.: Semantic reinforced attention learning for visual place recognition. In: International Conference on Robotics and Automation (ICRA) (2021) 3
85. Peng, G., Zhang, J., Li, H., Wang, D.: Attentional pyramid pooling of salient visual residuals for place recognition. In: International Conference on Computer Vision (ICCV) (2021) 3
86. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: OpenScene: 3D Scene Understanding with Open Vocabularies. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 10
87. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: European Conference on Computer Vision (ECCV) (2020) 4
88. Pion, N., Humenberger, M., Csurka, G., Cabon, Y., Sattler, T.: Benchmarking image retrieval for visual localization. In: International Conference on 3D Vision (3DV) (2020) 3
89. Plotz, T., Roth, S.: Automatic Registration of Images to Untextured Geometry Using Average Shading Gradients. International Journal on Computer Vision (IJCV) (2017) 3
90. Ponimatkin, G., Labbe, Y., Russell, B., Aubry, M., Sivic, J.: Focal length and object pose estimation via render and compare. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4
91. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 4, 6
92. 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. International Conference on Machine Learning (ICML) (2021) 10
93. Ramalingam, S., Bouaziz, S., Sturm, P.F., Brand, M.: SKYLINE2GPS: Localization in urban canyons using omni-skylines. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2010) 3
94. 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: International Conference on Robotics and Automation (ICRA) (2022) 4
95. 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. International Journal on Robotics Research (IJRR) (2021) 4, 5
96. 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) 4

97. Sarkar, S.D., Miksik, O., Pollefeys, M., Barath, D., Armeni, I.: SGAligner: 3D Scene Alignment with Scene Graphs. International Conference on Computer Vision (ICCV) (2023) 4, 6, 7
98. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
99. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3
100. 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: International Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 4
101. 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: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3
102. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2017) 3
103. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixe, L.: Understanding the limitations of cnn-based absolute camera pose regression. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
104. Savinov, N., Hane, C., Ladicky, L., Pollefeys, M.: Semantic 3d reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4
105. Schönberger, J.L., Pollefeys, M., Geiger, A., Sattler, T.: Semantic visual localization. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3
106. Sepulveda, G., Niebles, J., Soto, A.: A deep learning based behavioral approach to indoor autonomous navigation. In: International Conference on Robotics and Automation (ICRA) (2018) 4
107. Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Accurate geo-registration by ground-to-aerial image matching. In: International Conference on 3D Vision (3DV) (2014) 3, 4
108. Shubodh, S., Omama, M., Zaidi, H., Parihar, U.S., Krishna, M.: Lip-loc: Lidar image pretraining for cross-modal localization. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024) 4, 10, 11, 12, 13
109. Sibbing, D., Sattler, T., Leibe, B., Kobbelt, L.: SIFT-Realistic Rendering. In: International Conference on 3D Vision (3DV) (2013) 3
110. 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
111. Stückler, J., Behnke, S.: Multi-resolution surfel maps for efficient dense 3D modeling and tracking. Journal of Visual Communication and Image Representation (2014) 4
112. Svarm, L., Enqvist, O., Kahl, F., Oskarsson, M.: City-Scale Localization for Cameras with Known Vertical Direction. Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2017) 3

113. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In: International Conference on Neural Information Processing Systems (NeurIPS) (2023) 7, 10, 11, 12, 13
114. 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 (2022) 4
115. Tomesek, J., Cadik, M., Brejcha, J.: CrossLocate: Cross-modal Large-scale Visual Geo-Localization in Natural Environments using Rendered Modalities. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022) 3, 4
116. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2018) 2
117. 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? Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2021) 2
118. Valentin, J., Nießner, M., Shotton, J., Fitzgibbon, A., Izadi, S., Torr, P.: Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3
119. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. In: International Conference on Learning Representations (ICLR) (2018) 7
120. Ventura, J., Kukelova, Z., Sattler, T., Baráth, D.: Absolute Pose from One or Two Scaled and Oriented Features. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
121. Viswanathan, A., Rodrigues Pires, B., Huber, D.F.: Vision based robot localization by ground to satellite matching in GPS-denied situations. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2014) 4
122. Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-Based Localization Using LSTMs for Structured Feature Correlation. In: International Conference on Computer Vision (ICCV) (2017) 3
123. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: International Conference on Computer Vision (ICCV) (2019) 8, 9, 10, 11, 13, 14
124. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 4
125. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 11
126. Wang, S., Kannala, J., Barath, D.: DGC-GNN: Descriptor-free Geometric-Color Graph Neural Network for 2D-3D Matching. International Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
127. Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 3

128. Weder, S., Schonberger, J.L., Pollefeys, M., Oswald, M.R.: Neuralfusion: Online depth fusion in latent space. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
129. Workman, S., Souvenir, R., Jacobs, N.: Wide-Area Image Geolocalization with Aerial Reference Imagery. In: International Conference on Computer Vision (ICCV) (2015) 4
130. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4, 7, 12, 13
131. 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 (RA-L) (2023) 4
132. Zaffar, M., Garg, S., Milford, M., et al.: Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. International Journal on Computer Vision (IJCV) (2021) 2
133. Zeisl, B., Sattler, T., Pollefeys, M.: Camera pose voting for large-scale image-based localization. In: International Conference on Computer Vision (ICCV) (2015) 3
134. Zhang, C., Yu, J., Song, Y., Cai, W.: Exploiting edge-oriented reasoning for 3D point-based scene graph analysis. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
135. Zhang, G., Larsson, V., Barath, D.: Revisiting Rotation Averaging: Uncertainties and Robust Losses. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
136. Zhang, S., Hao, A., Qin, H.: Knowledge-inspired 3D scene graph prediction in point cloud. International Conference on Neural Information Processing Systems (NeurIPS) (2021) 4
137. Zhang, W., Kosecka, J.: Image based localization in urban environments. In: International Symposium on 3D Data Processing, Visualization, and Transmission (2006) 3
138. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. International Journal on Computer Vision (IJCV) (2020) 3
139. Zhao, L., Gatsis, K., Papachristodoulou, A.: Stable and Safe Reinforcement Learning via a Barrier-Lyapunov Actor-Critic Approach. In: IEEE Conference on Decision and Control (CDC) (2023) 3
140. Zhao, L., Miao, K., Gatsis, K., Papachristodoulou, A.: Stable and safe human-aligned reinforcement learning through neural ordinary differential equations. arXiv preprint arXiv:2401.13148 (2024) 3
141. Zheng, E., Wu, C.: Structure From Motion Using Structure-Less Resection. In: International Conference on Computer Vision (ICCV) (2015) 3
142. Zhou, Q., Agostinho, S., Ošep, A., Leal-Taixé, L.: Is Geometry Enough for Matching in Visual Localization? In: European Conference on Computer Vision (ECCV) (2022) 2, 3
143. Zurbrügg, R., Liu, Y., Engelmann, F., Kumar, S., Hutter, M., Patil, V., Yu, F.: ICGNet: A Unified Approach for Instance-Centric Grasping. In: International Conference on Robotics and Automation (ICRA) (2024) 2