IHF-RELOC: AN INSTANCE-AWARE HIERARCHICAL FRAMEWORK FOR 2D–3D VISUAL RELOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing visual relocalization methods typically follow a coarse-to-fine pipeline, establishing patch-level correspondences between a query image and a pre-built 3D map before refining them into point-to-pixel matches. However, in scenes with weak textures and occlusions, visually similar patches produce ambiguous responses. While recent methods exchange contextual information between image and point features to alleviate this ambiguity, they overlook the topological consistency among neighboring correspondences. In this paper, we propose IHF-ReLoc, an instance-aware hierarchical framework for 2D–3D visual relocalization, which progressively refines matching from instance-level associations to fine-grained point-to-patch correspondences. Specifically, IHF-ReLoc consists of two core modules, namely Scene Graph Matching (SGM) and Candidate Graph Reasoner (CGR). SGM performs matching over 2D and 3D instance graphs and exploits instance topology to enhance feature discriminability, thereby reducing cross-modal ambiguity. The CGR models candidate point-to-patch correspondences as graph nodes and their local geometric relations as edges. By propagating geometric information among neighboring candidates, CGR suppresses visually similar but incorrect matches. Experiments show that IHF-ReLoc achieves state-of-the-art performance and remains robust to temporal scene changes without relying on pose priors obtained through image retrieval.

1 INTRODUCTION

2D–3D visual relocalization, recovering the camera pose of a query image on a pre-built 3D map, is a fundamental capability for autonomous robot navigation and long-term operation (Taira et al., 2018; Li et al., 2023; Ye et al., 2026). The recovery requires direct correspondences between 2D observations and a point cloud map. However, a substantial modality gap exists between images and point clouds. This gap is further widened by weak textures, repetitive structures, and viewpoint changes, all of which hinder reliable matching. Beyond these perceptual challenges, changes in object placement and presence introduce additional discrepancies between the query observation and a map constructed at an earlier time (Wald et al., 2019; 2020).

Existing 2D–3D localization methods typically follow a coarse-to-fine strategy, identifying candidates at the region or patch level before refining them into accurate pixel-to-point correspondences (Kim et al., 2023; Li et al., 2023; Cheng et al., 2025). However, candidate selection heavily relies on cross-modal visual similarity. This reliance gives rise to high-confidence mismatches that then propagate into subsequent refinement stages.

These limitations reveal two key challenges in accuracy and robustness of 2D–3D matching. The first is how to reduce local appearance ambiguity and outliers caused by occlusion. Existing methods (Ye et al., 2026; Kim et al., 2023) organize coarse matching using planar regions or point cloud submaps. Such partitioning granularity may lead to semantic ambiguity in scenes with repetitive structures. In contrast, instances encode scene semantics and support inter-instance occlusion modeling, providing semantic and visibility constraints that reduce cross-modal ambiguity and interference from occluded map points. The second challenge is how to guide each 3D point toward the correct image patch. Graph-based methods enhance the consistency of the correct correspondence features and prune false correspondences (Bie et al., 2025). However, projection distortion changes the topology of local point sets. Moreover, topological information is propagated within each modal-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

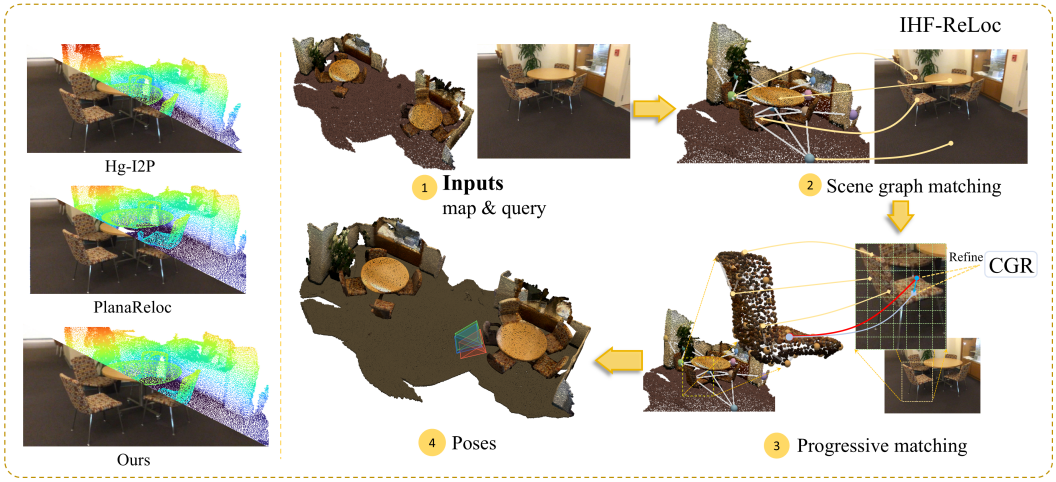


Figure 1: Overview of the coarse-to-fine matching and pose estimation pipeline of IHF-ReLoc. (1) A pre-built 3D map and a query image are provided as input. (2) Coarse matching establishes instance-level associations and identifies candidate regions. (3) Progressive matching searches for point-to-patch correspondences over increasingly finer patch grids and uses the Candidate Graph Reasoner to suppress incorrect matches. (4) The camera pose is estimated from the resulting high-confidence 2D–3D correspondences. Qualitative localization results of different methods are shown on the left.

ity before being transferred across modalities, making it difficult to directly and effectively constrain individual correspondences. By extracting the topology among projected points in the image plane and directly modeling the relations between point-to-patch candidates, topological constraints can directly guide candidate matching, improving both efficiency and accuracy.

Motivated by these observations, we propose IHF-ReLoc, an instance-aware hierarchical framework for 2D–3D visual relocalization. IHF-ReLoc consists of two core modules, namely Scene Graph Matching (SGM) and Candidate Graph Reasoner (CGR). As shown in Figure 1, SGM aggregates scene topology over 2D and 3D instance graphs to improve the discriminability of instance features. The derived instance-level associations constrain the search space for subsequent point-to-patch matching and filter out map points occluded by other instances. Graph attention first produces the top-K patch candidates for each 3D point, preserving potentially correct correspondences. CGR then treats each candidate match as a graph node, while the topology among projected 3D points is encoded into the graph edges. This mechanism allows neighborhood geometry to directly guide candidate selection, suppressing matches with high visual similarity but inconsistent topology. Progressive updates of the pose and projection geometry reduce the effect of initial pose errors and provide more reliable geometric priors for subsequent refinement. Experiments show state-of-the-art performance and robustness to temporal scene changes without pose priors from image retrieval.

The main contributions can be summarized as follows:

- We propose IHF-ReLoc, an instance-aware hierarchical framework for 2D–3D visual relocalization, which progressively refines matching from instance-level associations to fine-grained point-to-patch correspondences.
- We introduce Scene Graph Matching to reduce cross-modal ambiguity and occlusion interference. SGM performs matching over 2D and 3D instance graphs, using instance topology to enhance feature discriminability and instance-level visibility to filter occluded map points.
- We propose the Candidate Graph Reasoner to improve the reliability of point-to-patch matching. CGR models candidate correspondences as graph nodes and their local geometric relations as edges, allowing geometric consistency to suppress visually similar but incorrect matches.

2 Related Work

Image-retrieval methods rank posed reference images with global descriptors to identify likely locations for subsequent localization. Recent approaches improve descriptor aggregation and robustness to appearance ambiguity in challenging environments. SALAD combines DINOv2 features with optimal-transport aggregation (Izquierdo & Civera, 2024), while SemVPR incorporates multimodal semantic knowledge to learn compact, discriminative descriptors (Zhang et al., 2025a). EfficientVPR adapts pretrained features through scene-aware prompting (Tang et al., 2026), and A2GC introduces geometric constraints into descriptor aggregation (Li & Shang, 2026). Although these methods improve retrieval accuracy and efficiency, their localization performance depends on the coverage and known poses of reference images. IHF-ReLoc instead matches directly against a 3D point cloud map without retrieval-derived pose priors.

Scene Coordinate Regression (SCR) predicts the 3D scene coordinate corresponding to each query pixel and estimates the camera pose from these predictions, avoiding explicit feature matching and per-point 3D descriptors. NeuMap combines a scene-agnostic Transformer with scene-specific latent codes (Tang et al., 2023), while ACE pairs a shared feature encoder with a lightweight scene-specific prediction head to reduce mapping time (Brachmann et al., 2023). GLACE improves large-scale SCR through global–local encoding and co-visibility-aware feature diffusion (Wang et al., 2024a). ACE-G separates a generic coordinate regressor from the scene-specific map representation and uses cross-scene query pre-training to improve generalization to unseen viewpoints (Bruns et al., 2025). Despite compact map representations and efficient inference, SCR still depends strongly on the learned scene representation and remains sensitive to large viewpoint shifts and long-term scene changes.

Feature-matching methods explicitly establish 2D–3D correspondences between a query image and a 3D map and recover the camera pose through geometric estimation. Early learning-based methods, including 2D3D-MatchNet, LCD, and P2-Net, learn shared descriptors for pixels and points, but the cross-modal gap still causes ambiguous matches in weakly textured or repetitive regions (Feng et al., 2019; Pham et al., 2020; Wang et al., 2021). Recent patch- and region-based methods progressively refine coarse matches into pixel-to-point correspondences, whereas plane- and instance-based methods incorporate geometry, semantics, and spatial relations to constrain matching (Li et al., 2023; Kang et al., 2024; Cheng et al., 2025; Miao et al., 2024; Ye et al., 2026).

Another line of work filters matches using geometric relations among correspondences. GoMatch performs descriptor-free matching with directional geometry (Zhou et al., 2022), while DGC-GNN and A2-GNN encode distance, angular, and neighborhood consistency with graph neural networks (Wang et al., 2024c; Zhang et al., 2025b). GraphI2P selects reliable correspondences through graph-based consistency, and Hg-I2P constructs a heterogeneous graph over segmented 2D and 3D regions for cross-modal interaction and correspondence pruning (Bie et al., 2025; An et al., 2026). These methods nevertheless reason primarily over the original 2D or 3D elements, so topology constrains cross-modal correspondences only indirectly. In contrast, IHF-ReLoc treats point-to-patch candidate pairs as graph nodes, enabling their geometric consistency to be modeled and propagated explicitly.

3 IHF-ReLoc Method

3.1 Overview

Given a query image I^q, intrinsics $\mathbf{K}$, and a point cloud map $\mathcal{M} = \{\mathbf{P}_i\}_{i=1}^{N}$, IHF-ReLoc establishes 2D–3D correspondences and estimates the world-to-camera transformation $\mathbf{T}_{cw} = [\mathbf{R} \mid \mathbf{t}]$.

The framework has three components. Feature Embedding and Graph Construction (Section 3.2) extracts instance and multi-scale local features and builds graphs at both granularities. Scene Graph Matching and Progressive Matching (Section 3.3) retrieve 2D–3D instance pairs and refine point-to-patch correspondences within them. Graph Reasoning and Pose Refinement (Section 3.4) enforce neighborhood geometry among candidate matches and pass each stage-wise pose estimate to the next scale.

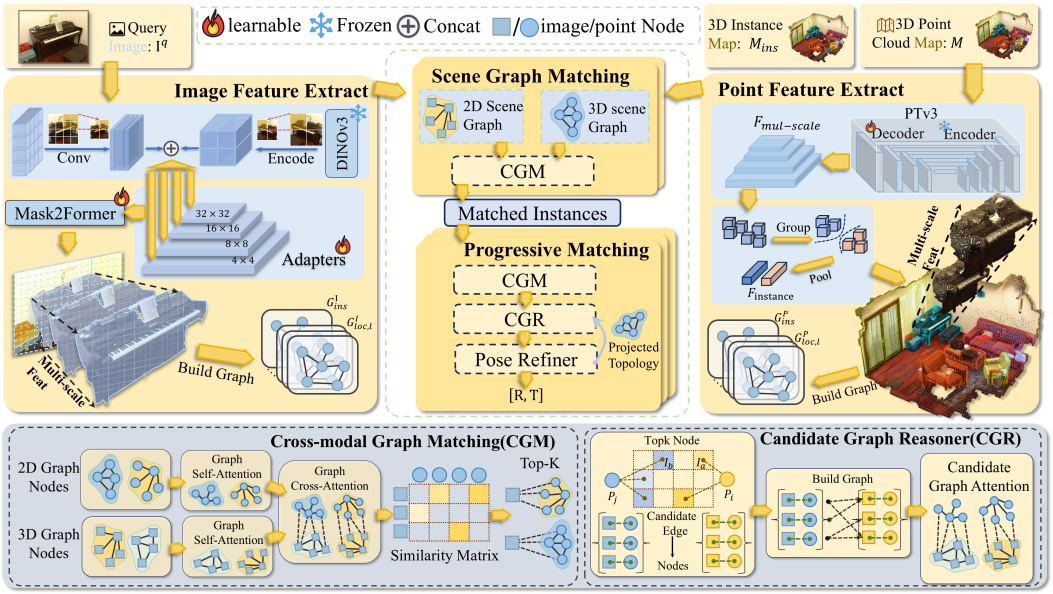


Figure 2: Overview of the IHF-ReLoc framework. The image and point-cloud branches extract instance-level representations and multi-scale local features, which are used to construct instance and local graphs. Scene Graph Matching first retrieves candidate 3D instances for each 2D instance and verifies them through local point-to-patch matching. Progressive Matching then refines the retained correspondences across multiple scales using Cross-modal Graph Matching, the Candidate Graph Reasoner, and pose refinement. The bottom panel details the two graph modules: CGM combines intra-modal graph attention with bidirectional cross-attention to generate top-K candidates, while CGR treats each candidate correspondence as a graph node and uses the topology of reprojected 3D points to refine matching scores.

3.2 Feature Embedding and Graph Construction

Image Feature Extraction. The image branch adopts DINOv3 ViT-B/16 as its visual backbone (Siméoni et al., 2025) and extracts tokens from multiple intermediate layers to obtain semantic representations with large receptive fields. Since the spatial resolution of DINOv3 features is insufficient for precise localization, as shown in Figure 2, an adapter fuses a CNN feature pyramid extracted from the original image to produce multi-scale features $\mathbf{F}_l^{2d}$, where l indexes the feature scale. A Mask2Former-based segmentation head (Cheng et al., 2022) sharing the image feature pyramid predicts 2D instance masks. Mask-wise pooling then produces the 2D instance feature $\mathbf{F}_{\text{ins}}^{2d}$. Thus, a single encoding pass provides both the instance representations for the scene graph and the multi-scale patch representations for Progressive Matching.

Point Feature Extraction. The point-cloud branch uses Point Transformer V3 (PTv3) (Wu et al., 2024) as its encoder and retains the outputs of multiple decoder layers, $\mathbf{F}_l^{3d}$, as multi-scale point features. Deeper features provide stable semantic context through larger spatial receptive fields, whereas shallower features preserve finer local geometry for Progressive Matching. Instance-wise aggregation based on the point cloud instance labels combines the features of points belonging to the same object into the 3D instance representation $\mathbf{F}_{\text{ins}}^{3d}$. Point-wise features from each layer are then used for point-to-patch matching at the corresponding scale.

Graph Construction. IHF-ReLoc builds KNN graphs within each modality by connecting every node to its k nearest neighbors according to inter-node distances, enhancing structural awareness across the image and point-cloud representations while reducing the computational cost of global interactions. The nodes encode the appearance and shape information of objects, while the edges characterize the relative directions, distances, and scale relationships between nodes. This yields an instance graph $\mathcal{G}_{\text{ins}}^D$ and multi-scale local graphs $\mathcal{G}_{\text{loc},l}^D$, where the local graphs are constructed only within the corresponding instances to capture finer-grained local structures.

3.3 Scene Graph Matching and Progressive Matching

Scene Graph Matching consists of instance retrieval and local verification. The 2D and 3D instance graphs are first fed into the Cross-modal Graph Matching (CGM) module, which computes the instance similarity score $S_{\text{inst}}(x, y)$ using cross-modally enhanced features, and retains the top-scoring 3D instances x for each 2D instance y. Keeping multiple candidates alleviates ambiguities caused by similar objects. The instance context is then injected into the nodes of the 2D graph $\mathcal{G}_{\text{ins}}^{2d}$ and 3D graph $\mathcal{G}_{\text{ins}}^{3d}$ for each candidate pair, and the resulting local graphs are fed into CGM to compute point-to-patch similarities. The similarities of valid points are aggregated into a local verification score $S_{\text{loc}}(x, y)$, which measures whether the candidate objects form stable correspondences in their local structures. The instance score first restricts the candidate set, while the local verification score selects the final instance match within the retained candidates, filtering out semantically similar objects with inconsistent internal structures before refinement.

Progressive Matching sequentially applies Cross-modal Graph Matching, the Candidate Graph Reasoner, and the Pose Refiner across multiple scales. CGM produces point-to-patch matching scores within the current candidate regions, the Candidate Graph Reasoner resolves local ambiguities using the relative projection relations between neighboring points, and the Pose Refiner estimates a pose at each stage from high-confidence correspondences and provides the projected locations and visibility information required at the next scale. For each 3D point, only high-scoring patches from the parent stage and their child regions are propagated to the finer scale, while multiple candidates are retained to reduce erroneous pruning caused by early mismatches. As the patches are progressively refined from 32×32 during Scene Graph Matching to 16×16 and 4×4, the search region contracts, ultimately yielding point-to-patch correspondences with continuous 2D coordinates and confidence scores.

3.4 Graph Reasoning and Pose Refinement

3.4.1 Cross-modal Graph Matching

Instance matching and point-to-patch matching share the same CGM module. As shown in Figure 2, it first aggregates neighborhood information within the 2D and 3D graphs through edge-conditioned self-attention, allowing each node representation to capture both its own features and local structure. Bidirectional cross-attention then exchanges context between the two modalities, and matching scores are computed from normalized feature similarities. Across levels, only the node features, edge features, and candidate masks change, while the interaction procedure remains the same. This design provides a unified computational framework for instance-level scene association and point-level local matching.

3.4.2 Candidate Graph Reasoner

Candidate Nodes. CGR treats each correspondence hypothesis between a 3D point and a candidate patch as a graph node. Based on the CGM scores, the model retains the top-K_c candidate patches for each 3D point and denotes a candidate node as $v = (\mathbf{P}_i, I_a)$, where $\mathbf{P}_i$ and I_a are the associated 3D point and patch, respectively. The candidate node feature $\mathbf{F}_v^{\text{node}}$ jointly encodes the corresponding point and patch features.

Topological Edges. The candidate graph inherits the connectivity of the 3D point graph. Specifically, an edge is created between candidate nodes v and u when their associated 3D points are neighbors. The edge feature $\mathbf{E}_{vu}$ consists of the relative displacement $\Delta\bar{\mathbf{x}}_{ij}^{P}$ between the projected 3D points $\bar{P}_i$ and $\bar{P}_j$, and the relative displacement $\Delta\mathbf{x}_{ab}^{I}$ between the centers of the corresponding image patches I_a and I_b. Both displacements encode direction and distance, and their consistency indicates whether the candidate correspondences agree with the local geometry under the current pose. This comparison explicitly measures the alignment between the projected 3D topology and the spatial arrangement of image patches, providing a direct geometric cue for candidate reasoning. Figure 3 illustrates how the candidate nodes, topological edges, and geometry-guided message passing are connected within CGR.

Message Passing. Given the node and edge features, CGR aggregates information from the candidate neighborhood $\mathcal{N}(v)$ through edge-conditioned message passing:

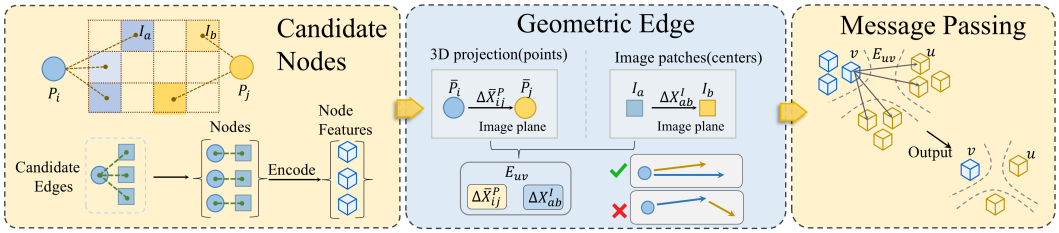


Figure 3: Illustration of the Candidate Graph Reasoner. Each node represents a correspondence between a 3D point and a candidate patch. Edge features compare the relative displacements of projected points and patch centers, and CGR updates the candidate features and matching scores through geometry-guided message passing.

$$\mathbf{m}_v = \sum_{u \in \mathcal{N}(v)} \alpha_{vu}\, f_{\text{msg}}(\mathbf{h}_u, \mathbf{E}_{vu})\,, \qquad \widetilde{\mathbf{h}}_v = f_{\text{upd}}(\mathbf{h}_v, \mathbf{m}_v)\,. \tag{1}$$

In Equation (1), $\mathbf{m}_v$ is the neighborhood message aggregated at node v, α_{vu} is the attention weight of neighboring node u, and $\mathbf{E}_{vu}$ encodes the geometric relation between the two candidates. The updated feature $\widetilde{\mathbf{h}}_v$ combines the node feature with neighborhood information, strengthening geometrically consistent candidates while suppressing inconsistent matches.

The updated node features predict geometric corrections, which are combined with the base CGM scores and the candidate evidence to determine the final ranking. A position head regresses a continuous 2D coordinate within each candidate patch, while a confidence head combines the candidate distribution, node features, and geometric support to predict the correspondence confidence ρ_v. Only high-confidence correspondences are used for pose estimation in subsequent stages. Unstable correspondences caused by object motion, occlusion, or incorrect instance association generally lack consistent feature and geometric support and are therefore unlikely to enter later geometric feedback. This confidence-based filtering reduces interference from dynamic regions during Progressive Matching.

3.4.3 Pose Refiner

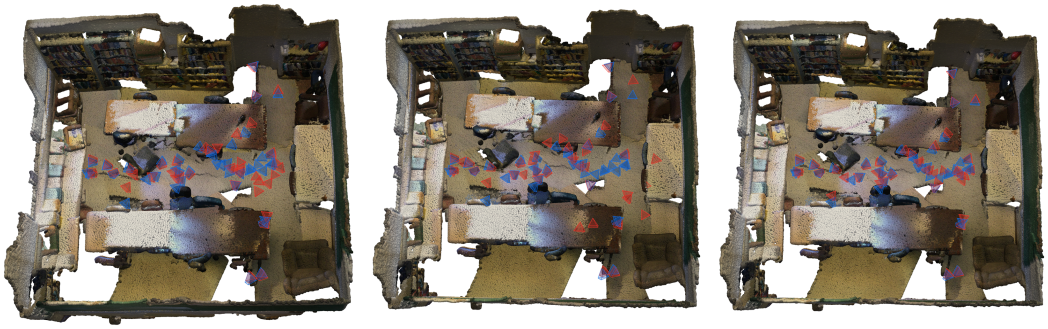
The Pose Refiner is placed at the end of both the scene graph matching stage and the progressive matching stage. It first collects high-confidence 2D–3D correspondences and estimates a coarse camera pose using PnP-RANSAC, after which the points of the candidate 3D instances are projected onto the image plane. Instance-based occlusion filtering constrains these projections by removing 3D points outside the viewing frustum or image bounds, and retaining only those projected inside the corresponding 2D instance regions. Since instance masks explicitly separate foreground objects from other objects and the background, this filtering effectively suppresses incorrect projections caused by occlusion, viewpoint changes, or overlapping instances, yielding a more reliable set of visible points. Finally, the topological relations among the retained projected points are used to construct candidate edges for the next scale, providing geometric constraints for subsequent fine-grained matching.

3.5 Loss Functions

IHF-ReLoc adopts multi-stage multi-task supervision. The cross-modal matching objective employs contrastive learning to pull matched pairs closer in the feature space while pushing unrelated pairs apart. This contrastive objective is applied to both instance-level scene graph matching and the subsequent progressive point-to-patch matching stages. The geometric objective supervises the relative geometry between projected points. Specifically, the direction loss supervises the relative directions between neighboring correspondences, while the length loss uses Smooth L1 to constrain their relative distances. These constraints encourage candidate correspondences to preserve consistent local geometry. The confidence branch is supervised with binary cross-entropy using soft targets. The soft target decreases with the pixel distance between the predicted and ground-truth locations, allowing the confidence score to reflect the actual reliability of each correspondence for pose estimation.

Table 1: Relocalization results on ScanNet v2. For each method, a ✓ denotes the use of coarse pose initialization or local-map truncation. We report rotation and translation errors, along with pose recalls—the ratios of successfully localized queries at different thresholds. We also provide the average runtime per query. Best and second-best accuracy results are shown in bold and underlined, respectively.

Methods	Coarse Init.	Local Map	ΔR (°) ↓		Δt (m) ↓		Pose Recall (%) ↑			Time (s/iter)
			Mean	Med.	Mean	Med.	(0.2 m, 10°)	(0.5 m, 15°)	(1.0 m, 30°)	
FreeReg		✓	32.8	25.6	1.67	0.85	2.6	10.5	37.8	∼10.9
2D3D-MATR		✓	17.1	5.8	1.42	0.34	36.8	60.6	70.3	∼0.2
CoFiI2P		✓	24.7	5.7	1.22	0.36	36.1	49.7	68.9	∼0.1
Hg-I2P		✓	<u>14.1</u>	<u>4.1</u>	<u>1.12</u>	<u>0.23</u>	<u>43.1</u>	<u>65.2</u>	<u>77.3</u>	∼0.2
PLANA3R	✓		20.7	8.3	1.38	0.48	20.4	41.7	57.9	∼0.4
PlanaReLoc			18.2	4.9	1.26	0.32	35.3	60.4	69.1	∼0.2
Ours			**10.7**	**3.7**	**0.92**	**0.16**	**47.6**	**77.8**	**86.4**	∼0.3



Hg-I2P PlanaReLoc Ours

Figure 4: Qualitative visual localization results on ScanNet v2. Red and blue camera frustums denote the estimated and ground-truth poses.

The overall objective is defined as

$$\mathcal{L}_{\text{total}} = \sum_{\ell=0}^{3} \lambda_\ell \mathcal{L}_{\text{con}}^{(\ell)} + \lambda_{\text{conf}} \mathcal{L}_{\text{conf}} + \lambda_{\text{dir}} \mathcal{L}_{\text{dir}} + \lambda_{\text{len}} \mathcal{L}_{\text{len}}. \tag{2}$$

Here, $\mathcal{L}_{\text{con}}^{(\ell)}$ denotes the contrastive loss at matching stage ℓ, where $\ell = 0$ corresponds to instance-level scene graph matching and $\ell \in \{1, 2, 3\}$ denotes the three progressive point-to-patch matching levels. $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\text{dir}}$, and $\mathcal{L}_{\text{len}}$ denote the confidence, relative direction, and relative length losses, respectively.

4 Experiments

Datasets. We evaluate visual localization performance on ScanNet v2 and assess dynamic localization performance on 3RScan. ScanNet v2 (Dai et al., 2017) contains indoor RGB-D sequences, 3D point clouds, and semantic instance annotations. Sampling every 20 frames and filtering yields 71,987 training images from 926 scans and 18,402 evaluation images from 214 scans (Appendix B.1). All images are resized to 640×480. 3RScan (Wald et al., 2019) contains 1,482 scans of 478 indoor environments and provides instance correspondences and scene-change annotations between reference scans and rescans. We construct the map from the reference scan and use rescan images captured at different times as cross-temporal queries. The query poses are transformed into the coordinate system of the reference map using the inter-scan registration transformations. Since some 3RScan images have a narrow field of view and contain only a few visible objects, we retain only images with more than 4 valid instances for training and evaluation.

Table 2: Localization on 3RScan reference scans and cross-temporal rescans. "Ref only" uses reference scans for adaptation, while "rescan train" also uses training-set rescans. Ref and Rescan report same-domain and cross-temporal performance.

Set	Ref					Rescan				
	ΔR (°) ↓		Δt (m) ↓		PnP (%) ↑	ΔR (°) ↓		Δt (m) ↓		PnP (%) ↑
	Mean	Med.	Mean	Med.	Success Rate	Mean	Med.	Mean	Med.	Success Rate
ref only	14.84	5.23	1.31	0.27	97.13	24.97	8.52	1.86	0.44	96.39
rescan train	14.36	5.09	1.24	0.27	99.08	17.63	6.15	1.73	0.36	98.11



Figure 5: Dynamic confidence under temporal scene changes. The top and bottom rows show the reference map and query images, respectively. Darker red indicates lower correspondence confidence; dashed boxes mark major changed regions.

Baseline Methods and Evaluation Metrics. We select FreeReg (Wang et al., 2024b), 2D3D-MATR (Li et al., 2023), CoFiI2P (Kang et al., 2024), Hg-I2P (An et al., 2026), PLANA3R (Liu et al., 2025), and PlanaReLoc (Ye et al., 2026) as baseline methods. Retrieval-free methods are selected to ensure the same level of map prior information as our method. We evaluate pose accuracy using the mean and median Relative Rotation Error (RRE) and Relative Translation Error (RTE), measured in degrees and meters, respectively. The evaluation also includes pose recalls R_1, R_2, and R_3 under thresholds of $(0.2\,\mathrm{m}, 10°)$, $(0.5\,\mathrm{m}, 15°)$, and $(1.0\,\mathrm{m}, 30°)$, respectively, indicating the percentage of successfully localized queries under different accuracy requirements.

Implementation Details. IHF-ReLoc is trained in multiple stages. We first train the 2D instance segmentation network, Mask2Former. The segmentation branch is then frozen while the Scene Graph Matching module is trained. After the instance-level candidates become stable, training proceeds to the progressive matching module. During the early stage of training, incorrect correspondences may lead to unstable pose feedback. Therefore, the Pose Refiner employs a fallback mechanism: when the predicted pose deviates beyond a reasonable range, the current estimate is replaced with the reference pose perturbed within a bounded range. The fallback mechanism is disabled during validation and testing. The full model is trained on a single 24-GB NVIDIA RTX 4090 GPU with a batch size of 2. For the graph-matching stages, the learning rate decays from 1×10^{-5} to 1×10^{-7}.

4.1 Indoor Visual Relocalization Results

Table 1 presents the visual relocalization results on ScanNet v2. IHF-ReLoc achieves the best results in rotation error, translation error, and pose recall under different thresholds. IHF-ReLoc not only has low median errors, but its mean errors are also significantly lower than those of most methods, indicating that the model can achieve robust localization without relying on image-retrieval priors. In terms of efficiency, the per-frame inference time of IHF-ReLoc is approximately 0.3 s, maintaining an acceptable computational cost while achieving higher localization accuracy. As shown in Figure 4, our estimated poses align more closely with the ground truth than those of the competing methods, demonstrating more accurate visual localization.

4.2 CROSS-TEMPORAL SCENE LOCALIZATION

Table 2 evaluates the localization ability of IHF-ReLoc under cross-temporal scene changes in 3RScan. When only reference scans are used for adaptation, the mean rotation error of the model on rescans increases significantly, indicating that instance movement, object addition and removal, and occlusion changes introduce severe incorrect correspondences. After rescan data are added for adaptation, the mean and median RRE on rescans decrease from 24.97° to 17.63° and from 8.52° to 6.15°, indicating that cross-temporal training reduces both severe localization failures and typical pose deviations. The dynamic-confidence visualization in Figure 5 also shows that the model can accurately identify regions affected by object movement, disappearance, and addition and actively assign lower matching confidence to these regions, thereby suppressing unstable correspondences produced by dynamic objects. Together, the quantitative and qualitative results show that IHF-ReLoc can effectively perceive cross-temporal scene changes and reduce the long-tailed localization errors caused by them.

4.3 ABLATION STUDY

Table 3: Ablation study of Scene Graph Matching, Candidate Graph Reasoner, and Occlusion Filtering on ScanNet v2. ✓ indicates that a module is enabled, × indicates that it is omitted, and ◯ indicates that Candidate Graph Reasoner is replaced by an alternative topology propagation design.

Scene Graph Matching	Candidate Graph Reasoner	Occlusion Filtering	ΔR (°) ↓		Δt (m) ↓		PnP Success Rate (%) ↑
			Mean	Med.	Mean	Med.	
×	×	✓	28.6	9.4	2.93	0.93	97.72
✓	×	✓	17.2	4.7	1.21	0.31	99.46
✓	◯	✓	12.5	4.1	1.14	0.17	99.86
✓	✓	×	14.8	3.8	1.07	0.20	99.71
✓	✓	✓	10.7	3.7	0.92	0.16	99.98

Table 3 reports the ScanNet ablation results. Scene Graph Matching provides the largest overall gain by narrowing the search space and reducing large-scale ambiguity. Occlusion Filtering lowers pose errors by excluding invisible or occluded map points from local matching. The ◯ configuration encodes the topology of projected 3D points in the edge features of Cross-modal Graph Matching (CGM) graph self-attention, without using Candidate Graph Reasoner (CGR). CGR achieves higher localization accuracy than this alternative. In the alternative design, topology is first propagated among point features before it influences cross-modal matching. Candidate Graph Reasoner instead represents each point-to-patch hypothesis as a graph node and passes geometric evidence directly between neighboring hypotheses. This allows a candidate to be evaluated against the geometry of nearby correspondences, helping suppress visually similar matches whose relative positions are inconsistent with the projected 3D structure. The lower pose errors suggest that this direct candidate-level reasoning uses topological constraints more effectively.

5 CONCLUSION

We presented IHF-ReLoc, an instance-aware hierarchical framework that progressively refines 2D–3D visual relocalization from instance-level associations to fine-grained point-to-patch correspondences. Scene Graph Matching exploits the topology of 2D and 3D instance graphs to reduce cross-modal ambiguity and uses instance-level visibility to filter out occluded map points. The Candidate Graph Reasoner models point-to-patch candidates and their local geometric relations as a graph, propagating geometric information to suppress visually similar but incorrect matches. Experiments on ScanNet v2 show state-of-the-art localization performance, and evaluations on 3RScan demonstrate robustness to temporal scene changes without pose priors from image retrieval. Together, these results show how instance-level visibility and candidate-level geometric consistency make 2D–3D matching more reliable.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and editing, including grammar and clarity checks and LaTeX formatting. They also helped identify relevant literature and check bibliographic details, such as titles and publication information, against original papers or publisher records. AI-assisted LaTeX edits were compiled and visually reviewed. The authors reviewed the AI-assisted content and take responsibility for the final text, citations, and claims.

REPRODUCIBILITY STATEMENT

Implementation details, training settings, and evaluation protocols are provided in the main paper and supplementary material. The source code, configuration files, and trained models will be publicly released upon acceptance.

REFERENCES

Pei An, Junfeng Ding, Jiaqi Yang, Yulong Wang, Jie Ma, and Liangliang Nan. Hg-I2P: Bridging modalities for generalizable image-to-point-cloud registration via heterogeneous graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39042–39051, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/An_Hg-I2P_Bridging_Modalities_for_Generalizable_Image-to-Point-Cloud_Registration_via_Heterogeneous_Graphs_CVPR_2026_paper.html.

Lin Bie, Shouan Pan, Siqi Li, Yining Zhao, and Yue Gao. GraphI2P: Image-to-point cloud registration with exploring pattern of correspondence via graph learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22161–22171, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bie_GraphI2P_Image-to-Point_Cloud_Registration_with_Exploring_Pattern_of_Correspondence_via_CVPR_2025_paper.html.

Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5044–5053, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Brachmann_Accelerated_Coordinate_Encoding_Learning_to_Relocalize_in_Minutes_Using_RGB_CVPR_2023_paper.html.

Leonard Bruns, Axel Barroso-Laguna, Tommaso Cavallari, Áron Monszpart, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. ACE-G: Improving generalization of scene coordinate regression through query pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26751–26761, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Bruns_ACE-G_Improving_Generalization_of_Scene_Coordinate_Regression_Through_Query_Pre-Training_ICCV_2025_paper.html.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Cheng_Masked-Attention_Mask_Transformer_for_Universal_Image_Segmentation_CVPR_2022_paper.html.

Zhixin Cheng, Jiacheng Deng, Xinjun Li, Baoqun Yin, and Tianzhu Zhang. Bridge 2D-3D: Uncertainty-aware hierarchical registration network with domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2491–2499, 2025. doi: 10.1609/aaai.v39i3.32251. URL https://ojs.aaai.org/index.php/AAAI/article/view/32251.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Mengdan Feng, Sixing Hu, Marcelo H. Ang, and Gim Hee Lee. 2D3D-MatchNet: Learning to match keypoints across 2D image and 3D point cloud. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 4790–4796, 2019. doi: 10.1109/ICRA.2019.8794415. URL https://arxiv.org/abs/1904.09742.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17658–17668, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Izquierdo_Optimal_Transport_Aggregation_for_Visual_Place_Recognition_CVPR_2024_paper.html.

Shuhao Kang, Youqi Liao, Jianping Li, Fuxun Liang, Yuhao Li, Xianghong Zou, Fangning Li, Xieyuanli Chen, Zhen Dong, and Bisheng Yang. CoFiI2P: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics and Automation Letters*, 9(11):10264–10271, 2024. doi: 10.1109/LRA.2024.3466068. URL https://doi.org/10.1109/LRA.2024.3466068.

Minjung Kim, Junseo Koo, and Gunhee Kim. EP2P-Loc: End-to-end 3D point to 2D pixel localization for large-scale visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21527–21537, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kim_EP2P-Loc_End-to-End_3D_Point_to_2D_Pixel_Localization_for_Large-Scale_ICCV_2023_paper.html.

Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2D3D-MATR: 2D-3D matching transformer for detection-free registration between images and point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14128–14138, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_2D3D-MATR_2D-3D_Matching_Transformer_for_Detection-Free_Registration_Between_Images_and_ICCV_2023_paper.html.

Zhenyu Li and Tianyi Shang. A2GC: Asymmetric aggregation with geometric constraints for locally aggregated descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19423–19431, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_A2GC_Asymmetric_Aggregation_with_Geometric_Constraints_for_Locally_Aggregated_Descriptors_CVPR_2026_paper.html.

Changkun Liu, Bin Tan, Zeran Ke, Shangzhan Zhang, Jiachen Liu, Ming Qian, Nan Xue, Yujun Shen, and Tristan Braud. PLANA3R: Zero-shot metric planar 3D reconstruction via feed-forward planar splatting. In *Advances in Neural Information Processing Systems*, 2025. URL https://lck666666.github.io/plana3r/.

Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Daniel Barath. SceneGraphLoc: Cross-modal coarse visual localization on 3D scene graphs. In *European Conference on Computer Vision*, pp. 127–150, 2024. doi: 10.1007/978-3-031-73242-3_8. URL https://link.springer.com/chapter/10.1007/978-3-031-73242-3_8.

Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. LCD: Learned cross-domain descriptors for 2D-3D matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11856–11864, 2020. doi: 10.1609/aaai.v34i07.6859. URL https://ojs.aaai.org/index.php/AAAI/article/view/6859.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7199–7209, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Taira_InLoc_Indoor_Visual_CVPR_2018_paper.html.

Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, and Yasutaka Furukawa. NeuMap: Neural coordinate mapping by auto-transdecoder for camera localization. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 929–939, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tang_NeuMap_Neural_Coordinate_Mapping_by_Auto-Transdecoder_for_Camera_Localization_CVPR_2023_paper.html.

Wenjing Tang, Chuanguang Yang, Zhulin An, Libo Huang, Boyu Diao, and Yongjun Xu. EfficientVPR: Toward efficient visual place recognition via scene-aware prompt tuning and adaptive feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33738–33748, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tang_EfficientVPR_Toward_Efficient_Visual_Place_Recognition_via_Scene-Aware_Prompt_Tuning_CVPR_2026_paper.html.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. 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. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Wald_RIO_3D_Object_Instance_Re-Localization_in_Changing_Indoor_Environments_ICCV_2019_paper.html.

Johanna Wald, Torsten Sattler, Stuart Golodetz, Tommaso Cavallari, and Federico Tombari. Beyond controlled environments: 3D camera re-localization in changing indoor scenes. In *European Conference on Computer Vision*, pp. 467–487, 2020. doi: 10.1007/978-3-030-58571-6_28. URL https://link.springer.com/chapter/10.1007/978-3-030-58571-6_28.

Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, and Andrew Markham. P2-Net: Joint description and detection of local features for pixel and point matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16004–16013, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wang_P2-Net_Joint_Description_and_Detection_of_Local_Features_for_Pixel_ICCV_2021_paper.html.

Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. GLACE: Global local accelerated coordinate encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21562–21571, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_GLACE_Global_Local_Accelerated_Coordinate_Encoding_CVPR_2024_paper.html.

Haiping Wang, Yuan Liu, Bing Wang, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. FreeReg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=BPb5AhT2Vf.

Shuzhe Wang, Juho Kannala, and Daniel Barath. DGC-GNN: Leveraging geometry and color cues for visual descriptor-free 2D-3D matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20881–20891, 2024c. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DGC-GNN_Leveraging_Geometry_and_Color_Cues_for_Visual_Descriptor-Free_2D-3D_CVPR_2024_paper.html.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4840–4851, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Point_Transformer_V3_Simpler_Faster_Stronger_CVPR_2024_paper.html.

Hanqiao Ye, Yuzhou Liu, Yangdong Liu, and Shuhan Shen. PlanaReLoc: Camera relocalization in 3D planar primitives via region-based structure matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26409–26421, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ye_PlanaReLoc_Camera_Relocalization_in_3D_Planar_Primitives_via_Region-Based_Structure_CVPR_2026_paper.html.

Sitao Zhang, Hongda Mao, Qingshuang Chen, and Yelin Kim. Efficient visual place recognition through multimodal semantic knowledge integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5601–5610, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Efficient_Visual_Place_Recognition_Through_Multimodal_Semantic_

Knowledge_Integration_ICCV_2025_paper.html.

Yejun Zhang, Shuzhe Wang, and Juho Kannala. A2-GNN: Angle-annular GNN for visual descriptor-free camera relocalization. In *International Conference on 3D Vision*, pp. 357–368, 2025b. URL https://arxiv.org/abs/2502.20036.

Qunjie Zhou, Sergio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. Is geometry enough for matching in visual localization? In *European Conference on Computer Vision*, pp. 407–425, 2022. doi: 10.1007/978-3-031-20080-9_24. URL https://link.springer.com/chapter/10.1007/978-3-031-20080-9_24.

A Supplementary Material Overview

This supplementary material provides additional details on dataset preparation, implementation, component-level evaluation, and qualitative analysis for IHF-ReLoc. Appendix B describes the preprocessing, coordinate handling, and query selection used for ScanNet v2 and 3RScan. Appendix C provides further implementation details for the image and point encoders, scene-graph matching, progressive matching, candidate-graph reasoning, training, and pose refinement. Appendix D reports component-level quantitative results for image-instance segmentation and 2D–3D instance association. Finally, Appendix E presents qualitative examples of final camera localization, including additional localization examples across diverse indoor scenes.

B Dataset Preparation

B.1 ScanNet v2

We start from the official ScanNet v2 split of 1,201 training scans and 312 validation scans (Dai et al., 2017). Each scan provides RGB images, camera calibration, camera poses, reconstructed geometry, and semantic and instance annotations. Images are sampled every 20 frames. After filtering, the training set contains 71,987 images from 926 scans, and the validation set contains 18,402 images from 214 scans. The resulting dataset comprises 90,389 images from 1,140 scans, as summarized in Table B.1.

Table B.1: ScanNet v2 dataset size after filtering.

Split	Scans	Images
Training	926	71,987
Validation	214	18,402
Total	1,140	90,389

The source images are center-cropped from 1296×968 to 1264×948 and resized to 640×480. Semantic and instance maps undergo the same spatial transformation, with nearest-neighbor interpolation used for discrete labels. The camera intrinsics $\mathbf{K}$ are updated consistently to account for the crop offset and resize scale, following the same camera model used in the main paper.

The input point cloud contains point coordinates and colors. Semantic and instance annotations are used during dataset preparation and training supervision rather than as point features. Semantic labels are mapped to 20 contiguous categories, and invalid labels are ignored. Scene-graph construction retains instances with at least 100 points and a valid semantic category. Each instance is represented by its points, bounding-box center and extent, and normalized local coordinates. The alignment loader applies 0.02 m grid sampling to the scene point cloud. The 2D masks and 3D instances share a consistent identity mapping, allowing projected map points to provide instance-pairing and point-to-patch supervision.

B.2 3RScan

For cross-temporal localization, the reference scan provides the 3D map, while reference and rescan trajectories provide same-time and cross-time query images, respectively (Wald et al., 2019). The split lists contain 385 training families and 47 validation families. The remaining 46 families belong to the test split and are not used in our experiments. A family consists of a reference scan and its associated rescans, and all scans belonging to the same family remain in one split. These family counts differ from the numbers of individual scans and retained query images.

Rescan camera poses are transformed into the reference-map coordinate system using the provided inter-scan registration and reference-axis alignment. After this transformation, all query poses are expressed using the same camera-pose convention $\mathbf{T}_{cw}$ as in the main paper.

The image pipeline center-crops the 960×540 source images to 720×540, resizes them to 640×480, and rotates them clockwise to obtain 480×640 portrait images. Intrinsics and camera axes are

transformed consistently with this rotation. The default frame interval is two. Ground-truth instance masks are generated by ray casting against the query mesh and remapping object identities into the reference-map identity space. Semantic labels use the 27-category RIO taxonomy.

The 3RScan query filtering is performed entirely from ground-truth annotations and does not use model predictions. For each query frame, we first inspect the ground-truth instance map generated by mesh rasterization; frames without a valid instance map are discarded. Background regions, instances absent from the reference scene graph, and instances covering fewer than 64 visible pixels are then removed. The reference scene graph retains only map instances with a valid RIO27 semantic label and at least 100 3D points. Finally, a query frame is kept only if it contains more than four valid visible instances. The number of ground-truth instances is used only for data filtering and result grouping, and is never used during instance segmentation, instance pairing, or pose estimation at inference time.

This selection is applied consistently to the compared methods. The training set retains 69,579 of 143,671 images (48.43%), and the validation set retains 9,516 of 19,368 images (49.13%). These counts aggregate the query images within each split. The resulting evaluation concerns the instance-rich subset rather than the complete trajectories.

Table B.2: 3RScan image counts before and after instance-count filtering.

Split	Before filtering	After filtering	Retention (%)
Training	143,671	69,579	48.43
Validation	19,368	9,516	49.13

C Implementation Details

C.1 Network Architecture

Multi-scale image encoding. The image branch uses a frozen DINOv3 ViT-B/16 backbone (Siméoni et al., 2025) and a trainable adapter to construct multi-scale image features $\mathbf{F}_l^{2d}$ at strides 4, 8, 16, and 32. Features at all scales are projected into a shared 128-dimensional matching space and are jointly used for instance segmentation and point-to-patch correspondence prediction.

Point and instance encoding. The point-cloud branch uses Point Transformer V3 (Wu et al., 2024) to extract multi-scale point features $\mathbf{F}_l^{3d}$. Its encoder is frozen, while the decoder provides point features for local matching. Instance-wise aggregation produces the 3D instance representation $\mathbf{F}_{\text{ins}}^{3d}$, which is projected into the same 128-dimensional space as the image features. Point coordinates normalized within each instance bounding box provide local positional information.

Mask2Former and 2D instance graph. The instance-segmentation branch is based on Mask2Former (Cheng et al., 2022) and directly reuses the shared image pyramid. The stride-4 feature map is used for mask prediction, while the stride-8, stride-16, and stride-32 features are used by the masked-attention decoder. The decoder maintains 40 object queries. After low-confidence filtering and mask NMS, features within each retained mask are pooled into the image-instance representation $\mathbf{F}_{\text{ins}}^{2d}$, which forms the nodes of the 2D instance graph.

Cross-modal graph matching. Both instance-level and local matching use the Cross-modal Graph Matching (CGM) module described in the main paper. CGM first aggregates intra-modal neighborhood structure through graph self-attention, then exchanges information between image and point-cloud features through bidirectional cross-attention. Matching similarities are computed in the shared 128-dimensional feature space, using the same interaction pattern at the instance and local levels.

Instance scoring and local verification. For an image instance y and a 3D map instance x, the instance score $S_{\text{inst}}(x, y)$ defined in the main paper is used to retain the three highest-scoring 3D candidates. Local point-to-patch matching is then performed for each candidate, and matching probabilities, visible-point support, and pooled local features are used to obtain the local verification

score $S_{\mathrm{loc}}(x, y)$. The two scores are applied sequentially: S_{inst} restricts the candidate set, while S_{loc} selects the final instance match within that set.

Progressive matching and candidate graph reasoning. Local matching is progressively refined from 32×32 to 16×16 and 4×4 patches. Finer stages search only within the child regions of high-scoring parent patches, progressively reducing the matching space. At the stride-16 and stride-4 stages, each 3D point retains $K_c = 4$ candidate patches. Following the notation in Section 3.4.2, each point-to-patch hypothesis forms a candidate node $v = (\mathbf{P}_i, I_a)$. The edge feature $\mathbf{E}_{vu}$ captures geometric consistency between neighboring projected 3D points and their candidate patches, including relative displacement, distance, and direction information. CGR uses these relations to refine the CGM matching scores and candidate ranking. The highest-scoring candidate is used for continuous 2D coordinate prediction, while the confidence head outputs the correspondence confidence ρ_v for pose estimation.

C.2 Training Protocol

IHF-ReLoc is trained in stages. We first train the Mask2Former segmentation branch, then freeze it and train Scene Graph Matching to obtain stable instance candidates. Finally, the progressive point-to-patch matching modules, CGR, and pose feedback are optimized jointly. The full model is trained with a batch size of 2 on a single 24-GB NVIDIA RTX 4090. For the graph-matching stages, the learning rate decays from 1×10^{-5} to 1×10^{-7}.

During training only, the Pose Refiner uses the fallback strategy described in the main paper when an intermediate pose is too inaccurate to provide reliable projection geometry. A bounded perturbation of the reference pose is then used to construct the geometric context for the next scale. The fallback is disabled during validation and testing, so all reported results rely only on the correspondences and poses predicted by the model.

C.3 Pose Estimation and Refinement

High-confidence correspondence selection and PnP. Each local matching stage selects valid correspondences according to the confidence ρ_v. Intermediate pose estimation uses a confidence threshold of 0.50 and retains up to eight high-confidence correspondences per instance, which prevents a densely sampled object from dominating the PnP input. If fewer than four valid correspondences remain, the filter is relaxed to retain enough matches for pose estimation.

PnP-RANSAC estimates the intermediate camera pose $\mathbf{T}_{cw}$ from the selected 2D–3D correspondences and the camera intrinsics $\mathbf{K}$. The intermediate solver uses a 16-pixel reprojection threshold, 1,000 RANSAC iterations, and confidence 0.99.

Projection, occlusion filtering, and geometric feedback. The intermediate pose $\mathbf{T}_{cw}$ is used to project candidate map points into the image. A shared 2D rotation and translation are used to lightly align the projected layout with high-confidence predicted image coordinates, reducing global displacement caused by residual pose error. This correction affects only the projected geometry used by subsequent matching and does not modify the PnP pose $\mathbf{T}_{cw}$.

The projected points are then filtered using the viewing frustum, image bounds, and the corresponding 2D instance masks. Points behind the camera, outside the image, or projected onto inconsistent instance regions are removed. The remaining visible projections provide the relative directions and distances used to construct the candidate graph at the next scale, thereby passing pose feedback to subsequent matching stages.

Final camera pose. After 4×4 matching and continuous coordinate prediction, high-confidence correspondences are pooled across instances for the final PnP-RANSAC solve. The final stage uses a confidence threshold of 0.50, retains up to 16 correspondences per instance, and applies a 6-pixel reprojection threshold with 1,000 iterations and confidence 0.99. At least four inliers are required. The output follows the same camera-pose convention $\mathbf{T}_{cw}$ as in the main paper.

D Additional Quantitative Evaluation

This section reports component-level results for image segmentation and 2D–3D instance association on ScanNet v2. All values are percentages.

D.1 Mask2Former Segmentation Quality

Mask2Former provides the image-instance masks used to construct the 2D scene graph. We report mask average precision at IoU thresholds of 0.50 and 0.75 (AP50 and AP75), average precision across multiple IoU thresholds (AP), and mean intersection over union (mIoU). AP50 and AP75 evaluate instance-detection and mask quality at different overlap requirements, while AP summarizes performance across thresholds and mIoU measures the overall mask overlap.

Table D.1: Image segmentation quality on the ScanNet v2 evaluation split (%). Higher is better for all metrics.

AP50 ↑	AP75 ↑	AP ↑	mIoU ↑
51.6	33.4	39.8	50.7

The segmentation branch achieves 51.6% AP50, 33.4% AP75, 39.8% AP, and 50.7% mIoU. These results show that it provides reliable instance masks for constructing the 2D scene graph. The decrease from AP50 to AP75 indicates that precise mask boundaries remain more challenging under a stricter overlap criterion and may affect subsequent instance association and local matching.

D.2 Instance-Pairing Quality

We evaluate instance association at two stages. Instance-Pairing Top-1 measures the percentage of image instances whose highest-scoring 3D candidate is correct, while Instance-Pairing Top-3 Recall measures whether the correct 3D instance is retained among the top three candidates. Local-Verification Refined Top-1 then measures the accuracy of selecting the correct instance from these top-three candidates when the ground-truth instance is present in the candidate set.

Table D.2: Instance-pairing results on ScanNet v2 (%). Refined Top-1 is evaluated on samples for which the correct instance is retained in the Top-3 candidates.

Metric	Result ↑
Instance-Pairing Top-1	71.75
Instance-Pairing Top-3 Recall	95.49
Local-Verification Refined Top-1	90.40

The results show that retaining three candidates substantially increases the chance of preserving the correct 3D instance, and local verification further selects the correct candidate with 90.4% accuracy when it is available in the Top-3 set.

E Qualitative Analysis

We provide additional camera localization visualizations to complement the qualitative results presented in the main paper. The examples show the estimated and ground-truth camera poses across diverse indoor scenes.

E.1 Additional Camera Localization Results

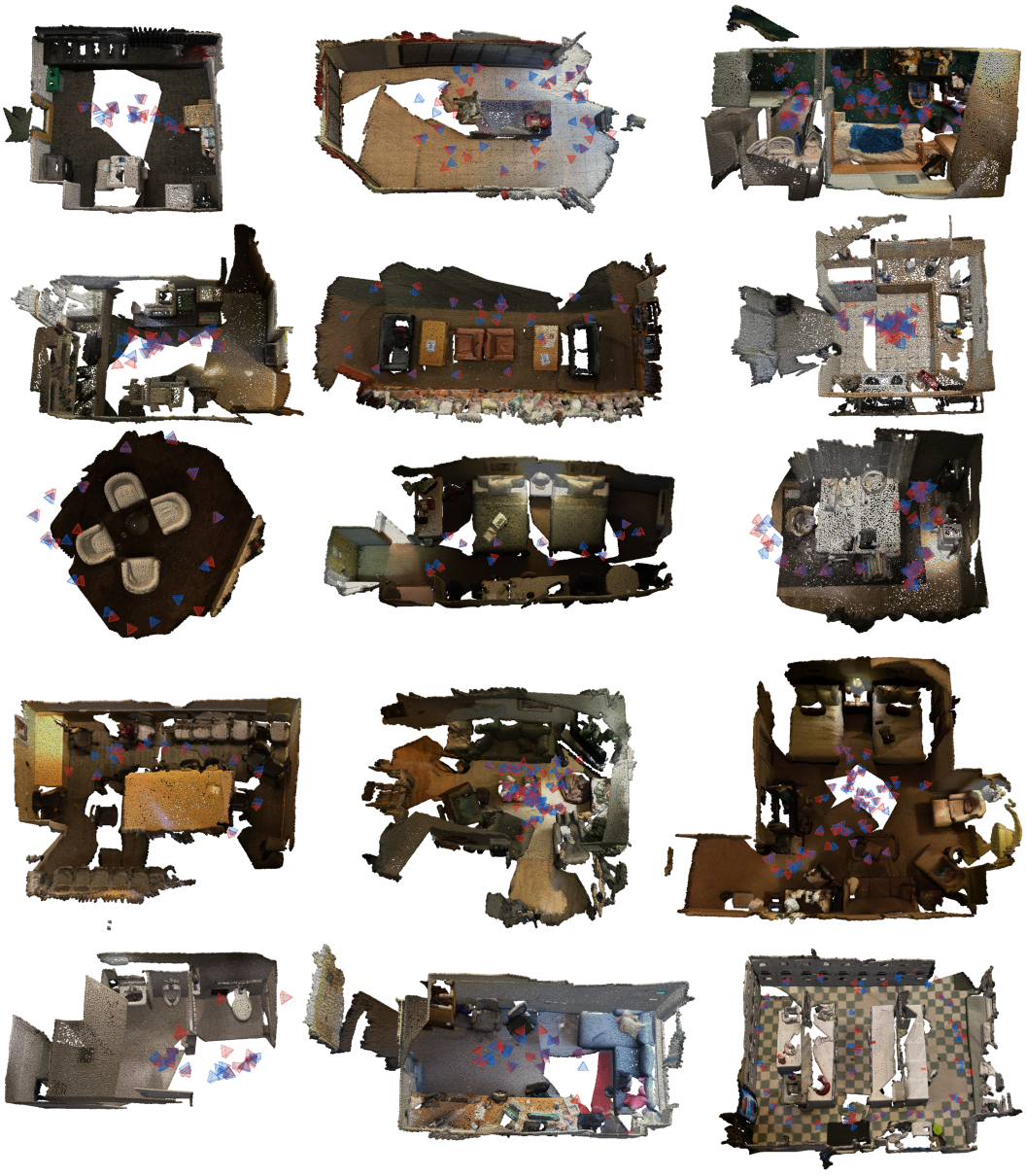


Figure E.1: Additional camera localization results across diverse indoor scenes. Red and blue camera frustums indicate the estimated and ground-truth poses.