OLA-Place

Cross-Modal Place Recognition without Global Descriptors

Tianyi Shang **Zhenyu Li**

NeurIPS 2026

PURDUE UNIVERSITY

齐鲁工业大学（山东省科学院）
QILU UNIVERSITY OF TECHNOLOGY SHANDONG ACADEMY OF SCIENCES

The task: find a 3D place from language

Input: a language query and a gallery of 3D scene cells
Goal: retrieve the matching cell and estimate coordinates.

Query excerpt

**“Pose is on-top of a green road.
Pose is east of a gray sidewalk.
Pose is north of a black vegetation.”**

Target place



Incorrect candidate

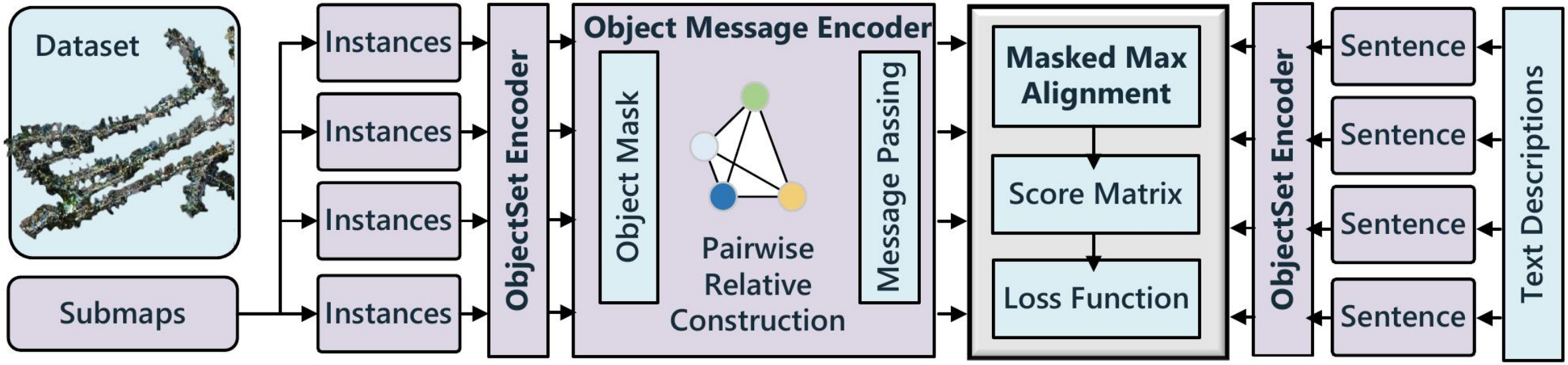


Language may describe only a few objects in a larger scene. Global compression can obscure the correspondences needed to distinguish similar places.

Focus: coarse retrieval. Fine localization follows the Text2Loc design.

Overview of the OLA-Place framework

Preserve one feature per object until the query-cell score



OSE: encode objects

Encode text units. Fuse 3D object geometry, semantics and appearance.

OME: add scene context

Refine 3D objects using neighboring objects and relative geometry.

MMA: match the sets

For each text unit, find its best valid 3D match and average the similarities.

Training uses query-cell pairs, without object correspondence annotations.

OSE: encode textual units and 3D objects

Normalized object features in a shared embedding space

Textual object units

Split each query into M textual units.
Encode units with a language backbone.
Contextualize tokens, then max-pool.
Project and aggregate unit context.
Keep one normalized embedding per unit.

3D object instances

Use PointNet2-style geometry features.
Encode RGB color, object center
and normalized point count.
Add optional class/color embeddings.
Normalize and concatenate components.
Fuse selected features with an MLP.

Text set: M × d features

Scene set: N × d features

OSE retains object granularity. Context refinement and cross-modal matching are performed by OME and MMA.

OME: refine objects with scene context

Scene-only refinement keeps one vector per object slot

1 Pad and mask object slots

Pad features and centers to N slots.
Mark valid slots and valid object pairs.

2 Construct relative geometry

Use pairwise position offsets and project them into initial relation features.

3 Update and normalize

Update object and relation streams.
Normalize the final object features.

Masks and relative positions

$M_n = 1$ for valid slots, else 0
$R_{ab} = M_a \times M_b$
$r_{ab} = p_a - p_b$

At each message layer

$O_{new} = \mathrm{Rec}_O(O_{prev}) \odot M$
$Q_{new,ab} = \tanh(W_r [o_a; o_b; q_{ab}]) \times R_{ab}$

o_a, o_b and q_{ab} are previous-layer inputs.

No language input is used in OME. No global scene descriptor is produced.

Masked Max Alignment

Each textual cue finds its best valid scene object

Textual unit	Car	Pole	Padding
“red car”	0.90	0.20	masked
“tall pole”	0.10	0.80	masked

1. Mask padded slots
2. Keep each row’s largest similarity
3. Average the maxima

Cell score = (0.90 + 0.80) / 2 = 0.85

$$S(q, C) = \text{mean}_m \max_{n:\, \text{mask}_n = 1} \cos(t_m, o_n)$$

Object order does not affect this score. Matching is not one-to-one.

Illustrative cosine similarities, not measured model outputs.

Training, retrieval and coordinate refinement

Score-based contrastive calibration

Build a $B \times B$ score matrix from all query-cell pairs in the mini-batch.

Normalize $\exp(S / \tau)$ by rows and columns to form two probabilities.

Exclude diagonal positives from mining. Select hard query-to-cell negatives.

Penalize selected negatives with distance weights and $-\log(1 - P + \varepsilon)$.

Inference and fine localization

Encode gallery cells as object features and occupancy masks.

Score every cell with the same MMA rule and rank by descending score.

Fuse point-cloud and text features with Text2Loc-style cascaded cross-attention.

An MLP predicts an offset from the cell anchor to recover the absolute location.

Pooled feature means provide auxiliary training weights. Retrieval uses MMA.

Datasets, training and evaluation

KITTI360Pose

15.51 km² across 9 city regions
28,689 train queries across 5 regions
3,187 validation queries in 1 region
11,505 test queries across 3 regions

StreetLoc

4.7 km driving trajectory
1,600 query-submap pairs
64-channel LiDAR and high-precision GPS
19 semantic object classes

Stage	Epochs	Batch size	Learning rate
Coarse retrieval	20	64	5×10^{-4}
Fine localization	100	32	3×10^{-4}

Coarse object embeddings have 256 dimensions.

Full localization: k = 1 / 5 / 10 at 5 m and 10 m thresholds.
Coarse retrieval: Recall@1 / @3 / @5.

KITTI360Pose: full localization results

Table 1. Each entry is recall at error < 5 m / 10 m

Method	k = 1	k = 5	k = 10
Text2Pos (CVPR'22)	0.13 / 0.20	0.33 / 0.42	0.43 / 0.61
RET (AAAI'23)	0.16 / 0.25	0.35 / 0.51	0.46 / 0.65
Text2Loc (CVPR'24)	0.33 / 0.48	0.60 / 0.75	0.70 / 0.84
IFRP-T2P (ACM MM'24)	0.22 / 0.40	0.47 / 0.68	0.58 / 0.78
MambaPlace (IROS'25)	0.38 / 0.52	0.66 / 0.79	0.76 / 0.87
CMMLoc (CVPR'25)	0.39 / 0.53	0.67 / 0.80	0.77 / 0.87
PMSH (ICCV'25)	0.39 / 0.55	0.68 / 0.81	0.78 / 0.89
OLA-Place (ours)	0.50 / 0.71	0.76 / 0.91	0.82 / 0.95

+11 pp at 5 m and +16 pp at 10 m for Top-1 versus PMSH.

StreetLoc: full localization results

Table 1. Each entry is recall at error < 5 m / 10 m

Method	k = 1	k = 5	k = 10
Text2Pos (CVPR'22)	0.14 / 0.17	0.27 / 0.31	0.36 / 0.52
RET (AAAI'23)	0.13 / 0.21	0.32 / 0.47	0.40 / 0.58
Text2Loc (CVPR'24)	0.29 / 0.42	0.47 / 0.66	0.55 / 0.72
IFRP-T2P (ACM MM'24)	0.20 / 0.33	0.45 / 0.63	0.53 / 0.68
MambaPlace (IROS'25)	0.32 / 0.42	0.50 / 0.70	0.58 / 0.76
CMMLoc (CVPR'25)	0.31 / 0.44	0.50 / 0.68	0.57 / 0.78
PMSH (ICCV'25)	0.30 / 0.43	0.52 / 0.70	0.59 / 0.77
OLA-Place (ours)	0.40 / 0.53	0.65 / 0.78	0.74 / 0.87

Top-1 reaches 0.40 / 0.53, versus best prior 0.32 / 0.44.

Coarse retrieval improves across all ranks

Table 2. Coarse submap retrieval recall

	KITTI360Pose			StreetLoc		
Method	**k = 1**	**k = 3**	**k = 5**	**k = 1**	**k = 3**	**k = 5**
Text2Pos	0.12	0.25	0.33	0.10	0.22	0.25
RET	0.15	0.29	0.37	0.14	0.23	0.30
Text2Loc	0.28	0.49	0.58	0.23	0.42	0.50
IFRP-T2P	0.23	0.39	0.48	0.19	0.31	0.40
MambaPlace	0.31	0.53	0.62	0.26	0.45	0.57
CMMLoc	0.32	0.53	0.63	0.25	0.47	0.55
PMSH	0.34	0.56	0.65	0.26	0.48	0.57
OLA-Place	0.49	0.72	0.79	0.41	0.62	0.70

Top-1 improves by 15 percentage points on both datasets.

Ablation: context, alignment and masking

Table 3. Coarse submap retrieval recall

	KITTI360Pose			StreetLoc		
Variant	k = 1	k = 3	k = 5	k = 1	k = 3	k = 5
w/o OME	0.31	0.55	0.64	0.24	0.44	0.53
w/o MMA	0.40	0.62	0.71	0.31	0.52	0.60
w/o Mask	0.38	0.61	0.70	0.32	0.50	0.61
Full OLA-Place	**0.49**	**0.72**	**0.79**	**0.41**	**0.62**	**0.70**

w/o OME: replace context encoding with a two-layer MLP.
w/o MMA: use directly pooled similarity.
w/o Mask: allow padded slots to participate in alignment.

Replacing OME causes the largest Top-1 drop: 18 / 17 pp.

Computational cost and remaining limitations

Reported coarse-stage cost under the same configuration

Method	Total params	Backbone	Other params	Inference
PMSH	372 M	336 M	36 M	≈39 ms
OLA-Place	392 M	336 M	56 M	≈43 ms

Improved retrieval with ≈4 ms additional coarse inference time

The method still depends on a fixed semantic label space.
Geometrically similar places can remain difficult to distinguish.

Future work in the paper: language models and open-vocabulary
3D perception to reduce reliance on rigid object categories.