Towards Global Localization Using Multi-Modal Object-Instance Re-Identification

Aneesh Chavan
c.aneesh.1203@gmail.com
RRC, IIIT Hyderabad
India

Vaibhav Agrawal*
vaibhav.agrawal@research.iiit.ac.in
RRC, IIIT Hyderabad
India

Vineeth Bhat*
vineeth.bhat@students.iiit.ac.in
RRC, IIIT Hyderabad
India

Sarthak Chittawar*
sarthak.chittawar@research.iiit.ac.in
RRC, IIIT Hyderabad
India

Siddharth Srivastava
siddharth.sri89@gmail.com
Typeface Inc.
India

Chetan Arora
chetan@cse.iitd.ac.in
IIT Delhi
India

K Madhava Krishna
mkrishna@iiit.ac.in
RRC, IIIT Hyderabad
India

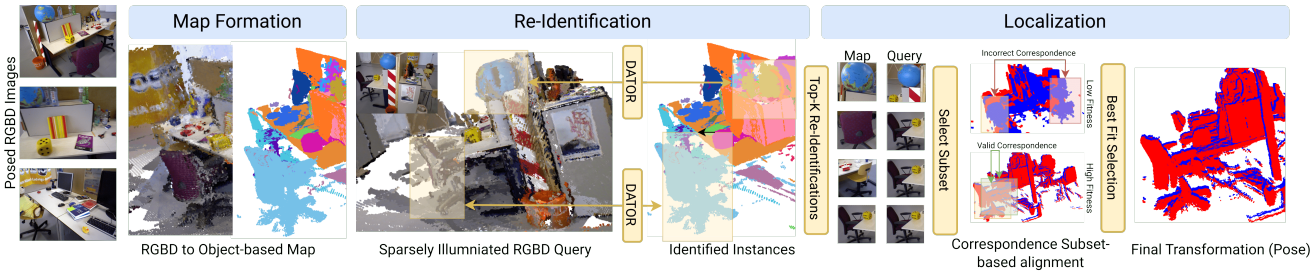

Figure 1: *Overview of the Localization Framework.* Our localization framework takes in posed-RGBD images and forms an object-based map consisting of instances and their descriptors. Given a query RGB-D image, we identify the objects within it and identify correspondences within our map using the ReID module (DATOR). The correspondences with the best fitness score are used to calculate the pose.

Abstract

Re-identification (ReID) is a critical challenge in computer vision, predominantly studied in the context of pedestrians and vehicles. However, robust object-instance ReID, which has significant implications for tasks such as autonomous exploration, long-term perception, and scene understanding, remains underexplored. In this work, we address this gap by proposing a novel dual-path object-instance re-identification transformer architecture that integrates multimodal RGB and depth information. By leveraging depth data, we demonstrate improvements in ReID across scenes that are cluttered or have varying illumination conditions. Additionally, we develop a ReID-based localization framework that enables accurate camera localization and pose identification across different viewpoints. We validate our methods using two custom-built RGB-D datasets, as well as multiple sequences from the open-source TUM RGB-D datasets. Our approach demonstrates significant improvements in both object instance ReID (mAP of 75.18) and localization accuracy (success rate of 83% on TUM-RGBD), highlighting the essential role of object ReID in advancing robotic perception. Our models, frameworks, and datasets have been made publicly available. [a]

CCS Concepts

• **Computing methodologies** → **Vision for robotics**; *Object identification.*

Keywords

Object re-identification, Localization, Multi-modal, Instance-based

ACM Reference Format:
Aneesh Chavan, Vaibhav Agrawal, Vineeth Bhat, Sarthak Chittawar, Siddharth Srivastava, Chetan Arora, and K Madhava Krishna. 2025. Towards Global Localization Using Multi-Modal Object-Instance Re-Identification. In *Proceedings of Make sure to enter the correct conference title from your*

*These authors contributed equally to this research.

AIR '25, Jodhpur, India

ACM ISBN 978-1-4503-XXXX-X/2025/07
https://doi.org/XXXXXXX.XXXXXXX

[a]https://instance-based-loc-machine.github.io

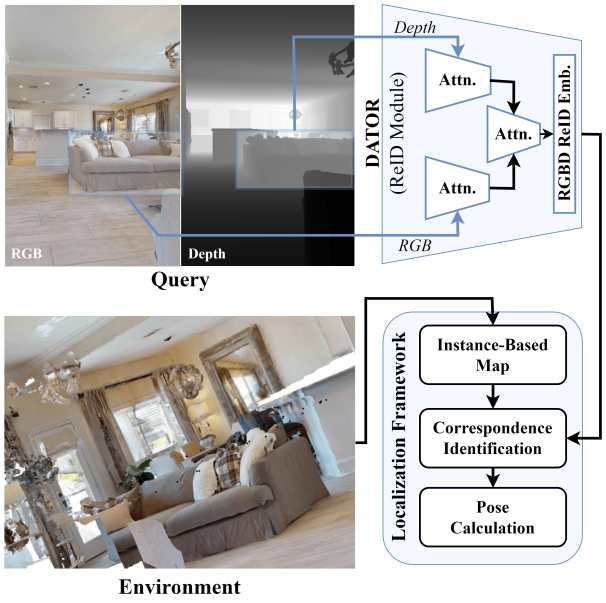

Figure 2: *Overview.* We propose a novel dual path transformer architecture, DATOR, combining cues from both RGB and depth modalities for effective object-instance ReID. Our localization framework generates an instance based map and uses our ReID model in conjunction with it to localize unseen views.

rights confirmation email (AIR '25). ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Objects in an environment can serve as important landmarks and offer significant cues for spatial awareness and orientation. They provide valuable information for understanding both an agent's general location and its precise orientation. However, the reliable re-identification of objects — formally known as the object-instance re-identification task — remains underexplored, particularly in the context of robotics.

Object-instance re-identification (ReID), often referred to simply as object ReID, is the task of reliably recognizing and matching identical instances of an object across different perspectives and environmental conditions. For example, in a warehouse setting, object ReID could be used to track the same piece of equipment across multiple camera views, even if the lighting or the equipment's position changes. While extensive research has been conducted on ReID for specific categories such as people [43, 56, 58] and vehicles [3, 23, 63], which often leverage domain-specific features like gait patterns or vehicle parameters, the broader domain of object ReID presents unique challenges. Objects vary widely in structure, appearance, and type, lacking a common unifying feature. Foundational models such as DINOv2 [30] and vision-language models like CLIP [36] provide general classification into broad categories but fall short in re-identifying specific instances within these categories. Their ability to generalize to new scenes does not offer the fine-grained recognition needed for precise object-based applications.

In robotics, the ability to accurately re-identify objects can be widely utilized for various tasks. Global relocalization, in particular, is a critical application where accurate object ReID can significantly enhance performance. This task is especially challenging in environments with repetitive scenes or numerous objects and rooms, where both local and global registration difficulties are common. Traditional global relocalization approaches often rely on aligning entire point clouds [5, 9] or extensive collections of images [2, 59] to maximize available information. However, a significant portion of this information may be redundant or not informative for effective localization.

To tackle these challenges, we introduce a Dual Path Attention Transformer for Object Re-identification (DATOR), a deep object-ReID model that leverages RGB and depth sensors commonly used with mobile robots. DATOR employs a dual-path transformer architecture to significantly enhance its ReID capabilities across multiple views. This architecture extracts and refines features from both modalities, integrating them to produce a robust final embedding. By effectively combining these modalities, DATOR ensures high accuracy in object-ReID, maintaining performance across varying illumination conditions and diverse environmental settings.

Building on the fine-grained ReID capabilities of DATOR, we introduce an object-instance based global localization framework. (Fig. 1, 2) This framework operates effectively in diverse indoor environments without requiring manual object annotation. Drawing inspiration from human navigation in familiar environments, our method constructs an instance-based map by mapping visible objects, aligning with principles similar to those discussed in [28, 57]. We catalog and encode objects using our ReID model, preserving visual and structural information, while maintaining positional data through point clouds of individual objects. For localization, we process query RGB-D views to detect and match visible objects with those in our object map, optimizing alignment for accurate localization.

To validate our framework, we provide a real-world dataset from a large, object-rich laboratory environment with multiple instances per object class (e.g., tables, chairs), presenting a challenging ReID scenario. Additionally, we provide a real as well a synthetic indoor dataset with multiple objects of myriad classes for benchmarking global localization. We also benchmark our approach against sequences from the TUM [42] RGB-D datasets.

In conclusion, we make the following contributions:

- A multimodal RGB-D object-instance ReID model (DATOR) achieving an mAP of 75.18, higher than other SoTA models.
- An object-instance ReID-based global localization framework, without manual annotation, for high accuracy in indoor environments, successfully localising 83.01% of the time on dense, publicly available datasets.
- A comprehensive object-instance ReID dataset with multiple indoor object instances under varying lighting conditions.
- A real as well as a synthetic dataset for benchmarking global localization in complex indoor environments.

2 Background

Re-Identification (ReID). ReID is a computer vision task that focuses on the recognition and matching of specific objects or

individuals across varying contexts and multiple camera views, *distinguishing it from general recognition and classification tasks* that typically involve identifying or categorizing objects [55]. Extensive research has been done in person [8, 21, 24, 38, 43, 56, 58] and vehicle ReID [3, 23, 35, 61, 63] to achieve near human performance on multiple datasets. Further, existing ReID methods also dive into the RGB-D and crossmodal domains [21, 25, 27, 46] as well as ReID for animals and buildings [1, 16, 53].

To the best of our knowledge, existing works that attempt to tackle object-instance ReID either use only one modality [28, 45, 57], benchmark against person/vehicle ReID datasets rather than indoor objects [20, 37, 62], focus on identifying evolving descriptions dependent on nearby objects [17] or operate on LiDAR scans [34].

Foundational image models. Foundational models refer to a category of models consisting of a very large number of parameters trained on a large and diverse variety of datasets for a particular task. Our framework uses three particular foundational models, Recognise Anything (RAM) [60], Segment Anything (SAM) [19] and Grounding DINO [26]. RAM is a captioning model that outputs a list of captions describing objects present in an input image. SAM is a general purpose segmentation model that can generate precise segmentation masks in challenging conditions given an image. DINOv2 [30] is all-purpose vision model trained at scale, and its variant Grounding DINO generates tight bounding boxes around an object.

Localization. Localization is a well-studied problem in robotics with diverse solutions, including one-step methods [7, 31] and multi-step approaches [49]. Methods typically depend on scene as well as object recognition [7, 12, 22, 32, 51]. There are also approaches tailored for outdoor environments, which often deal with large-scale and diverse settings [15, 34], while indoor environments require more precise methods [31, 54]. More novel approaches to localization have begun using neural fields [4, 44].

In contrast to the above methods, ReID-based localization focuses on matching individual object instances, disregarding the broader scene or surrounding objects. For example, [28] effectively demonstrates the capabilities of vision-language models (VLMs) like CLIP for object ReID, but struggles with scalability due to the use of generalized embeddings and the need for manual dataset annotation. On the other hand, [57] is better suited for large-scale datasets but treats its environment as a collection of prior scans rather than a unified map, and relies solely on depth data, limiting its robustness in more complex scenarios.

3 Approach

3.1 Dual Path Transformer Architecture

Network Architecture: We propose a novel architecture for utilizing information from both the RGB and depth modalities (Fig. 3). The network has an RGB pathway and a depth pathway, taking an RGB and a depth image as input respectively. Within the network, information is exchanged between both the pathways, and finally features from both the pathways are combined to give a final embedding.

Specifically, the RGB and depth images are first input to their respective backbone networks to extract features $f^{\text{enc}}_{\text{depth}}, f^{\text{enc}}_{\text{RGB}}$ of size $(H \times W \times E)$, where H, W represent the spatial dimensions and E is the hidden dimension of the model. We use a ViT pretrained on ImageNET [39] as the backbone for both the modalities. Additionally, we augment the last 2 encoder layers of both backbone ViTs with LoRA adapters [14].

These features $(f^{\text{enc}}_{\text{depth}}, f^{\text{enc}}_{\text{RGB}})$ are further refined through specially designed attention modules. We use deformable attention [52, 64] for our attention modules, an attention scheme that has been shown to address several limitations of using standard attention [47]. Unlike in standard attention, where the final attention outputs are produced by a weighted sum of the *entire* value feature map, in deformable attention, each query feature *selects* a fixed number of positions (K) on the value feature map, and only those elements are used to calculate attention scores. For further details on deformable attention, we refer the reader to [64].

We linearly project RGB features $f^{\text{enc}}_{\text{RGB}}$ to compute queries q_R, and values v_R. Similarly, depth features $f^{\text{enc}}_{\text{depth}}$ are projected to obtain queries q_D and values v_D. These query and value vectors are used for all subsequent attention calculations.

The RGB and the depth pathways undergo the following transformations:

$$f_{\text{R}} = f^{\text{enc}}_{\text{RGB}} + \text{Attn}_{\text{R2R}}(q_R, v_R) + \text{Attn}_{\text{D2R}}(q_D, v_R)$$

$$f_{\text{D}} = f^{\text{enc}}_{\text{depth}} + \text{Attn}_{\text{D2D}}(q_D, v_D) + \text{Attn}_{\text{R2D}}(q_R, v_D)$$

where $f_{\text{R}}, f_{\text{D}}$ are of size $(H \times W \times E)$. Finally, we perform a learned weighted sum of features as

$$f_{\text{combined}} = \alpha * f_{\text{R}} + (1 - \alpha) * f_{\text{D}}$$

with matrix α of size $(H \times W)$ effectively encoding the importance of each modality at each position on the feature map when making the final prediction. α is learnt by a CNN directly from the encoder feature representations $f^{\text{enc}}_{\text{RGB}}$ and $f^{\text{enc}}_{\text{depth}}$ (see Fig. 3 – Weighing Network). Finally, global average pooling is performed on this feature map to obtain the final embedding, a vector of dimension E.

Loss Function: Similar to earlier works [13, 29], we use a cross entropy loss and a triplet loss, and the final loss function is given as $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{triplet}}$.

Modality Dropout: When training both the modalities, it is possible that one of the modality dominates the other. This may result in collapse for one of the two pathways. To prevent this, we employ *modality dropout* [41] during training, where for each training sample, we randomly zero-out features from one of the two modalities. The dropout is represented as:

$$f^{\text{enc}}_{\text{RGB}}, f^{\text{enc}}_{\text{depth}} = \begin{cases} f^{\text{enc}}_{\text{RGB}}, 0 & \text{with } p_{\text{RGB}} \\ 0, f^{\text{enc}}_{\text{depth}} & \text{with } p_{\text{depth}} \\ f^{\text{enc}}_{\text{RGB}}, f^{\text{enc}}_{\text{depth}} & \text{with } 1 - p_{\text{RGB}} - p_{\text{depth}} \end{cases}$$

3.2 Localization Framework

Our localization framework (Fig. 1) aims to first build an object-instance based map of the environment, encoding information from each object it sees into separate reidentifiable units of information using embeddings from our ReID model. Then, when given an RGB-D image, it consults this map to find correspondences between visible objects and objects in it's memory. These correspondences are used to determine the RGB-D image's pose.

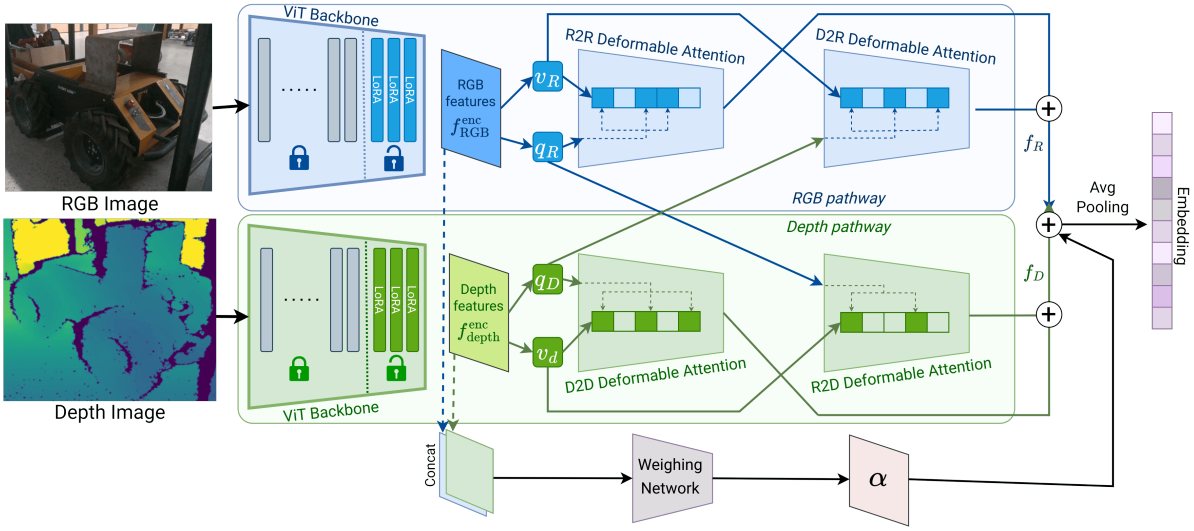

Figure 3: *Proposed model DATOR*: The model can take paired RGB and depth images of an object, and utilize cues from both the modalities to give an embedding which can be used for object ReID.

Memory formation. Given a sequence of n posed RGB-D images, $\{(I_i, D_i, t_i)\}_{i=1}^{n}$, with I_i, D_i, and t_i representing the RGB image, depth image, and pose respectively, we build an object map $\mathcal{M} = \{O_i\}_{i=1}^{n_{\text{objects}}}$, where each object O_i is stored as a tuple of its point cloud and embeddings (*object info tuple*). RAM [60] outputs captions c_i, Grounding DINO [26] generates bounding boxes b_i, and SAM [19] produces segmentation masks m_i. *Filtering* removes captions that do not represent objects directly, like adjectives ("dark", "industrial", "wooden") or descriptions of the whole scene itself ("living room", "workspace").

$$\{c_1, \ldots, c_p\} = \text{Filtering}(\text{RAM}(I_i))$$
$$\{b_1, \ldots, b_j\} = \text{unique}(\text{GDINO}(I_i, \{c_1, \ldots, c_p\}))$$
$$m_k = \text{SAM}(I_i, b_k) \text{ for } 1 \le k \le j$$

Using the RGB-D pair (I_i, D_i), camera matrices K, world-frame transformation T_i, and segmentation mask m_k, we backproject each object and form its object info tuple:

$$\mathcal{P}_k^{\text{global}} = f(I_i, D_i, m_k, K, T_i) \text{ for } 1 \le k \le j$$
$$e_k = \text{ReID}(I_i, D_i, b_k) \text{ for } 1 \le k \le j$$
$$O_k = (\mathcal{P}_k^{\text{global}}, [e_k]) \text{ for } 1 \le k \le j$$

We repeat this for all RGB-D pairs in the sequence, generating the object memory from the object info tuples of detected objects in each image.

Memory consolidation and post-processing. In many cases, objects in a scene may be represented by multiple object information tuples after processing a sequence We can improve registration accuracy during localization by grouping together object tuples that belong to the same object to form more complete pointclouds and more representative sets of embeddings. To address this, we perform a postprocessing step that clusters object info tuples based on semantic similarity and spatial proximity. When combining n tuples, their point clouds are combined, and all associated sets of embeddings are combined. We refer to this as *grouping* tuples henceforth.

$$O_i + O_j := \{\mathcal{P}_i \oplus \mathcal{P}_j, [e_{i_1}, \ldots, e_{i_m}, e_{j_1}, \ldots, e_{j_n}]\}$$

Algorithm 1 Object info tuple clustering

$\mathcal{M} \leftarrow \{O_1, \ldots, O_n\}$
$\epsilon_{\text{IoU}} \leftarrow 0.25$
$\epsilon_{L_2} \leftarrow 0.5$
Let IoU $\in \mathbb{R}^{n \times n}$
for all (O_i, O_j) **in** $\mathcal{M} \times \mathcal{M}$ **do**
  IoU$[i][j]$ = GetIoU(O_i, O_j)
end for
$\{\mathcal{L}_1, \ldots, \mathcal{L}_n\} \leftarrow$ AggClustering$(\mathcal{M}, \text{IoU}, \epsilon_{\text{IoU}})$
for k **in unique values in** $\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}$ **do**
  $O'_k \leftarrow \sum_{i=1}^{n} O_i$ if $\mathcal{L}_i = k$
end for
$\mathcal{M} \leftarrow \{O'_1, \ldots, O'_m\}$
Let PairwiseDist $\in \mathbb{R}^{m \times m}$
for all (O'_i, O'_j) **in** $\mathcal{M} \times \mathcal{M}$ **do**
  PairwiseDist$[i][j]$ = GetL2Distance(O'_i, O'_j)
end for
$\{\mathcal{L}'_1, \ldots, \mathcal{L}'_m\} \leftarrow$ AggClustering$(\mathcal{M}, \text{PairwiseDist}, \epsilon_{L_2})$
for k **in unique values in** $\{\mathcal{L}'_1, \ldots, \mathcal{L}'_m\}$ **do**
  $\{d_1, \ldots, d_a\} \leftarrow$ DBSCAN$(\{O'_i \mid \mathcal{L}_i = k\})$
  for l **in unique values in** $\{d_1, \ldots, d_p\}$ **do**
    $O''_{k,l} \leftarrow \sum_{i=0}^{m} O'_i$ if $\mathcal{L}_i = k, d_i = l$
  end for
end for
$\mathcal{M} \leftarrow \{O''_{1,1}, \ldots, O''_{m,l}\}$
return $\mathcal{M}$

Clustering occurs in stages (Described in Algorithm 1) to prevent errors from single-step grouping. Using just mean embedding-based

clustering ignores positional data, potentially grouping distant objects together. To fix this, we apply DBSCAN [10] within each semantic cluster, ensuring object-instances are grouped based on both proximity in 3D space and embedding similarity. By performing clustering, each object has a more complete set of structural information, leading to more accurate point-feature generation and improved localization.

Localization. We begin by applying the RAM-Grounding DINO-SAM pipeline used during memory formation. By doing so, we obtain an object info tuple for each object.

Since all partial point clouds are backprojected from the same RGB-D image, their relative positions remain unchanged after any rigid transformation. Therefore, the rigid transform aligning detected point clouds with those in object memory also localizes the RGB-D image in the global frame. We then seek the most accurate assignment between detected objects and those in memory using ReID embeddings, comparing assignments based on the $L2$ norm between ReID embeddings of detected objects from those in memory.

We score an assignment by the product of embedding distances between correspondences and select the k lowest scores to avoid ICP failure from poor initialization. Using subsets of detected objects (containing at least 3 detections), we compute assignments to efficiently determine localization. Each assignment provides a transform that aligns detected objects with those in object memory. We combine point clouds from detected and memory objects into single *detected* and *memory* point clouds, respectively, and register them using RANSAC [11] followed by colored ICP. We enhance this process with custom features, including FPFH [40] and one-hot encoded object indices, to prioritize matching between corresponding points. The quality of each pose is evaluated by the overlap between transformed detected point clouds and memory point clouds, with the best assignment being the one with the highest overlap.

Dataset	Sequence Count	Mean seq. length	No. of classes	Mean instances per class
DATOR-lab	1	1703	19	10
DATOR-synth	6	3766	9	5
TUM	5	1331	29	2-3

Table 1: *Localization Dataset Metrics*

3.3 Dataset Generation

We present *DATOR-ReID*, a new, real-world, object re-identification dataset of RGB-D images designed to benchmark our object ReID model. Captured with an Intel RealSense [18] camera, it includes 8 types of objects, such as chairs and tables, with upto 5 distinct instances of each object type. We use approximately 60 images per instance under varying environmental conditions.

We also present two of our datasets, *DATOR-lab* and *DATOR-synth*, designed to evaluate our localization system. *DATOR-lab* is a real-life localization dataset consisting collected on a P3DX robot in a large indoor laboratory. The RGB-D images were collected using an Intel RealSense camera and P3DX's wheel odometry. *DATOR-synth*, is a set of sequences generated using the ProcTHOR [6] API in a multi-room environment with various items of furniture. Both datasets feature a large variety of objects and multiple instances of visually similar structures, resulting in a challenging localization benchmark. We note that *DATOR-lab* and *DATOR-ReID* were recorded in different locations. All of our data has been made publicly available [b].

Table 1 collates metrics about the variety and number of objects in the each of the datasets we use in our localization experiments.

4 Experimental Setup

Our ReID module training as well as the localization pipeline requires only 12 GB VRAM, allowing it to run on commercially available GPUs. Localization tests were conducted on an NVIDIA RTX A4000 GPU and an AMD Ryzen 9 7950X 16-Core Processor.

Object ReID: DATOR was trained for 240 epochs, using SGD optimizer and cosine LR scheduler, with an initial learning rate of 0.008. A batch size of 64 was used for all the experiments.

Localization: We evaluate our localization pipeline on *DATOR-lab*, 6 *DATOR-synth* sequences, as well as 5 TUM RGB-D sequences, (namely *fr{_desk,_desk2,_room}*, *freiburg2_desk* and *freiburg3 _long_office_household*).

For *DATOR-synth*, we create an object memory for each sequence by sampling RGB-D pairs every 30 frames. Additional unseen RGB-D pairs are sampled every 30 frames with a 15-frame offset. The only information available during localization is the generated object memory. Each frame is individually localized using the sequence's object memory. Sampling strategies for other datasets are detailed in our released implementation.

We compare each RGB-D pair's estimated pose with its corresponding ground truth pose, measuring 3D translation error (TE) in meters and rotation error (RE) in radians. An RGB-D pair is said to be correctly localized, if the TE and RE are below certain thresholds (see caption under table 2). We report the success rates for various baselines along with our method for comparison.

The implementations and annotation tools used for [57] and [28] are not available. We release a re-implementation of [28] but observe suboptimal results using this approach (elaborated in section 5).

5 Results

Object ReID Results

We benchmark DATOR against some popular vehicle-ReID and person-ReID baselines, the results are presented in table 3. We train all the RGB-only baselines on the RGB images in the *DATOR-ReID* dataset. To evaluate the effectiveness of each individual pathway in DATOR, we also show RGB-only inference results, in which the depth features are zeroed (similar to modality dropout) and depth-only inference results, wherein the RGB features are zeroed. Note that DATOR achieves competitive results even when using only one of the modalities, while at the same time achieving much better performance (last row in table 3) when both the modalities are used. This shows that there is effective exchange of information between both the pathways in the network.

Further, we show a qualitative result in Fig. 4 where DATOR re-identifies the correct instance in a low-illumination setting while

[b] https://github.com/instance-based-loc/instance-based-loc

Dataset	ReID Backbone	Avg Errors (m, rad)	Median Errors (m, rad)	Success Rate (%)
TUM RGB-D	**CLIP**	1.83/0.9479	1.60/1.121	12.02
	DINOv2	1.11/0.629	0.85/0.637	41.69
	ViT-b	1.32/0.720	1.18/0.788	24.70
	DATOR	**0.50/0.252**	**0.29/0.028**	**83.01**
DATOR-lab	**CLIP**	4.99/1.002	4.74/1.211	0.00
	DINOv2	4.96/0.912	4.69/1.01	3.38
	ViT-b	5.28/0.602	4.74/0.322	0.00
	DATOR	**2.25/0.373**	**1.43/0.024**	**41.18**
DATOR-synth	**CLIP**	4.03/0.327	2.26/0.00176	33.33
	DINOv2	3.66/**0.192**	3.09/**0.00099**	15.38
	ViT-b	5.11/0.341	4.270.0017	12.82
	DATOR	**2.98** / 0.4552	**0.93** / 0.0976	**47.05**

Table 2: *Localization results across ReID techniques.* Avg. and Median Errors have been represented as (Translation Error/Rotation Error) in (m/rad) units. We consider a success as a pose prediction within 0.6m of translation error and 0.3 radians of rotation error. Our results show that DATOR outperforms other models by a significant margin across multiple datasets. Best results are in bold and second-best in underlined text.

Method	mAP Score
CLIP-ReiD (CNN) [25]	54.1
TransReID [13]	55.7
LoGoViT [33]	60.2
CLIP-ReID (ViT) [25]	61.2
NFormer [48]	63.2
PADE [50]	66.0
DATOR (RGB Inference)	61.41
DATOR (Depth Inference)	64.11
DATOR (Full)	**75.18**

Table 3: *ReID mAP Metrics.* We benchmark DATOR against several ReID methods trained our *DATOR-ReID* dataset. We demonstrate the effectiveness of using both modalities for object ReID. Best result is in bold and second-best is underlined.

the existing best performing method, PADE [50], retrieves a similar but incorrect instance.

Localization Results

Table 2 summarizes the effectiveness of each model in our localization architecture. We include both mean and median errors to illustrate the effect of high error misalignments as compared to successful alignments, that are near perfect predictions of pose. The large jump in success rate from other baselines to DATOR points to more successful instance ReID leading to more accurate registration. Our results demonstrate the challenging nature of our datasets and the potential improvements left to future works. A common trend observed across all models is the negative effect of large, flat, texture-less walls that interfere with ICP matching. For example, unlike the other TUM sequences, *fr1_room* dataset from TUM has large, textureless, flat objects that result in low success rates for all benchmarked methods.

In comparison, results shown in [28] are only for two TUM sequences (*fr2_desk*, *fr3_lo__household.*) achieving close to only 80% while we achieve success rates of 88.5% and 100% respectively. This is despite of us employing stricter margins and also considering rotation error (See Fig. 4 in [28]).

Figure 4: *Qualitative Analysis of DATOR.* Given a query in a low-illumination scene, DATOR reidentifies the robot instance successfully, while PADE, identifies it as a different robot that is missing an overhead attachment. This gain can be attributed to DATOR's use of depth information.

6 Conclusion

We introduce an object-based localization framework that generalizes across diverse indoor environments without the need for manual annotations, marking a significant step forward in autonomous indoor navigation. Our approach demonstrates accurate and robust localization in both real-world and synthetic environments. Additionally, our ReID architecture achieves a high mean Average Precision (mAP) of 75.18 on a challenging dataset with varying illumination, demonstrating its adaptability to real-world scenarios. The embeddings generated by our model allow for more accurate object-instance re-identification localization success rate of other large-scale image encoding models, localizing successfully in 83.01% of the cases averaged across multiple sequences of TUM-RGBD. Moreover, we release a challenging, object-rich pair of real and synthetic relocalization datasets, as well as an object ReID dataset featuring varying illumination conditions.

Future work includes exploring expanding our framework to effectively recognise significant non-object landmarks, function in outdoor environments, enhancing its robustness to more extreme variations in lighting and occlusions and integrating it into mobile robotics pipelines.

References

[1] Adam, L., Čermák, V., Papafitsoros, K., and Picek, L. Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024).

[2] Ali-Bey, A., Chaib-Draa, B., and Giguere, P. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (2023).

[3] Amiri, A., Kaya, A., and Keceli, A. S. A comprehensive survey on deep-learning-based vehicle re-identification: Models, data sets and challenges. *arXiv preprint arXiv:2401.10643* (2024).

[4] Aoki, K., Koide, K., Oishi, S., Yokozuka, M., Banno, A., and Meguro, J. 3d-bbs: Global localization for 3d point cloud scan matching using branch-and-bound algorithm. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (2024), IEEE.

[5] Besl, P. J., and McKay, N. D. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures* (1992), vol. 1611, Spie.

[6] Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Salvador, J., Ehsani, K., Han, W., Kolve, E., Farhadi, A., Kembhavi, A., and Mottaghi, R. Procthor: Large-scale embodied ai using procedural generation, 2022.

[7] Dong, S., Wang, S., Zhuang, Y., Kannala, J., Pollefeys, M., and Chen, B. Visual localization via few-shot scene region classification. In *2022 International Conference on 3D Vision (3DV)* (2022), IEEE, pp. 393–402.

[8] Dou, Z., Wang, Z., Li, Y., and Wang, S. Identity-seeking self-supervised representation learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023).

[9] Elbaz, G., Avraham, T., and Fischer, A. 3d point cloud registration for localization using a deep neural network auto-encoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4631–4640.

[10] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining* (1996).

[11] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM 24*, 6 (1981).

[12] Gard, N., Hilsmann, A., and Eisert, P. Spvloc: Semantic panoramic viewport matching for 6d camera localization in unseen environments. *arXiv preprint arXiv:2404.10527* (2024).

[13] He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. Transreid: Transformer-based object re-identification, 2021.

[14] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

[15] Humenberger, M., Cabon, Y., Pion, N., Weinzaepfel, P., Lee, D., Guérin, N., Sattler, T., and Csurka, G. Investigating the role of image retrieval for visual localization: An exhaustive benchmark. *International Journal of Computer Vision 130*, 7 (2022), 1811–1836.

[16] Jiao, B., Liu, L., Gao, L., Wu, R., Lin, G., Wang, P., and Zhang, Y. Toward re-identifying any animal. *Advances in Neural Information Processing Systems 36* (2024).

[17] Keetha, N. V., Wang, C., Qiu, Y., Xu, K., and Scherer, S. Airobject: A temporally evolving graph embedding for object identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[18] Keselman, L., Woodfill, J. I., Grunnet-Jepsen, A., and Bhowmik, A. Intel realsense stereoscopic depth cameras, 2017.

[19] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv:2304.02643* (2023).

[20] Lee, H., Eum, S., and Kwon, H. Negative samples are at large: Leveraging hard-distance elastic loss for re-identification. In *European Conference on Computer Vision* (2022), Springer.

[21] Lejbolle, A. R., Nasrollahi, K., Krogh, B., and Moeslund, T. B. Multimodal neural network for overhead person re-identification. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)* (2017).

[22] Li, F., Gao, D., Huang, Q., Li, W., and Yang, Y. Magiccubepose, a more comprehensive 6d pose estimation network. *Scientific Reports 13*, 1 (2023), 6923.

[23] Li, H., Chen, J., Zheng, A., Wu, Y., and Luo, Y. Day-night cross-domain vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 12626–12635.

[24] Li, H., Ye, M., Wang, C., and Du, B. Pyramidal transformer with conv-patchify for person re-identification.

[25] Li, S., Sun, L., and Li, Q. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels, 2023.

[26] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024.

[27] Martini, M., Paolanti, M., and Frontoni, E. Open-world person re-identification with rgbd camera in top-view configuration for retail applications. *IEEE Access 8* (2020).

[28] Matsuzaki, S., Sugino, T., Tanaka, K., Sha, Z., Nakaoka, S., Yoshizawa, S., and Shintani, K. Clip-loc: Multi-modal landmark association for global localization in object-based maps. *International Conference on Robotics and Automation (ICRA)* (2024).

[29] Ning, E., Wang, C., Zhang, H., Ning, X., and Tiwari, P. Occluded person re-identification with deep learning: a survey and perspectives. *Expert Systems with Applications* (2023), 122419.

[30] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2024.

[31] Panek, V., Kukelova, Z., and Sattler, T. Meshloc: Mesh-based visual localization. In *European Conference on Computer Vision* (2022), Springer.

[32] Peng, Y., Zhai, Z., and Feng, M. Slmsf-net: A semantic localization and multi-scale fusion network for rgb-d salient object detection. *Sensors 24*, 4 (2024), 1117.

[33] Phan, N., Huy, T. D., Duong, S. T. M., Hoang, N. T., Tran, S., Hung, D. H., Nguyen, C. D. T., Bui, T., and Truong, S. Q. H. Logovit: Local-global vision transformer for object re-identification. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023).

[34] Pramatarov, G., Gadd, M., Newman, P., and De Martini, D. That's my point: Compact object-centric lidar pose estimation for large-scale outdoor localisation. *arXiv preprint arXiv:2403.04755* (2024).

[35] Qian, W., Luo, H., Peng, S., Wang, F., Chen, C., and Li, H. Unstructured feature decoupling for vehicle re-identification. In *European Conference on Computer Vision* (2022), Springer, pp. 336–353.

[36] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

[37] Rao, Y., Chen, G., Lu, J., and Zhou, J. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021).

[38] Ren, L., Lu, J., Feng, J., and Zhou, J. Uniform and variational deep learning for rgb-d object recognition and person re-identification. *IEEE Transactions on Image Processing 28*, 10 (2019), 4970–4983.

[39] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge, 2015.

[40] Rusu, R. B., Blodow, N., and Beetz, M. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation* (2009), pp. 3212–3217.

[41] Shi, B., Hsu, W.-N., Lakhotia, K., and Mohamed, A. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184* (2022).

[42] Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (2012), IEEE, pp. 573–580.

[43] Tan, W., Ding, C., Jiang, J., Wang, F., Zhan, Y., and Tao, D. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024).

[44] Tang, G., Jatavallabhula, K. M., and Torralba, A. Efficient 3d instance mapping and localization with neural fields. *arXiv preprint arXiv:2403.19797* (2024).

[45] Thérien, B., Huang, C., Chow, A., and Czarnecki, K. Object re-identification from point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), pp. 8377–8388.

[46] Uddin, M. K., Lam, A., Fukuda, H., Kobayashi, Y., and Kuno, Y. Fusion in dissimilarity space for rgb-d person re-identification. *Array 12* (2021), 100089.

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

[48] Wang, H., Shen, J., Liu, Y., Gao, Y., and Gavves, E. Nformer: Robust person re-identification with neighbor transformer, 2022.

[49] Wang, S., Laskar, Z., Melekhov, I., Li, X., and Kannala, J. Continual learning for image-based camera localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 3252–3262.

[50] Wang, Z., Huang, H., Zheng, A., Li, C., and He, R. Parallel augmentation and dual enhancement for occluded person re-identification, 2024.

[51] Wen, B., Yang, W., Kautz, J., and Birchfield, S. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 17868–17879.

[52] Xia, Z., Pan, X., Song, S., Li, L. E., and Huang, G. Vision transformer with deformable attention. In *IEEE/CVF CVPR* (2022).

[53] Xue, F., Budvytis, I., Reino, D. O., and Cipolla, R. Efficient large-scale localization by global instance recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (6 2022).

[54] Yang, L., Shrestha, R., Li, W., Liu, S., Zhang, G., Cui, Z., and Tan, P. Scenesqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022).

[55] Ye, M., Chen, S., Li, C., Zheng, W.-S., Crandall, D., and Du, B. Transformer for object re-identification: A survey, 2024.

[56] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence 44*, 6 (2021), 2872–2893.

[57] Zhang, L., Digumarti, T., Tinchev, G., and Fallon, M. Instaloc: One-shot global lidar localisation in indoor environments through instance learning. *Robotics: Science and Systems* (2023).

[58] Zhang, Q., Wang, L., Patel, V. M., Xie, X., and Lai, J. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 22000–22009.

[59] Zhang, X., Wang, L., and Su, Y. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition 113* (2021), 107760.

[60] Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514* (2023).

[61] Zhao, J., Zhao, Y., Li, J., Yan, K., and Tian, Y. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 205–214.

[62] Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., and Shan, Y. Dual cross-attention learning for fine-grained visual categorization and object re-identification, 2022.

[63] Zhu, X., Luo, Z., Fu, P., and Ji, X. Voc-reid: Vehicle re-identification based on vehicle-orientation-camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 602–603.

[64] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

Received 11 February 2025; accepted 18 April 2025