Supplementary Material
SceneGraphLoc

Abstract. In the supplementary material, we provide additional information of SceneGraphLoc:

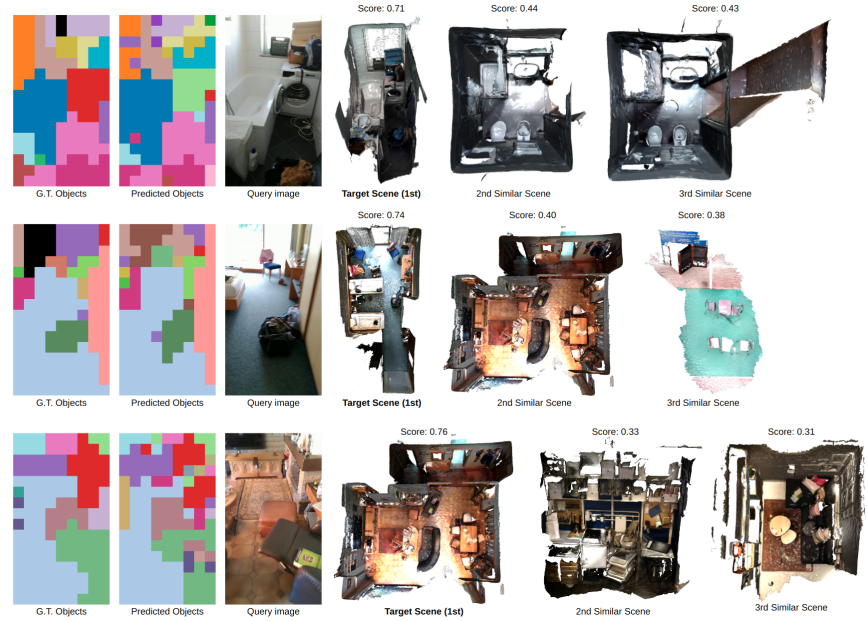
1. Qualitative results to further understand the performance of SceneGraphLoc (Section A).
2. Ablation studies and analysis of experiment results (Section B).
3. Details on implementation (Section C).

A Qualitative Results

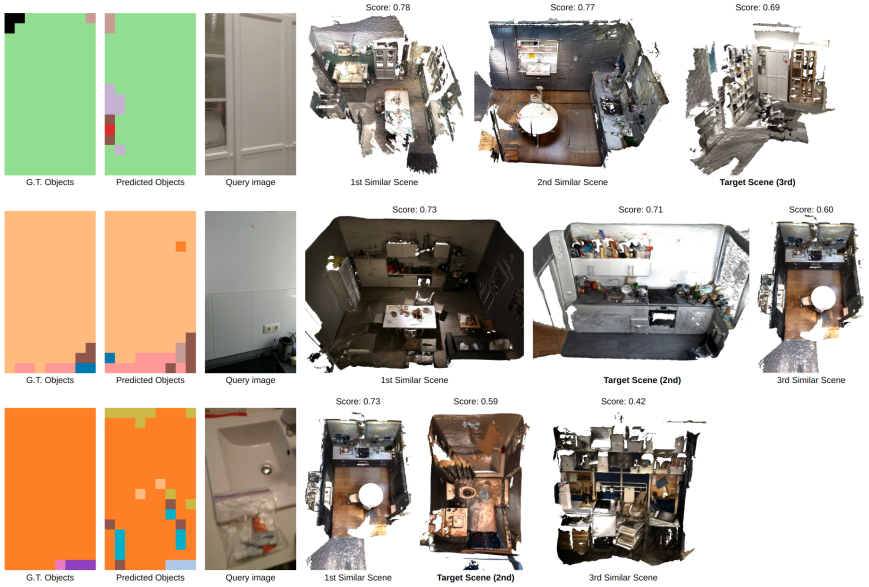
In this section, we provide additional qualitative results of successful and failed cases of room retrieval with R^t@1 in different scenarios.

In Fig. 6a, we show successful cases with the target scene ranked as No.1 in the pool of 10 candidate scenes. From the left images in Fig. 6a, we can see that the majority of the patches is correctly assigned to corresponding objects given a pool of objects within the target scene graph. Furthermore, the similarity score gap of the target scene is significantly larger than the second most similar scene, showing the effectiveness of the proposed similarity score in distinguishing the target scene and other scenes.

In Fig. 6b, we show failure cases with the target scene not ranked as No.1 in the pool of candidate scenes. Compared to Fig. 6a, we can see that the query images in Fig. 6b have a limited field of view and a limited diversity of observed distinct objects. The intuition is that the localization performance of the query image is related to the diversity of objects observed in the image: the more diverse and distinctive objects are in the query image, the easier it is for the query image can be correctly matched to the target scene, as shown in Fig. 6a. Conversely, if the query image is dominated by non-unique objects (i.e., wall), then it can be difficult to retrieve the target scene graph, as shown in Fig. 6b. The dependence on the object number will be shown in the next section. This tendency can be exploited in practice as confidence in the predicted results, assigning high confidence if many objects are seen from the query image.



(a) Successful cases of top-1 scene retrieval



(b) Failures for top-1 scene retrieval

Fig. 6: Successful and **failed** cases for scene retrieval with R^t@1. On the left are the G.T. and predicted objects of query image within the target scene graph. On the right are the top-3 retrieved scenes with their image-scene similarity scores.

Table 5: Ablation study on the methods generating image embeddings for the map.

K_{view}	R^t@1	@3	@5
1	83.9	96.6	99.4
3	84.2	96.7	99.5
5	85.4	97.4	99.5
7	86.7	97.0	99.4
10	**88.5**	97.7	99.6
15	86.3	**97.8**	99.6
20	86.8	**97.8**	**99.7**

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

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

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

Fig. 7: Shannon entropy $\mathcal{H}_I$, denoting the diversity of objects observed in the query image.

B Additional Ablation Study

B.1 Image Modality Embedding $\mathcal{I}$

SceneGraphLoc integrates multi-view image features for object embedding of image modality $\mathcal{I}$, as shown in Fig.3 in the main paper. In Table 5, we explore the impact of the number (K_{view}) of views considered when creating the multi-view embedding and the employed image fusion methods on the localization performance. Table 5a shows that by using more views for modality $\mathcal{I}$, the localization performance improves, and this trend stoping when K_{view} reaches 15 and 20. Furthermore, Table 5b shows that the localization performance benefits from the transformer encoder with positional encoding followed by max-pooling. The intuition behind positional encoding with image poses is to integrate spatial context with the multi-view visual information for more-informed visual embedding of objects within the scene graph.

B.2 Correlation between variables and the recall performance

In Table 6, we report the correlation between multiple factors and the localization performance R^t@1 and Acc_q^t under multiple settings of modalities. The following notations are defined:

– Scalar $|\mathcal{V}_0^t|$ represents the number of object nodes within the target scene graph with potential temporal changes $\mathcal{G}_I^t$.

Table 6: Statistics Analysis on the val. split of 3RScan [5], analysing the correlation between multiple factors ($|\mathcal{V}_0^t|$, $\mathcal{H}_I$ and s_I) and the performance of coarse localization (R^t@1 abbreviated as R_1^t and Acc_q^t) under multiple modalities.

Map modalities					R^t@1	$\rho_{(\lvert\mathcal{V}_0^t\rvert, R_1^t)}$	$\rho_{(\mathcal{H}_I, R_1^t)}$	$\rho_{(s_I, R_1^t)}$	Acc_q^t	$\rho_{(\lvert\mathcal{V}_0^t\rvert, Acc_q^t)}$	$\rho_{(\mathcal{H}_I, Acc_q^t)}$
$\mathcal{P}$	$\mathcal{I}$	$\mathcal{A}$	$\mathcal{S}$	$\mathcal{R}$							
✓					43.9	0.20	0.16	0.02	49.2	-0.10	0.03
✓		✓			54.8	0.21	0.19	0.07	53.8	-0.06	0.05
✓		✓	✓		56.5	0.29	0.20	0.11	55.9	-0.04	0.03
✓		✓	✓	✓	62.7	0.38	0.22	0.19	54.8	-0.07	0.06
	✓				80.2	0.15	0.06	0.19	55.6	-0.06	-0.07
✓	✓				84.7	0.21	0.19	0.20	61.1	-0.06	-0.01
✓	✓	✓	✓	✓	88.5	0.28	0.15	0.17	64.2	-0.07	-0.03

- Scalar $\mathcal{H}_I$ represents the Shannon entropy of object information observed in image patches $q \in \mathcal{Q}_I$, defined in Eq. 9.
- Scalar $s_I = s(\mathcal{G}_I^t, I)$ represents the similarity score between $\mathcal{G}_I^t$ and the query image.
- Scalar Acc_q^t represents the percentage of image patches q_I that are correctly assigned to the objects in the scene graph given Eq. 3 in the main paper.
- Scalar $\rho(a, b) \in [-1, 1]$ represents the Pearson Correlation coefficient between two variables a and b. Parameter $\rho > 0$ represents positive correlation while $\rho < 0$ means negative correlation.

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

$\mathcal{O}_I^{gt}$ is the ground truth set of objects observed in query image I and $p_I(o)$ is the frequency of patches observing the object o. Scalar $\mathcal{H}_I$ denotes the diversity of objects observed in I, as illustrated in Fig. 7.

From the table, we can see that:

- Values $\rho_{(\lvert\mathcal{V}_0^t\rvert, R_1^t)}$ and $\rho_{(\lvert\mathcal{H}_I\rvert, R_1^t)}$ are greater than 0 by a not negligible amount, denoting positive correlation between $|\mathcal{V}_0^t|$ and R_1^t, and the positive correlation between $\mathcal{H}_I$ and R_1^t. The intuition is that the more objects observed in the query image and located in the target 3D scene graph, the easier the query image can be localized. This correlation agrees with the qualitative results in Section A.
- Noticeably, with integration of modalities $\{\mathcal{S}, \mathcal{R}\}$, the correlation $\rho_{(\lvert\mathcal{V}_0^t\rvert, R_1^t)}$, $\rho_{(\mathcal{H}_I, R_1^t)}$ increases. The intuition is that by incorporating $\{\mathcal{S}, \mathcal{R}\}$, the proposed modules learn to leverage scene-context information, e.g., the relationship between objects, for object embedding and coarse localization. Thus, the localization accuracy R_1^t benefits from more context information (larger $\rho_{(\lvert\mathcal{V}_0^t\rvert, R_1^t)}$ and $\rho_{(\mathcal{H}_I, R_1^t)}$).

Table 7: Ablation study performed on the val. split of ScanNet [2] with SceneGraphLoc with ground truth 3D instance segmentation and predicted instance segmentation from [6].

Map modalities		G.T. Seg			Predicted Seg [6]		
$\mathcal{P}$	$\mathcal{I}$	R^t@1	@3	@5	R^t@1	@3	@5
✓		62.0	86.9	94.5	53.1	85.1	93.4
✓	✓	79.3	96.3	99.4	68.7	94.9	98.8

– For patch-object association accuracy Acc_q^t, all modalities except $\mathcal{R}$ have contributions to improving Acc_q^t. On the other hand, there is a slightly negative correlation between $|V_0^t|$ and Acc_q^t, denoting that the more diversity of the objects in the scene graph, the slightly harder for the image patches to be correctly assigned to certain objects. Noticeably, with integration of image modality $\mathcal{I}$, the correlation $\rho_{(\mathcal{H}_I, Acc_q^t)}$ turns from slightly positive to slightly negative, denoting that with object embedding of $\mathcal{I}$, the diversity of objects observed in the query image affects the patch-object matching accuracy.

B.3 The Impact of 3D Instance Segmentation Accuracy

In the main paper, the experiments on ScanNet [2] with predicted scene graph with [6] shows that there is a performance gap between SceneGraphLoc and the image-retrieval-based methods. One potential reason for the gap is the inaccurate instance segmentation from [6], as illustrated in Fig. 8, as the object embedding of modalities $\mathcal{P}$ and $\mathcal{I}$ requires a 3D model of each object node within the scene graph. In order to understand the impact of 3D instance segmentation accuracy in the performance, we compare the performance of SceneGraphLoc with predicted and ground truth 3D instance segmentation under the modalities of object embedding ($\mathcal{P}$ and $\mathcal{I}$). From Table 7 we can see that by using ground truth instance segmentation, the performance of SceneGraphLoc improves by a large margin, implying that the performance in Table 2 in the main paper can be potentially improved by applying more accurate 3D instance segmentation methods when creating the reference map of the environment.

B.4 Confusion Matrix

In SceneGraphLoc, each patch of the query image $q \in \mathcal{Q}_I$ is assigned to an object node $v \in \mathcal{V}_I$ in the scene graph. We compute and visualize confusion matrices of semantic categories of (q, v) pairs, as illustrated in Fig. 10. From the figure, we can see that as more modalities are integrated (from Fig. 10a to Fig. 10f), the confusion matrix is closer to the identity matrix, denoting that the patch-object matching becomes more accurate, which agrees with the trend of Acc_q^t shown in Table 6. Fig. 10f shows that with all modalities integrated, there are still objects of certain categories with non-trivial probabilities of being mismatched: (i) image patches of *counter* can be assigned to nodes of *other structure*; (ii)

(a) G.T. Instance Segmentation. **(b)** Predicted Instance Segmentation [6].

Fig. 8: Comparison of G.T. and predicted instance segmentation in ScanNet dataset [2]. The left image shows that SceneGraphFusion [6] applied in the Section 4 in the main paper can output inaccurate instance segmentation (red box) and under-reconstruction (white boxes) results.

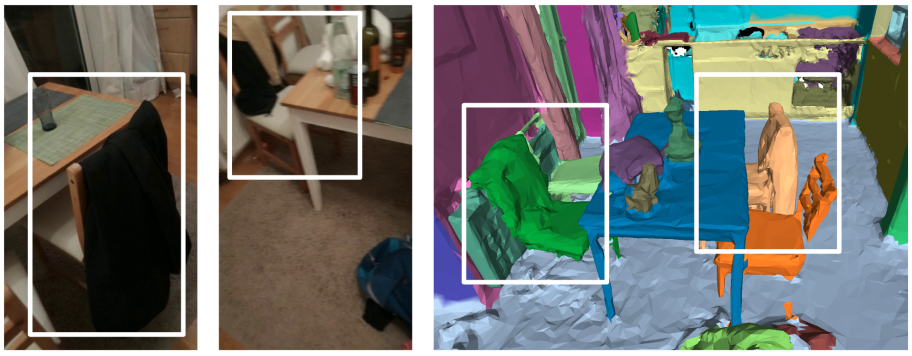


Fig. 9: The two left image show that the clothes hangs on the chair back and the right image shows the under-segmentation of the clothes and the chairs.

patches of *door* can be assigned to nodes of *wall* and (iii) patches of clothes can be assigned to nodes of *chair* due to the inaccurate instance segmentation when a cloth is hanging on the back of a chair, as shown in Fig. 9.

B.5 Patch Size

In order to further understand the impact of patch size selection in the performance, we further conduction ablation study with various patch sizes on test set of 3RScan dataset: As shown in table 8, the proposed method is robust to the patch size of the query images. With small (15px) or large (135px) patches

Table 8: Ablation study on patch size of the query image.

patch size (px)	135	90	60	30	15
R^t@1	78.4	81.6	81.5	80.2	79.3

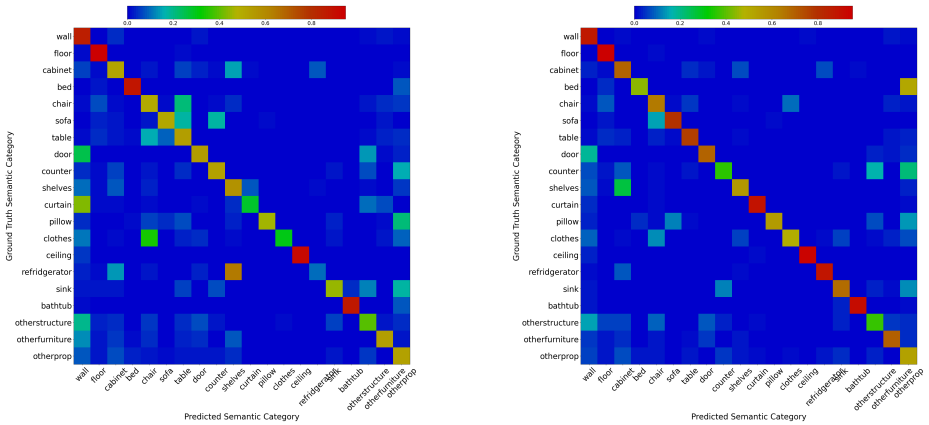
used for query image encoding, the accuracy dropped slightly from 81.5% (with patch size of 60 px) to 78.4% and 79.3%.

B.6 Sub-sampled database of images

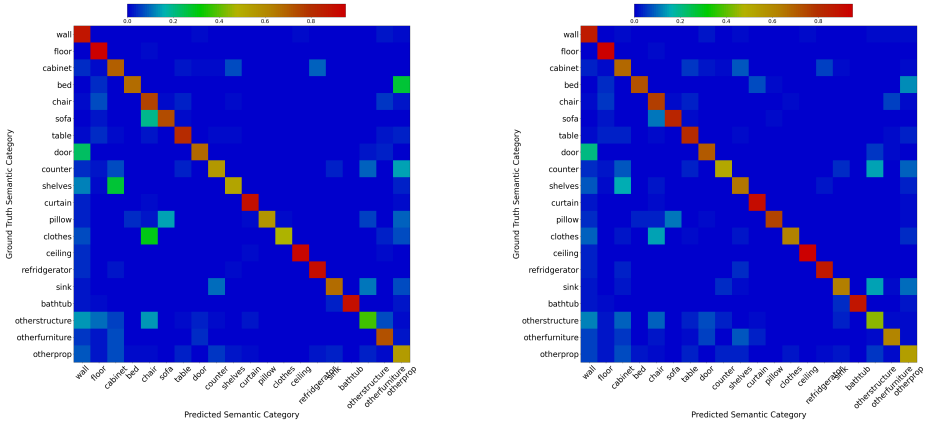
It is noticeable that there is a performance gap between the proposed method and the heavy image-based methods (*i.e.*, AnyLoc). One assumption is that image-based methods (AnyLoc) have better accuracy due to the exhaustively sampled image database, especially in the ScanNet dataset. Thus, we show comparisons against image-based methods within sub-sampled image databases (each scan preserves at least one image). Additionally, we evaluate NetVLAD [1], the traditional learning-based image-retrieval method in the proposed task. As shown in the Fig. 11, as the database is down-sampled, the gap between our method and others narrows and even reverses, while our method maintains low memory cost. Additionally, SceneGraphLoc outperforms NetVLAD in both accuracy and storage on 3RScan. On ScanNet, the performance of SceneGraphLoc can be significantly improved with better segmentation (GT in this case).

C Implementation Details

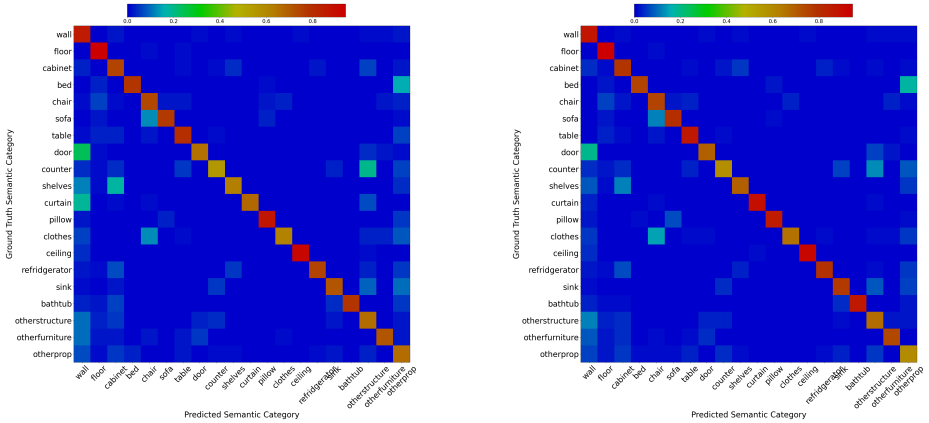
Machine. All the experiments of the room retrieval tasks during the inference phase are implemented on a machine with an Intel-12700K CPU, a Nvidia RTX3090 GPU and 64 GB RAM. For time measurement, the time t_{e_q} of encoding the query image is measured by using the GPU and the time t^N_{retr} of implementing room retrieval task is measured by using the CPU.
Models and Training. We use $L = 3, K_{view} = 10$ for multi-level multi-view image embedding of objects as depicted in Section 3.1 in the main paper. We use $\alpha = 0.5$ as the weight between static loss and temporal loss. The dimension D^k for the embedding e^v of each modality $k \in \{P, S, R, A\}$ is 100 and the dimension for image modality is 256. The dimension of unified embedding D is 400. We train SceneGraphLoc our model with a batch size of 16 using Adam [3] optimizer. Learning rate is 0.0011 with the step learning rate scheduler.
Dataset. In 3RScan dataset [5], the query images are with resolution of 960×540 pixels and are resize to 224×126 pixels before feeding into the Dino [4] backbone, which then extract 16×9 patches features from the image. In ScanNet dataset [2], images of 1296×968 are resized to 448×338 and 24×32 patch features are extracted.



(a) $\mathcal{P}$ **(b)** $\mathcal{P} + \mathcal{A}$

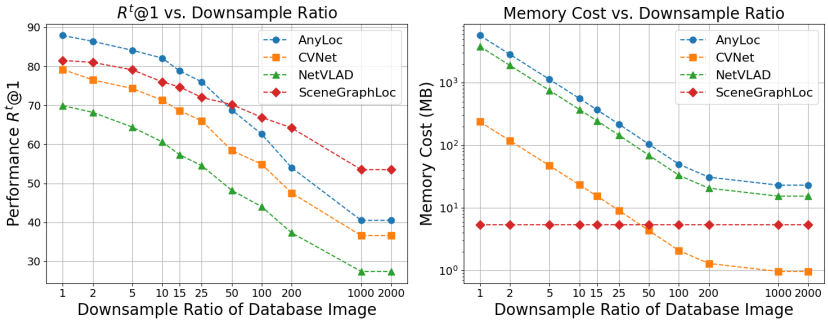


(c) $\mathcal{P} + \mathcal{A} + \mathcal{S}$ **(d)** $\mathcal{P} + \mathcal{A} + \mathcal{S} + \mathcal{R}$

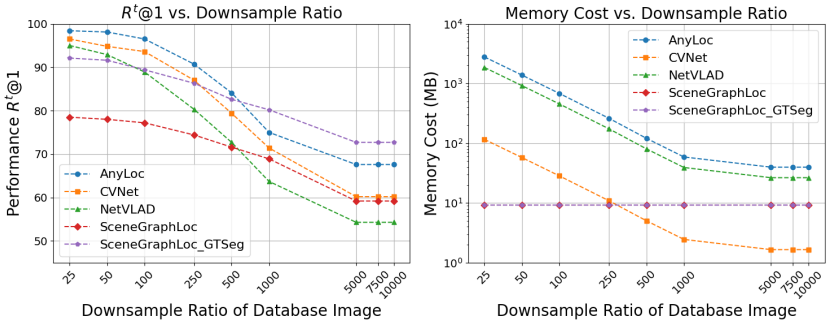


(e) $\mathcal{P} + \mathcal{I}$ **(f)** $\mathcal{P} + \mathcal{I} + \mathcal{A} + \mathcal{S} + \mathcal{R}$

Fig. 10: Confusion Matrices of 6 modality combinations of SceneGraphLoc. The y-axis represents the ground truth semantic category of the image patch q of query image and the x-axis represents the semantic category of the object note v in scene graph matched to q. Ideally, the confusion matrix should be the identity matrix.



(a) 3RScan - Test Set



(b) ScanNet - Test Set

Fig. 11: Ablation study on down-sampled database of images.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
5. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: International Conference on Computer Vision (ICCV) (2019)
6. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2021)