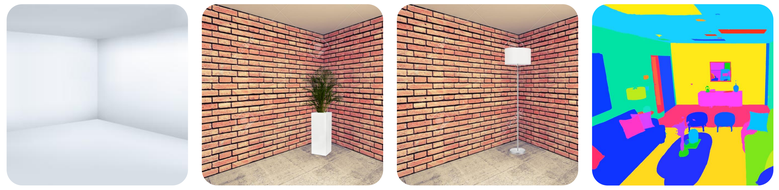
Spatial Perception by Object-Aware Visual Scene Representation

Chung-Yeon Lee[1,2], Hyundo Lee[1], Injune Hwang[1], Byoung-Tak Zhang[1,2]

Seoul National University[1], Surromind Robotics[2]

Motivation

- Visual SLAM based on geometric representations is good, except some bad conditions of the environment.
- Lack of reliable descriptions → Fail to relocalization
- Using object labels to augment the scene representation



Visual Perception Framework

Semantically-Augmented Scene Representation

$$\mathbb{L}_f := \{\langle l, \omega\rangle \,|\, l \in L_f,\ \omega = \gamma |P_f^l|\} \cup \{\langle v, \omega\rangle \,|\, v \in B_f,\ \omega = \eta_v |P_f^v|\}, \quad (1)$$

Improved Similarity Measure of Descriptors

$$dist(p,q) := \delta_{pq} \cdot \|p,q\|_n, \quad (2)$$

$$\delta_{pq} := \begin{cases} 1 & \text{if } l_p = l_q \\ \xi & o.w. \end{cases}$$

Efficient Selection Algorithm for Keyframe Candidates

$$C_f^1 := \left\{ \langle k, n\rangle \,|\, k \in KF,\ L_f \cap L_k \neq \emptyset \right\} \quad (3)$$

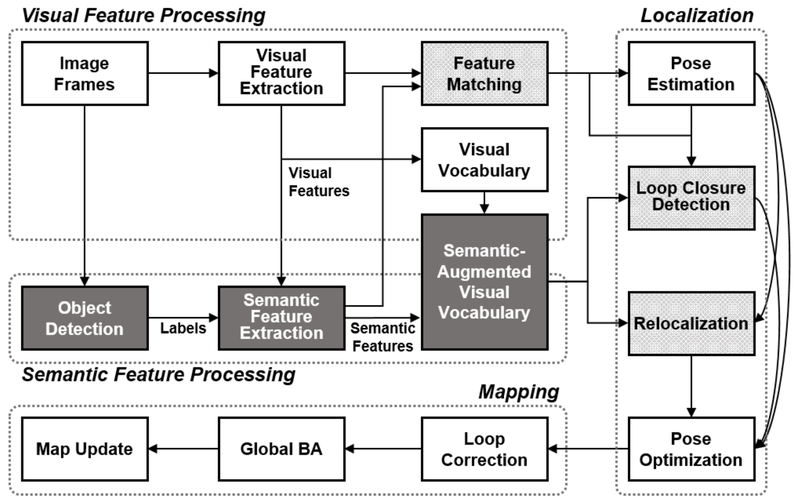
$$n = |L_f \cap L_k|$$

$$C_f^2 := \left\{ k \,|\, \langle k, m\rangle \in C_f^1,\ m > \alpha \cdot \max_{\langle k,n\rangle \in C_f^1}(n) \right\} \quad (4)$$

$$C_f^3 := \left\{ \langle k, a\rangle \,|\, k_c \in C_f^2,\ k = \underset{k_c'}{argmax}(\|\mathbb{L}_f, \mathbb{L}_{k_c'}\|_1) \right\} \quad (5)$$

$$a = \textstyle\sum_{k_c'} \|\mathbb{L}_f, \mathbb{L}_{k_c'}\|_1$$

$$C_f := \left\{ k \,|\, \langle k, b\rangle \in C_f^3,\ b > \beta \cdot \max_{\langle k,a\rangle \in C_f^3}(a) \right\} \quad (6)$$

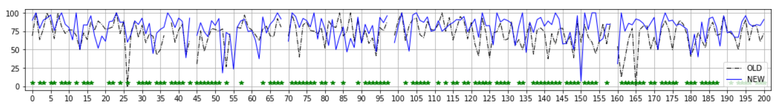


Experimental Results

Evaluation Procedure

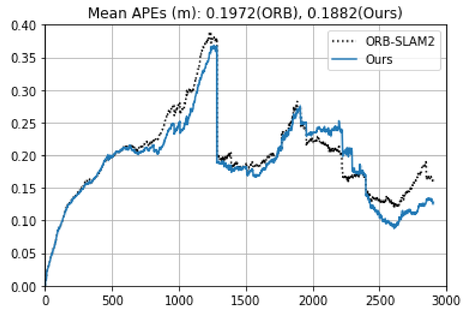
- Success rates of the relocalization task with its distance errors of the estimated positions are evaluated.
- RGBD image sequences and labels from ScanNet dataset (Dai et al. 2019) are used as input frames for VSLAM.

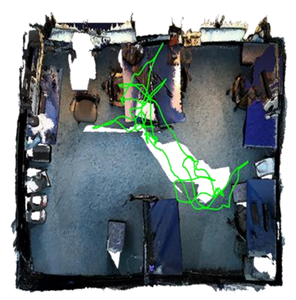
Relocalization Performance



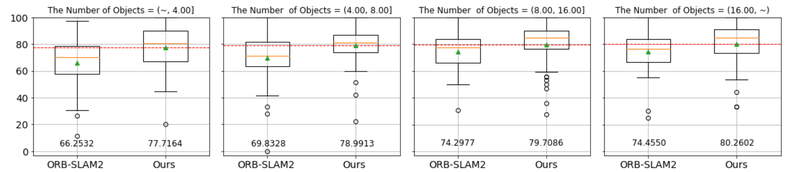
Environments	Num. of Scenes	Absolute Position Error (m)		Relative Position Error (m)		Relocalization Success Rate		Mean Distance (m)	
		Ours	ORB-SLAM2	Ours	ORB-SLAM2	Ours	ORB-SLAM2	Ours	ORB-SLAM2
Bathroom	27	0.0929	0.0898	0.0110	0.0117	**0.7534**	0.6918	0.0277	0.0166
Bedroom	25	0.1419	0.1455	0.0124	0.0129	**0.7985**	0.6607	0.0357	0.0283
Bookstore	10	0.2594	0.2437	0.0114	0.0114	**0.8381**	0.7276	0.0104	0.0095
Classroom	7	0.2077	0.2300	0.0134	0.0132	**0.8717**	0.8248	0.0133	0.0360
Conference Room	13	0.6619	0.6808	0.0139	0.0147	**0.7192**	0.6441	0.0388	0.0727
Copy/Mail Room	7	0.4954	0.4799	0.0164	0.0163	0.7028	**0.8159**	0.0161	0.0186
Hallway	8	0.1149	0.1238	0.0126	0.0129	**0.8417**	0.8076	0.0125	0.0150
Kitchen	16	0.2055	0.2112	0.0103	0.0108	**0.8251**	0.7461	0.0183	0.0159
Living room	34	0.1561	0.1479	0.0132	0.0130	**0.7580**	0.6418	0.0294	0.0257
Lobby	8	0.2228	0.2153	0.0143	0.0132	**0.8274**	0.5860	0.0162	0.0153
Office	22	0.1360	0.1457	0.0104	0.0105	**0.8486**	0.7778	0.0117	0.0253
Misc.	24	0.1510	0.1457	0.0112	0.0111	0.7766	**0.7791**	0.0126	0.0234
Total	201	0.1994	0.1997	0.0121	0.0123	**0.7895**	0.7117	0.0226	0.0252

Robustness over Trajectory Length





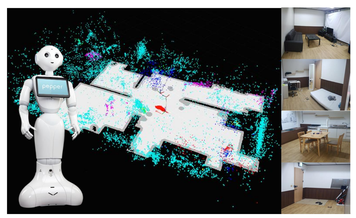
Effects related to the Number of Objects

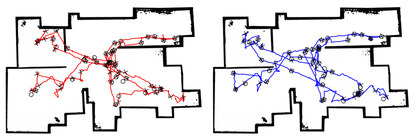


Efficiency of the Keyframe Candidate Selection

- Single keyframe (66.18%, average 1.89) remained after the whole selection process performed on initial keyframes.
- The number of keyframes decreased by 99.36%, 2.18%, and 0.93% from the total number of keyframes detected in each scene, throughout three steps of the algorithm.

Evaluation on Real Environment





	Reloc. success rate (%)	Average distance (m)
Ours	79.25	0.0512
ORB-SLAM2	71.43	0.0945

Conclusion

- Add-on feature augmentation module that fuses geometric representations with corresponding semantic representations
- Evaluated using a mobile robot platform in real world and a large-scale indoor 3D photo-realistic dataset

ICCV-2019 Workshop on Deep Learning for Visual SLAM | Seoul, Korea | November 2, 2019