2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

KR-Net: A Dependable Visual Kidnap Recovery Network for Indoor Spaces

Janghun Hyeon[1*], Dongwoo Kim[1*], Bumchul Jang[1,2*], Hyunga Choi[1*],
Dong Hoon Yi[3], Kyungho Yoo[3], Jeongae Choi[3], and Nakju Doh[4]

***Abstract*— In this paper, we propose a dependable visual kidnap recovery (KR) framework that pinpoints a unique pose in a given 3D map when a device is turned on. For this framework, we first develop indoor-GeM (i-GeM), which is an extension of GeM [1] but considerably more robust than other global descriptors [2]–[4], including GeM itself. Then, we propose a convolutional neural network (CNN)-based system called KR-Net, which is based on a coarse-to-fine paradigm as in [5] and [6]. To our knowledge, KR-Net is the first network that can pinpoint a wake-up pose with a confidence level near 100$\%$ within a 1.0$\,m$ translational error boundary. This dependable success rate is enabled not only by i-GeM, but also by a combinatorial pooling approach that uses multiple images around the wake-up spot, whereas previous implementations [5], [6] were constrained to a single image. Experiments were conducted in two challenging datasets: a large-scale (12,557$\,m^2$) area with frequent featureless or repetitive places and a place with significant view changes due to a one-year gap between prior modeling and query acquisition. Given 59 test query sets (eight images per pose), KR-Net successfully found all wake-up poses, with average and maximum errors of 0.246$\,m$ and 0.983$\,m$, respectively.**

I. INTRODUCTION

Recent advances in visual localization within a given 3D map have shown significant improvements with the application of deep-learning technology [5]–[10]. A pioneering work was conducted by Kendall *et al.* [7], [8], commonly known as PoseNet. However, according to [11], PoseNet shows lower performance in its accuracy and robustness, compared to current state-of-the arts approaches. Recent studies [12], [13] have developed systems that demonstrated enhanced performance, but the systems still have a weakness with regard to scalability.

Breakthroughs [5], [6] that simultaneously enhance accuracy, robustness, and scalability have demonstrated successful visual localization from a single image. These breakthroughs rely mainly on a coarse-to-fine paradigm that conducts global retrieval [1], [2] to obtain location hypotheses and local feature matching [9], [10] within those candidates. However, previous implementations of the coarse-to-fine paradigm are not suitable for our target task of kidnap recovery (KR). For the KR task, a critical requirement is unique pose pinpointing to determine, with high confidence, where a device (such as a robot or mobile phone) is being initiated. Considering our purpose of commercialization, the level of confidence should be near 100$\%$, which is far beyond the success rates of previous approaches such as the systems developed in [5] and [6].

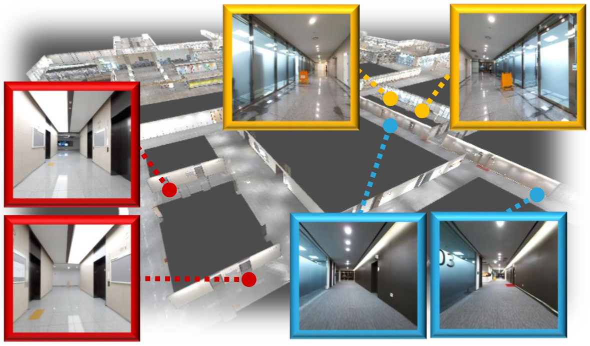


Fig. 1. One of our test sites, called M-site, where featureless and/or repetitive places are spread over the entire large map whose size is 12,557$\,m^2$. Here, images with the same color borders yield the same GeM scores, although their locations are quite different.

In this study, we propose a system called KR-Net, on the basis of the coarse-to-fine paradigm, which is sufficiently robust for commercialization. Here, the challenging problem is acquiring *abundant* as well as *reliable* features at the same time. A trade-off exists between the characteristics of *abundance* and *reliability*, mainly because of the attributes of objects. Naturally, there exist rich visual features in objects. However, those features could significantly impair localization when the objects' locations are not the same in the prior map and query images.

To overcome this problem, we exploit *structures*, which yield situation-invariant features in a way that fully utilizes their explicit visual features as well as implicit depth information as follows: First, regarding a prior 3D map, we adopt a structure-oriented map [14], [15] called TeeVR, which is a photo-realistic modeling of structures. Second, regarding a global feature for initial hypothesis retrieval, we suggest indoor-GeM (i-GeM), which is an extension of GeM [1], to make the best use of indoor structures in two ways: making them equivariant to feature location in order to distinguish similar images (as shown in Fig. 1) whose GeM scores are the same and embedding pixel-wise depth information.

To integrate these structure-oriented strategies into a monolithic network, we designed a convolutional neural

[1] Authors are with the School of Electrical Engineering, Korea University, Seoul, Republic of Korea.

[2] Author is with TeeLabs, Seoul, Republic of Korea.

[3] Authors are with LG Electronics, Seoul, Republic of Korea.

[4] N. Doh is a professor with the School of Electrical Engineering, Korea University, Seoul, Republic of Korea. He is also the CEO of TeeLabs. nakju@korea.ac.kr

*Authors contributed equally to this work.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

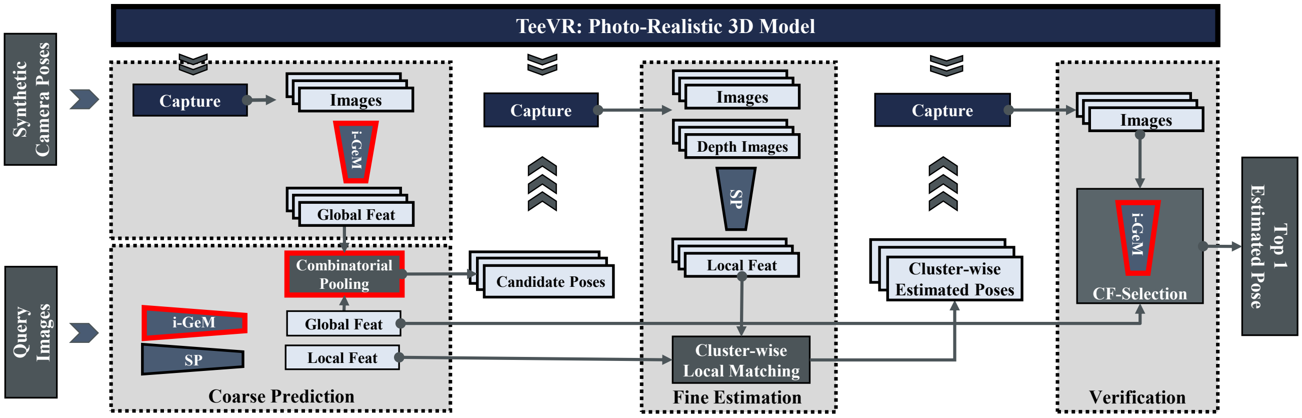


Fig. 2. KR-Net consists of a priori modeling (top) and on-line pose estimation (bottom). Pose estimation consists of three modules: coarse prediction (left), fine estimation (middle), and verification (right). This structure inherits the coarse-to-fine paradigm of HF-Net with additional verification from InLoc. Here, our key contributions are i-GeM and the combinational pooling as indicated in red.

network (CNN), called KR-Net, that successfully conducts kidnap recovery in challenging environments. As in previous implementations [5], [6], KR-Net is also based on the coarse-to-fine paradigm. However, it is designed to take multiple images around the initial pose, whereas previous implementations were constrained to use a single image that may contain insufficient information. To provide multiple image information for reliable pose decisions, we insert a combinatorial pooling between the coarse retrieval and the fine estimation, so that multiple independent retrievals are condensed into a single probability distribution over the entire map. Furthermore, KR-Net inherits the concept of view-verification and co-visibility from InLoc [5] and HF-Net [6], respectively, not only for increased accuracy but also for dependability at the cost of additional computation.

Experiments were conducted on two challenging datasets constructed at the KU-plaza and the M-site. Here, the KU-plaza dataset (1,930 m^2) contains dramatic changes of sign boards, object locations, and even structural changes due to large-scale remodeling. By contrast, there are few changes in the M-site. However, this large map (12,557 m^2) contains many featureless and/or repetitive regions over the entire area.

By these experiments, it was verified that the retrieval performance of i-GeM is superior to that of NetVLAD [2] by approximately 20% and 18% (within a 1.0 m recall threshold) in the KU-plaza and the M-Site, respectively. In addition, KR-Net successfully pinpointed its initial pose: The success rate was 100% (within the 1.0 m recall threshold) and the average accuracy was 0.246 m.

In summary, our contributions are as follows:

- Proposing a new global descriptor, i-GeM, which shows superior performance compared with other methods (such as NetVLAD and GeM) for structures in indoor spaces.
- Proposing a multiple image-based kidnap recovery network with two novelties: combinatorial pooling and integration of verification and co-visibility.
- Experimental validation of the robustness of KR-Net (success rate of 100%, even in featureless or repetitive places or under dramatic view changes), accuracy (average error of 0.246 m), and scalability (tested in areas up to 12,557 m^2).

This paper is organized as follows. In Section II, the proposed method is explained step-wise. Section III discusses the experimental validation results, and conclusions follow in Section IV.

II. Method

In this section, we propose a structure-oriented pipeline for the dependable kidnap recovery network called KR-Net, whose pipelines are shown in Fig. 2. This network consists of a priori modeling (top in Fig. 2) and on-line pose estimation (bottom). The pose estimation consists of three modules: coarse prediction (left), fine estimation (middle), and verification (right).

This structure inherits the coarse-to-fine paradigm of HF-Net, in which GeM and Superpoint (SP in Fig. 2) are used for the coarse prediction and fine estimation, respectively. KR-Net, however, substitutes GeM with i-GeM (Section II-A), which makes the best use of indoor structures. In addition, KR-Net includes combinatorial pooling (Section II-B) within the coarse prediction, which conveys multiple independent image data to a condensed single probability distribution. Given candidate poses from the coarse prediction, virtual images are captured from TeeVR, and the fine prediction (Section II-C) is initiated. Then, the verification (Section II-D), whose concept is adopted from InLoc for accuracy enhancement at the cost of additional computation, is conducted on all candidate poses from the coarse prediction and fine estimation.

A. *i-GeM*

In this section, we explain a new global descriptor, i-GeM, which makes the best use of explicit visual information as well as implicit depth information at indoor spaces. i-GeM is an extension of GeM in two directions, as shown in Fig. 3.

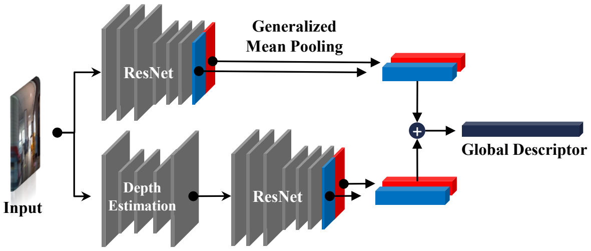
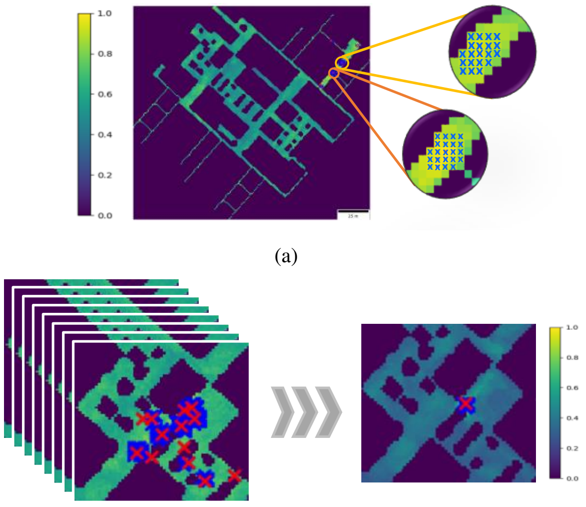


Fig. 3. Schematic diagram of i-GeM, which is an extension of the generalized mean pooling [1] in two directions. First, to discriminate different images with the same GeM scores, region-wise pooling is conducted for the left (red) and the right (blue) regions. Second, to embed depth information, a monocular depth prediction network [16] is exploited (bottom), in which depth is decoded by GeM. Finally, a total of four GeM features (two from visual and two from depth) constitutes i-GeM.



(a)

(b)

Fig. 4. Probability distributions over the entire map, from (a) a single image input and (b) multiple image inputs after a weighted-sum process.

First, to discriminate different images with the same GeM scores as shown in Fig. 1, region-wise pooling is conducted. In selecting the region, we focus on the fact that target images are horizontally flipped. Thus, we conduct GeM pooling from the left half and the right half twice so that the descriptor is equivariant to feature location. This scheme is easily implemented as the last layer of the backbone networks (VGG [17], ResNet [18]), directly corresponding to the input image in its location.

Second, to embed implicit depth information, a monocular depth prediction network [16] is exploited as in the bottom part of Fig. 3. Given the pixel-wise depth information, we conduct the same region-wise GeM pooling for two purposes: making the descriptors equivariant to feature location, and conserving features not from objects but from structures as GeM does.

Finally, the proposed i-GeM is constructed by combining four GeM features: two from the left and right images and the remaining two from the depth images.

Algorithm 1: Combinatorial Pooling

```
$P_{sim} \leftarrow f_q^T * f_{db},$
$S_{map} \leftarrow map(P_{sim}, {}^{map}I_{db})$
$S_{xy} \leftarrow max(S_{map}, axis = 0)$
$Arr_{ang} \leftarrow argmax(S_{map}, axis = 0)$
$\hat{S}_{xy} \leftarrow maxpool(S_{xy})$
$T_{peak} \leftarrow (S_{xy} = \hat{S}_{xy})$ and $(S_{xy} > \lambda \times max(S_{xy}))$
$T_{cluster}$=[ ]
for $idx$ in $T_{peak}$ do
    $C \leftarrow (S_{xy}[idx] = \hat{S}_{xy})$ and $(\hat{S}_{xy} > \lambda \times max(S_{xy}))$
    $T_{cluster}.append(C)$
end
return $T_{cluster}$, $Arr_{ang}$
```

B. Combinatorial Pooling

In this subsection, we explain the combinatorial pooling, which conveys multiple independent visual data to a condensed probability distribution over the entire map.

First, given two global descriptors from a query (f_q) and a database image (f_{db}), a similarity value (P_{sim}) is calculated as the inner product ($P_{sim} = f_q^T * f_{db}$). Then, we project these similarities into the grid-map. Considering orientation, we conserve best matched orientation information by selecting the largest P_{sim} among the similarities from the each database position.

Then, we construct clusters whose values are higher than a certain threshold ($\lambda = 0.85$). Candidate poses within each cluster are indicated by blue crosses in Fig. 4. As shown in Fig. 4(a), the probability map contains many clusters because of insufficient information considering the large size of the map. However, if we use multiple independent images and conduct a weighted-sum process, only a few candidate groups appear, from which we can pinpoint an initial pose as shown in Fig. 4(b).

C. Fine Prediction

Given clusters of pose candidates, we conduct fine prediction for two purposes: unifying local features within a cluster into a single pose, and updating the single pose for higher precision.

For this stage, it should be noted that our structure-based approach has both advantages and disadvantages. The advantage is that only situation-invariant features from structures are used; the disadvantage is that the number of features is significantly reduced as shown in Fig. 5. However, because we use multiple images, we can obtain sufficient features to accurately estimate the pose.

The fine prediction consists of two steps. First, all local features, Superpoint [10] in our approach, of multiple poses in each candidate cluster are aggregated into a set of features. This can be easily conducted in TeeVR, because this map provides any-view corresponding images as well as pixel-wise depth. Second, given a set of features, we apply a 2D–3D matching using the PnP method [19] to accurately refine the pose.

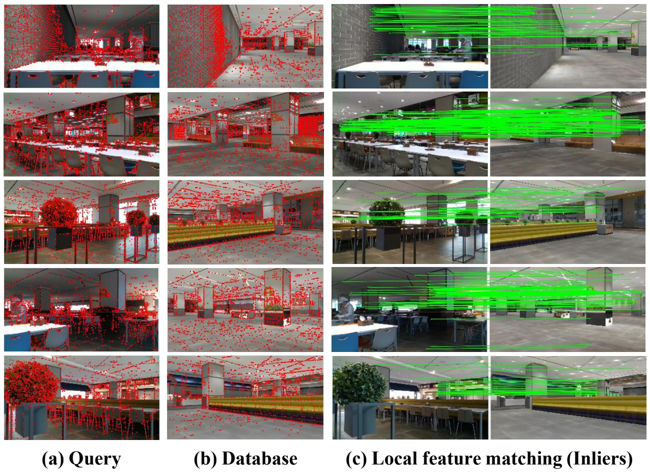


Fig. 5. Characteristics of our structure-oriented approach, where given many local features (Superpoint) of query images, only reliable features that overlap those of database images as in (b) are selected to be inliers as in (c).

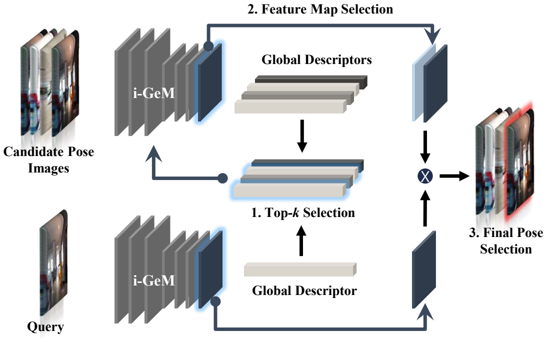


Fig. 6. Intensive verification for whole candidates from the query and all candidate poses.

Note that the overall scheme described in this subsection is basically the same as that of HF-Net, the so-called co-visibility approach.

D. Verification

For high dependability, we further adopt the verification concept from InLoc [5] at the cost of additional computation. Its pipeline, as shown in Fig. 6, works as follows.

- 1. Given images from estimated poses, compute similarities using i-GeM from a query and the estimated pose images.
- 2. Select the top K images. Here, K varies with respect to the relative threshold.
- 3. Conduct a rigorous verification of the K candidates that compares the similarity in a pixel-by-pixel manner. Here, we also compute error values utilizing the equivariant discrimination as discussed in Section II-A but in a more rigorous pixel-by-pixel manner.
- 4. Select the final pose whose image has the minimum error value from step 3.

III. Experimental Evaluation

A. Dataset

Because the open datasets [20]–[22] cannot be converted to TeeVR format, we built the two datasets of KU-plaza and M-site. First, the KU-plaza, of size 1,930 m^2, is a representative place of significant view changes, because there was a one-year gap between the 3D modeling and query acquisition. Some examples of view changes are board-sign changes (Fig. 7 (a), blue box), illumination changes (Fig. 7 (a), green box), and even structural changes (Fig. 7 (a), red box).

Second, the M-Site, of size 12,557 m^2, is a representative large-scale place (Fig. 7 (b)) with many featureless (Fig. 7 (b), yellow box), and/or repetitive places (Fig. 7 (b), red box and blue box), as well as object-dominant places (Fig. 7 (b), green box). This site is suitable for testing the following characteristics of visual localization:

- *Reliability* in many featureless spots.
- *Ambiguity discrimination* among many repetitive spots.
- *Scalability* in large-scale environments, which also makes the aforementioned reliability and ambiguity discrimination far more difficult.
- *Individual contributions of objects and structures* in visual localization.

B. Implementation Details

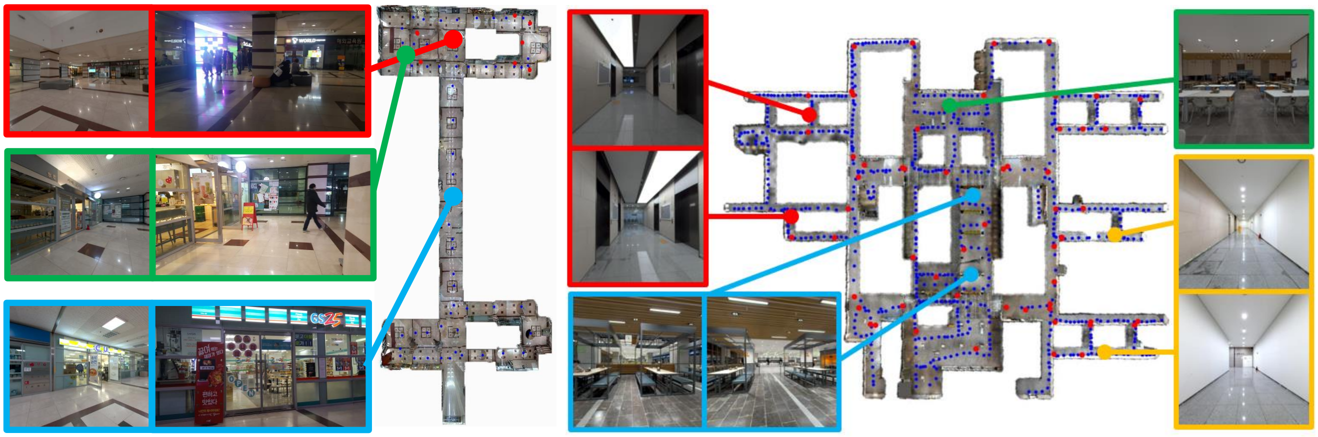
3D Map Modeling

For a prior 3D map generation, a structure-oriented photo-realistic modeling, namely TeeVR [14], is adopted. It consists of two modules: data acquisition and modeling.

First, a scanning robot (as shown in Fig. 8) [23] equipped with two LiDARs (Velodyne VLP-16) and a 360° camera (Ladybug5 Plus) roams around the space at 2 km/h recording images at 10 *fps*. At this speed, it takes 20 minutes and 6 hours to cover the full area of the KU-plaza and the M-site, respectively. Second, accurate pose estimation [24] is conducted using pointcloud data, which is followed by automatic structure generation [25] and structure-oriented image inpainting [15].

Here, this modeling strategy is not only suitable in our structure-oriented approach but also useful for other visual localization schemes. For example, the dense colored pointcloud data and key frame images of InLoc [5] is easily selected after the pose estimation step of TeeVR's pipeline, whereas those data were laboriously acquired in [5]. In addition, because the scanning robot acquires images at 10 *fps*, a sufficient number of images for SfM in [6] are easily provided.

Furthermore, TeeVR can provide pixel-wise depth information that can be used for i-GeM's depth training, and its compact storage (146.0 MB and 431.3 MB for the KU-plaza and the M-site, respectively) increases the feasibility of on-device implementation.



(a)

(b)

Fig. 7. Datasets were generated in two target sites. (a) The first site is KU-plaza (1,930 m^2) where the prior map (left square) is significantly different from the query (right rectangle): A new corridor is found in the query (red), light conditions are different (green), and a few board signs were changed (blue). (b) The second site is the M-site (12,557 m^2), which includes many featureless (yellow) and/or repetitive places (red and blue) as well as object-dominant places (green) spread over the entire space.

TABLE I
IMAGE RETRIEVAL FEATURE EVALUATION

	KU-Plaza									M-Site								
	Top 1			Top 5			Top 10			Top 1			Top 5			Top 10		
	1.0 m	3.0 m	5.0 m	1.0 m	3.0 m	5.0 m	1.0 m	3.0 m	5.0 m	1.0 m	3.0 m	5.0 m	1.0 m	3.0 m	5.0 m	1.0 m	3.0 m	5.0 m
NetVlad [2]	0.000	0.051	0.103	0.026	0.077	0.205	0.026	0.128	0.333	0.180	0.341	0.388	0.350	0.519	0.566	0.430	0.610	0.659
GeM [1]	0.051	0.231	0.487	0.179	0.410	0.615	0.282	0.513	0.667	0.197	0.381	0.407	0.386	0.508	0.525	0.441	0.561	0.585
GeM+Eq	0.128	0.359	0.538	0.282	0.487	0.615	0.333	0.564	0.667	0.309	0.483	0.513	0.487	0.585	0.606	0.576	0.648	0.669
i-GeM	**0.205**	**0.538**	**0.590**	**0.333**	**0.564**	**0.641**	**0.385**	**0.615**	**0.769**	**0.364**	**0.532**	**0.547**	**0.572**	**0.665**	**0.680**	**0.625**	**0.718**	**0.746**

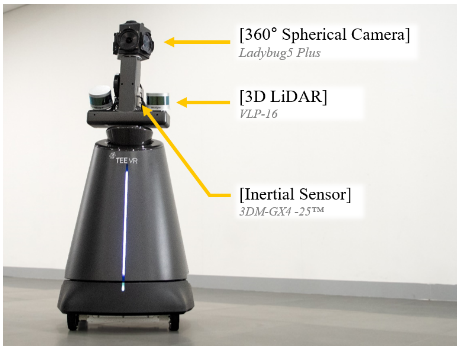


Fig. 8. Sensor system for data acquisition. Sensor measurements, acquired by a spherical camera, two 3D LiDARs, and an inertial sensor are employed to generate realistic 3D modeling [14].

Database Generation

From TeeVR itself or its raw data after pose estimation, we constructed three sets of databases for InLoc [5], HF-Net [6], and KR-Net.

First, we generated an InLoc dataset of 36 perspective RGB-D images for each 360° image. In this generation, sampling strides for yaw and pitch angles were set to 30° and ± 30°, respectively. As a result, we built a total of 2,100 and 27,000 RGB-D image sets for the KU-plaza and the M-site, respectively.

Second, we built an SfM model of the KU-plaza with COLMAP [26], [27] using 1,300 images with known camera poses, and followed a model construction process as HF-Net instructed for a test of HF-Net. In the case of the M-site, it is difficult to build an SfM model of the place because it is a large-scale indoor space with many self-similar features (such as repetitive patterns), although we used known camera poses. Thus, we tested only in the KU-plaza for the comparison of HF-Net.

Finally, for the validation of KR-Net, we captured virtual images from the TeeVR map. Because the map provides an image from any point of view given a pre-specified field of view (FoV), we extracted images for every 50 cm in 36 yaw directions. Here, the FoV was set to be the same as that of the camera (Realsense) used in query image generation. As a result, a total of 130,000 and 240,000 image databases were constructed for the KU-plaza and the M-site, respectively.

TABLE II

POSE ESTIMATION RESULTS USING SINGLE IMAGE

		Top 1				
		$0.3\,m$	$0.5\,m$	$1.0\,m$	$3.0\,m$	$5.0\,m$
KU-Plaza	InLoc [5]	0.0769	0.2821	0.4103	0.5385	0.5897
	HF-Net [6]	0.1026	0.1538	0.3333	**0.8462**	**0.9231**
	KR-Net	**0.2051**	**0.5128**	**0.6923**	0.7692	0.7692
M-Site	InLoc [5]	0.4068	0.5678	**0.6864**	**0.7564**	**0.7606**
	KR-Net	**0.4703**	**0.5890**	0.6610	0.7225	0.7309

TABLE III

POSE ESTIMATION RESULTS USING MULTIPLE IMAGES

	Top 1				avg. error (m)
	$0.1\,m$	$0.3\,m$	$0.5\,m$	$1.0\,m$	
success rate	0.288	0.763	0.881	1.000	0.246

Query Sets Acquisition

Query image sets were acquired so that multiple images around a spot were obtained. In the KU-plaza, we took four images using a smartphone (Samsung Galaxy Note 5) for every 90° at 10 places. In the M-site, however, we acquired eight images using a Realsense at a resolution of 45° because the camera's FoV (Realsense) was narrower than that of the smartphone. In 59 arbitrarily selected places (indicated by the red spot in Fig. 7), a total of 472 images were acquired, of which the ground truth was identified by a manual one-to-one matching in a way that fully overlaps two images: TeeVR and the query.

C. Feature Evaluation

For the evaluation of i-GeM, evaluations including an ablation study were conducted for two datasets in a way that identifies the accuracy of retrieval given a single image as the recall threshold changes.

Table I shows the results of NetVLAD [2], GeM [1], GeM with equivariant-descriptor (GeM+Eq), and i-GeM (GeM+Eq+Depth). Here, Top 1, Top 5, and Top 10 indicate how many top hypotheses were included in evaluating the success rate (%) as recall distance changes (*1.0 m*, *3.0 m*, and *5.0 m*).

For the KU-plaza dataset, it was shown that GeM (the second row) yields superior performance compared with NetVLAD (the first row) against significant view changes. In addition, it was shown that only depth information in i-GeM (the fourth row) significantly improves retrieval performance, whereas the equivariant discrimination contributes little. By contrast, both the equivariant discrimination and the depth information similarly contribute in the M-site, because there are many repetitive places over the entire area. Overall, it was shown that i-GeM is superior in describing explicit or implicit features of structures compared with NetVLAD, and that i-GeM is a suitable choice in the sense of robustness.

D. Comparison

Because of the difference in the number of image inputs between the current state-of-the art techniques (InLoC [5] and HF-Net [6]) and KR-Net, a fair comparison cannot be conducted. As an alternative, comparisons were conducted using a single image for two datasets as shown in Table II.

In the results for KU-plaza, KR-Net shows superior performance for recall thresholds of $0.3\,m$, $0.5\,m$, and $1.0\,m$. In contrast, HF-Net shows the best results at $3.0\,m$ and $5.0\,m$. However, the results of HF-Net cannot be trusted because the SfM map for HF-Net was successfully generated only for a relatively small indoor place ($965\,m^2$), which, in return, provides favorable situation for the $3.0\,m$ and $5.0\,m$ criteria. Thus, overall, it can be said that KR-Net shows superior results under significant view changes even when using a single image.

In the results for the M-site, we were unable to generate the SfM model for HF-Net for the reason mentioned in Section III-B (Database Generation). Thus, comparisons were conducted only for InLoc and KR-Net; it can be shown that KR-Net exhibits better precision within a tight threshold ($\sim 0.5\,m$), whereas InLoc exhibits better recall within a wide threshold ($\sim 5.0\,m$). Considering that InLoc utilizes features both from objects and structures (whereas KR-Net uses only those from structures) and those from objects whose locations were not changed, KR-Net's baseline performance with a single image can be said to be competitive to that of InLoc.

E. Kidnap Recovery Evaluation

The objective of KR-Net, a dependable KR with multiple images, was tested for 59 sample sets in the M-site. Table III shows recall rates at different distances *w.r.t.* the ground truth for all the query sets. Because all the wake-up poses are successfully found within the $1.0\,m$ bound, and the average error is $0.246\,m$, we conclude that our structure-based approach with multiple images yields highly reliable results and fits well into our commercialization plan.

IV. CONCLUSIONS

In this paper, we propose a highly dependable KR-Net that successfully works in two challenging datasets: one (KU-plaza) with significant view changes, and the other (M-site) with many featureless and/or repetitive places over the entire large map. To our knowledge, this is the first system that pinpoints an initial pose with a success rate near 100%, which is far beyond that of the current state-of-the-art techniques.

These significant improvements were enabled by two main contributions. One is a structure-oriented global descriptor (i-GeM) that provides reliable information, and the other is a CNN network (KR-Net) that can make the best use of abundant information from multiple images.

In the future, we plan to conduct more rigorous testing in various places with a goal of commercialization. Meanwhile, we will add information regarding some objects that may not be moved, such as furniture or heavy statues. In addition, we will optimize the overall algorithm so that it can run online in a device, not in a cloud as is done now.

ACKNOWLEDGMENT

This research was supported by the Brain Korea 21 Plus project in 2020.

REFERENCES

[1] F. Radenović, G. Tolias, and O. Chum, “Fine-tuning cnn image retrieval with no human annotation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[3] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, “Deep image retrieval: Learning global representations for image search,” in *European Conference on Computer Vision*, 2016.

[4] G. Tolias, R. Sicre, and H. Jégou, “Particular object retrieval with integral max-pooling of cnn activations,” *arXiv preprint arXiv:1511.05879*, 2015.

[5] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, “Inloc: Indoor visual localization with dense matching and view synthesis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[6] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, “From coarse to fine: Robust hierarchical localization at large scale,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[8] A. Kendall and R. Cipolla, “Geometric loss functions for camera pose regression with deep learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, “D2-net: A trainable cnn for joint detection and description of local features,” *arXiv preprint arXiv:1905.03561*, 2019.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[11] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, “Understanding the limitations of cnn-based absolute camera pose regression,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] A. Valada, N. Radwan, and W. Burgard, “Deep auxiliary learning for visual localization and odometry,” in *International Conference on Robotics and Automation(ICRA)*, 2018.

[13] N. Radwan, A. Valada, and W. Burgard, “Vlocnet++: Deep multitask learning for semantic visual localization and odometry,” *IEEE Robotics and Automation Letters*, 2018.

[14] N. Doh, H. Choi, B. Jang, S. Ahn, H. Jung, and S. Lee, “Teevr: spatial template-based acquisition, modeling, and rendering of large-scale indoor spaces,” in *ACM SIGGRAPH 2019 Emerging Technologies*, 2019.

[15] J. Hyeon, H. Choi, J. Kim, B. Jang, J. Kang, and N. Doh, “Automatic spatial template generation for realistic 3d modeling of large-scale indoor spaces,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[16] W. Yin, Y. Liu, C. Shen, and Y. Yan, “Enforcing geometric constraints of virtual normal for depth prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] L. Kneip, D. Scaramuzza, and R. Siegwart, “A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[20] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The oxford robotcar dataset,” *The International Journal of Robotics Research*, 2017.

[21] A. Bansal, H. Badino, and D. Huber, “Understanding how camera configuration and environmental conditions affect appearance-based localization,” in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 2014.

[22] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, “Image retrieval for image-based localization revisited.” in *BMVC*, 2012.

[23] J. Kang and N. L. Doh, “Automatic targetless camera–lidar calibration by aligning edge with gaussian mixture model,” *Journal of Field Robotics*, 2020.

[24] K. Lee, S.-H. Ryu, S. Yeon, H. Cho, C. Jun, J. Kang, H. Choi, J. Hyeon, I. Baek, W. Jung, *et al.*, “Accurate continuous sweeping framework in indoor spaces with backpack sensor system for applications to 3-D mapping,” *IEEE Robotics and Automation Letters*, 2016.

[25] G. Lim, Y. Oh, D. Kim, C. Jun, J. Kang, and N. Doh, “Modeling of Architectural Components for Large-scale Indoor Spaces from Point Cloud Measurements,” *IEEE Robotics and Automation Letters*, 2020.

[26] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[27] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, “Pixelwise view selection for unstructured multi-view stereo,” in *European Conference on Computer Vision*, 2016.