2020
ACM multimedia Seattle

Zero-Shot Multi-View Indoor Localization via Graph Location Networks

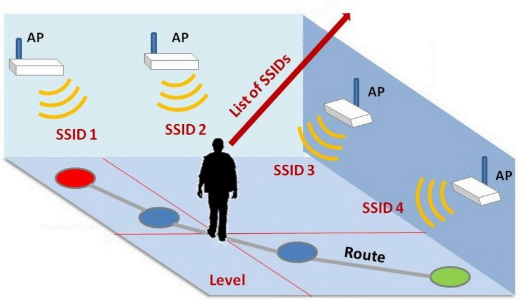
Meng-Jiun Chiou[1], Zhenguang Liu[2*], Yifang Yin[1], An-An Liu[3], Roger Zimmermann[1]

[1]National University of Singapore [2]Zhejiang Gongshang University [3]Tianjin University

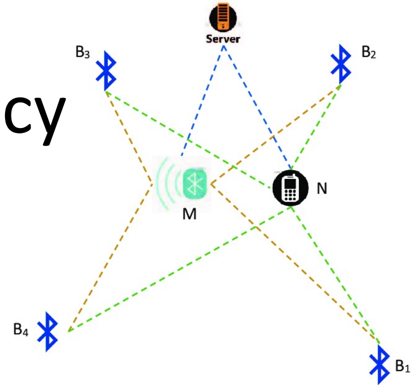
*Corresponding author

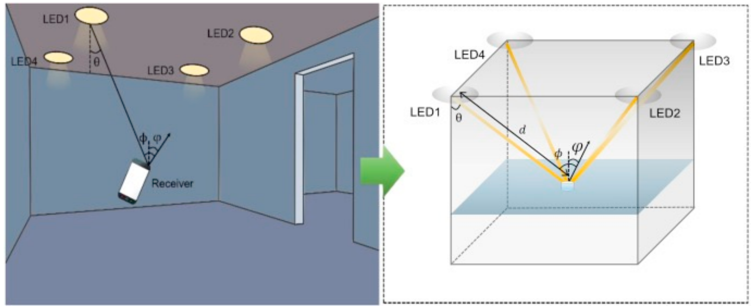
Background & Motivation

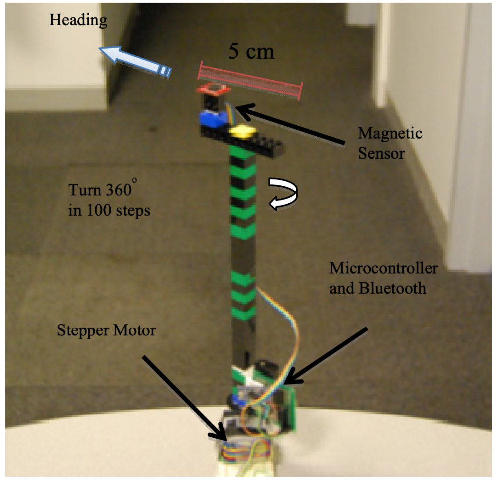
Localization



- Objective: Accurate positioning of user(s)
- Outdoor: Global Positioning System (GPS)[1]: 3-5 meters accuracy
- Indoor:
  - Signal based: WiFi[2], Bluetooth[3], magnetism[4], optics[5], etc.
    - Require additional infrastructures (WiFi access points/transmitters/receivers)
  - Image based (this work)
    - Require only minimal setup







[1] https://www.gizmochina.com/2017/09/26/broadcoms-highly-accurate-gps-chip-arriving-smartphones-next-year/

[2] https://eloquentarduino.github.io/2019/12/wifi-indoor-positioning-on-Arduino/

[3] Li et al. Indoor Positioning Algorithm Based on the Improved RSSI Distance Model. Sensors. 18. 2820

[4] Chung, Jaewoo, et al. "Indoor location sensing using geo-magnetism." Proceedings of the 9th international conference on Mobile systems, applications, and services. 2011.

[5] Chen, Hao, et al. "Indoor high precision three-dimensional positioning system based on visible light communication using modified genetic algorithm." Optics Communications 413 (2018)

Motivation I: Feature Propagation between Views

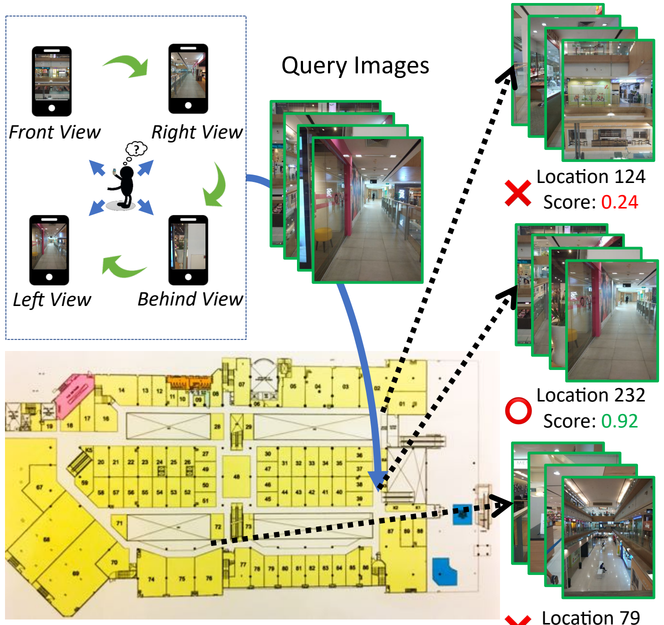
- Multi-view: we assume there are four views at each location. Namely, *front*, *right*, *back* and *left*.

*Images are from WCP dataset

Motivation I: Feature Propagation between Views

- Multi-view: we assume there are four views at each location. Namely, *front*, *right*, *back* and *left*.

*Images are from WCP dataset

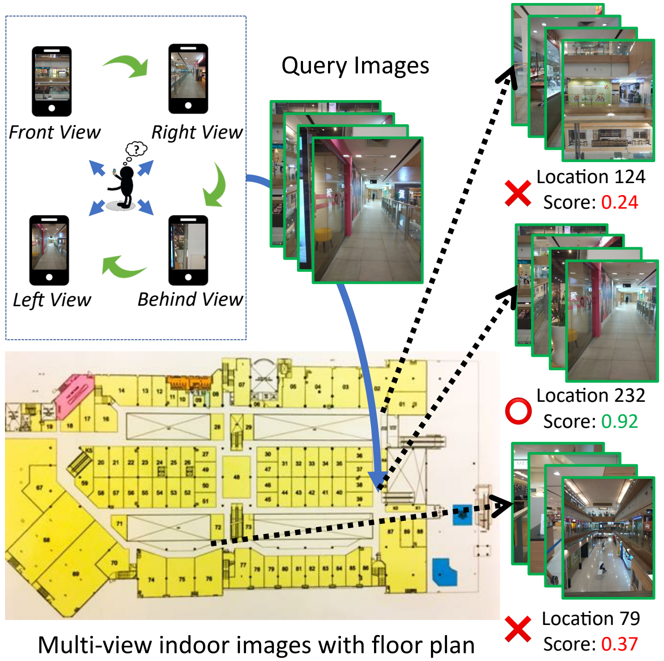


Multi-view indoor images with floor plan

Motivation I: Feature Propagation between Views

- Multi-view: we assume there are four views at each location. Namely, *front*, *right*, *back* and *left*.

*Images are from WCP dataset



Multi-view indoor images with floor plan

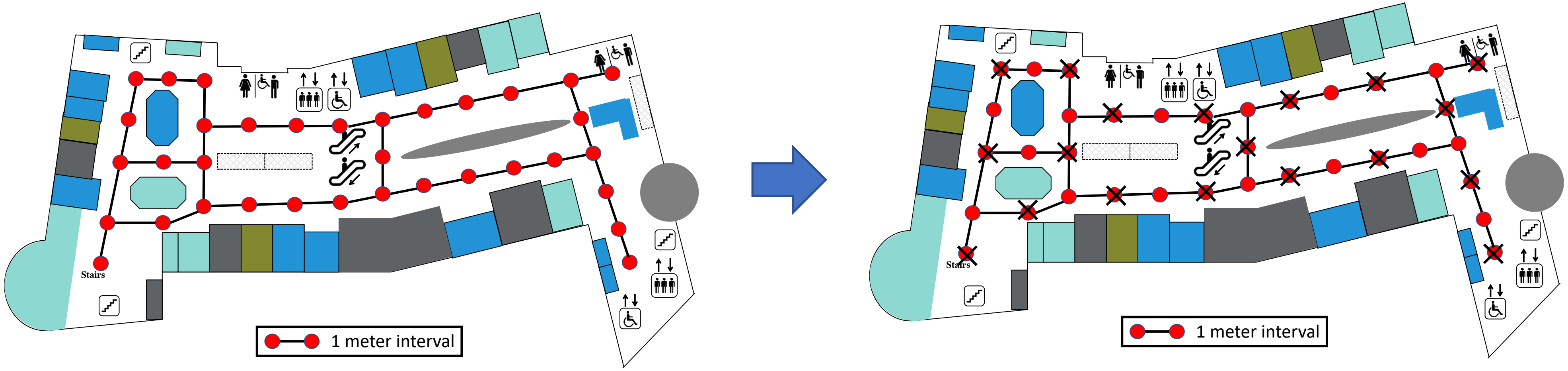
- Shouldn’t we treat these views **differently**?
  - Each view features more/less information needed for location retrieval. We should treat them differently to get a holistic representation of a location.

Motivation II: Reduce Costs via Zero-shot Learning

- Is it possible to reduce labelling costs (training data collection), by not collecting some of the locations?

Motivation II: Reduce Costs via Zero-shot Learning

- Is it possible to reduce labelling costs (training data collection), by not collecting some of the locations?

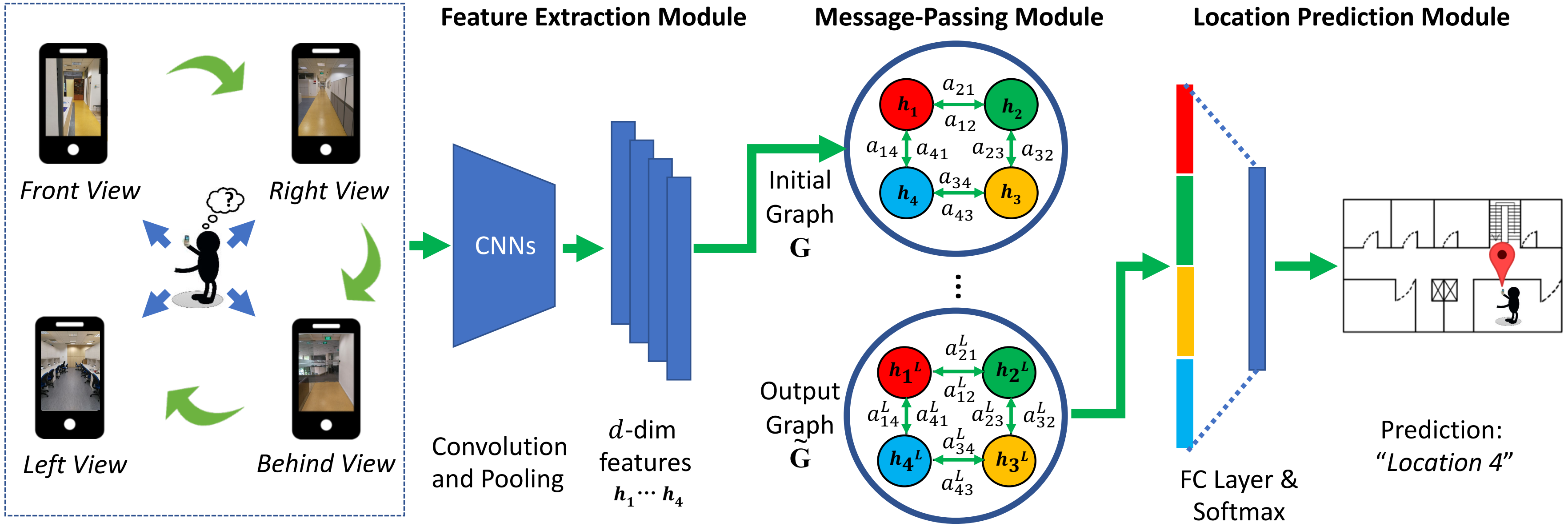


- We assume the unseen locations are in-between seen locations. Thus, the training data is reduced by ~50%

Methodology

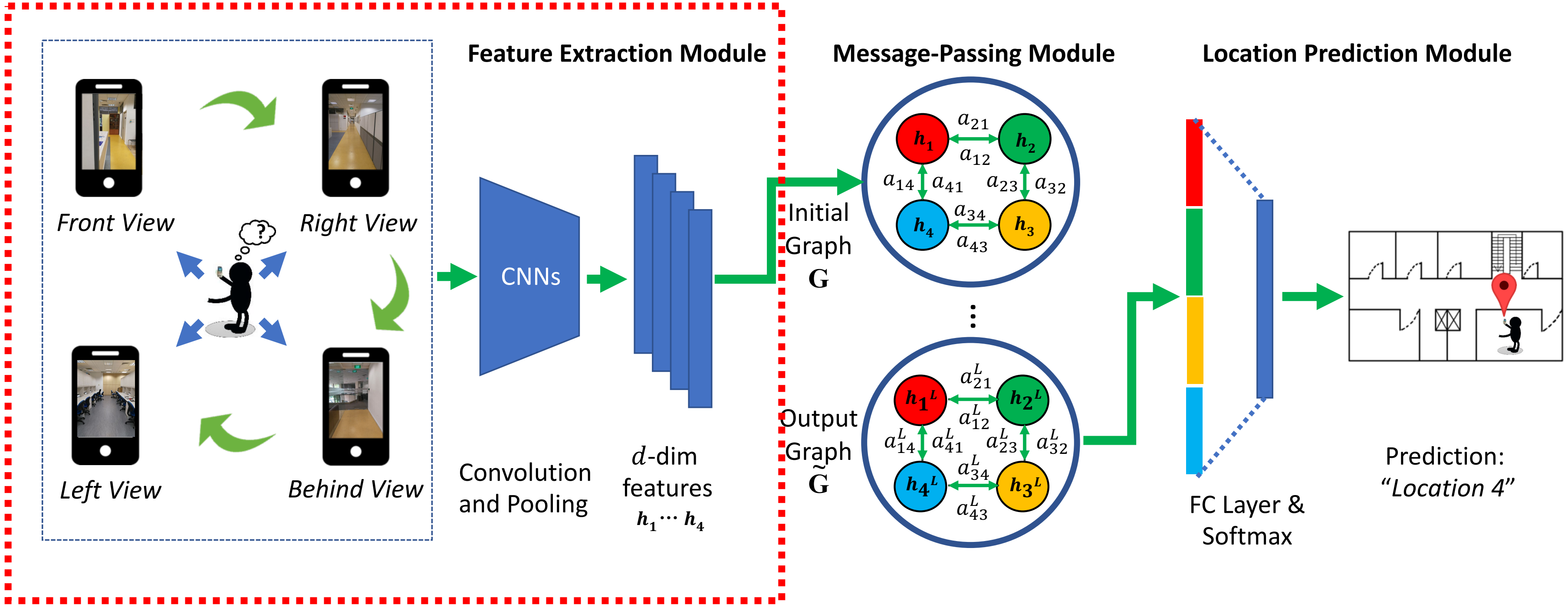
Graph Location Networks (GLN)

- We propose a multi-view image based localization method that utilizes Graph Neural Networks (GNNs) to propagate distinct features of different views.



Graph Location Networks (GLN)

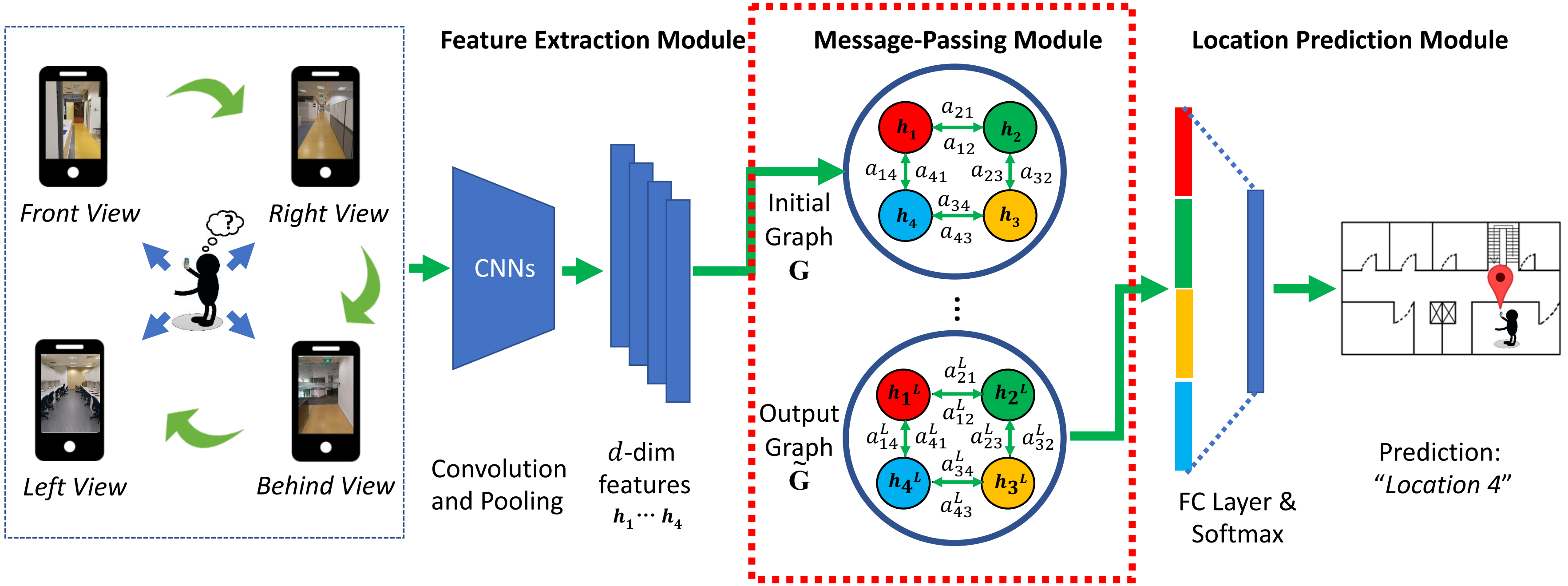
- We propose a multi-view image based localization method that utilizes Graph Neural Networks (GNNs) to propagate distinct features of different views.



Extracting image features with ResNet-152

Graph Location Networks (GLN)

- We propose a multi-view image based localization method that utilizes Graph Neural Networks (GNNs) to propagate distinct features of different views.



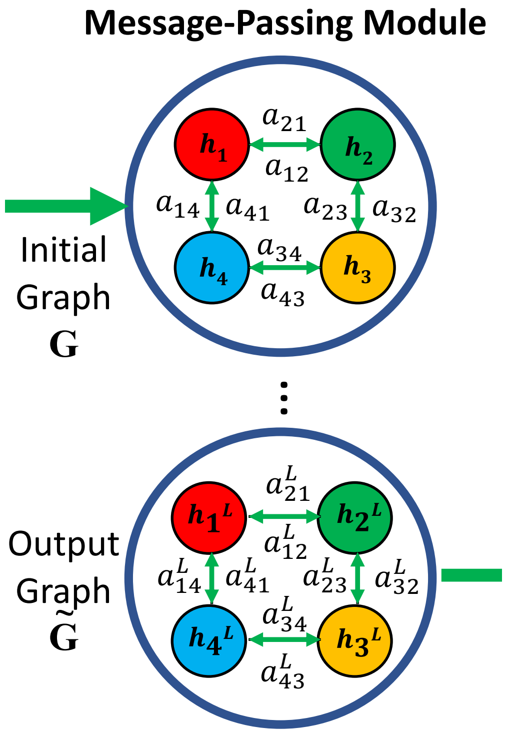
Message-passing with GNNs
(w/ or w/o attention mechanism)

GNNs & Attention Mechanism

- Graph Conv Nets: $h_i^l = \begin{cases} r_i, & \text{if } l = 1 \\ \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{\alpha_{ij}} W^{l-1} h_j^{l-1}\right), & \text{otherwise} \end{cases}$

- Where normalization constant

$$\alpha_{ij} = \sqrt{|\mathcal{N}(i)\mathcal{N}(j)|}$$



Message-Passing Module

GNNs & Attention Mechanism

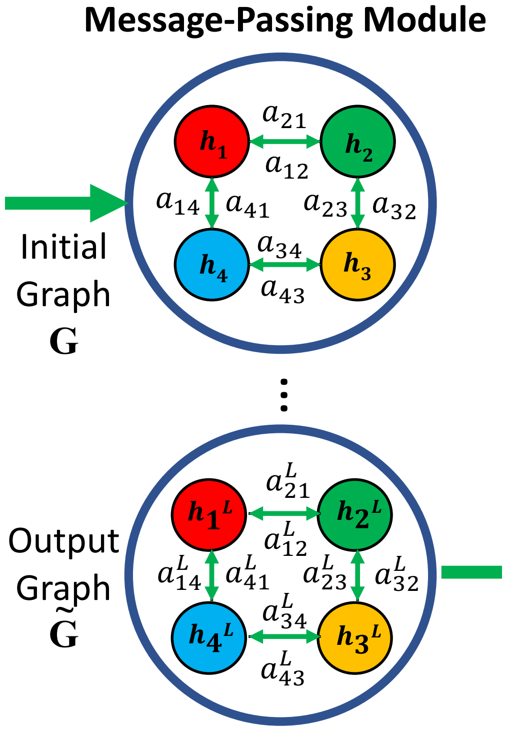
- Graph Conv Nets: $h_i^l = \begin{cases} r_i, & \text{if } l = 1 \\ \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{\alpha_{ij}} W^{l-1} h_j^{l-1}\right), & \text{otherwise} \end{cases}$

- Where normalization constant

$$\alpha_{ij} = \sqrt{|\mathcal{N}(i)\mathcal{N}(j)|}$$

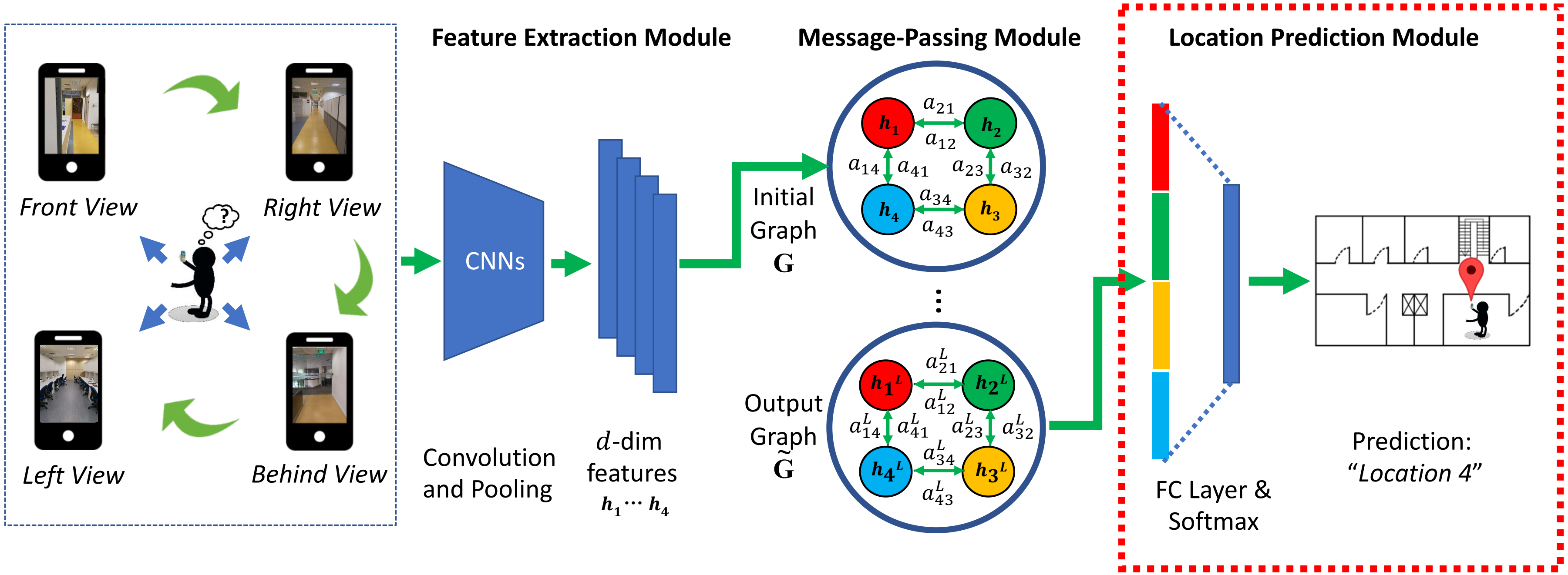
- We try to employ attention mechanism to actively assign different weights for views

$$\alpha_{ij}^l = \frac{exp(\sigma(a[W^l h_i^l, W^l h_j^l]))}{\sum_{k \in \mathcal{N}(i)} exp(\sigma(a[W^l h_i^l, W^l h_k^l]))},$$



Graph Location Networks (GLN)

- We propose a multi-view image based localization method that utilizes Graph Neural Networks (GNNs) to propagate distinct features of different views.


Front View
Right View
Left View
Behind View
Feature Extraction Module
CNNs
Convolution and Pooling
d-dim features
$h_1 \cdots h_4$
Message-Passing Module
Initial Graph
$\mathbf{G}$
h_1 h_2 h_3 h_4
a_{21} a_{12} a_{14} a_{41} a_{23} a_{32} a_{34} a_{43}
Output Graph
$\tilde{\mathbf{G}}$
$h_1{}^L$ $h_2{}^L$ $h_3{}^L$ $h_4{}^L$
a^L_{21} a^L_{12} a^L_{14} a^L_{41} a^L_{23} a^L_{32} a^L_{34} a^L_{43}
Location Prediction Module
FC Layer & Softmax
Prediction:
"Location 4"

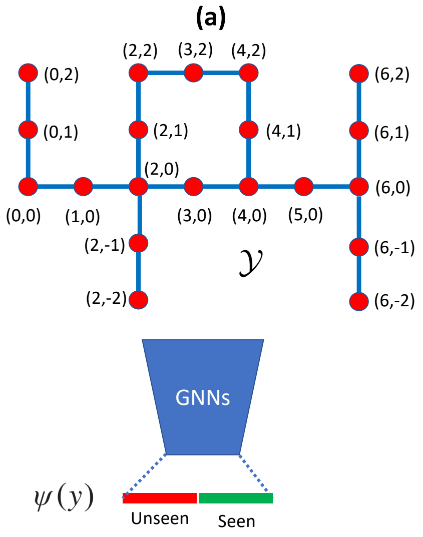
Training against cross-entropy loss
Inference by picking the most confident location

Enabling Zero-Shot Indoor Localization

- A Three-step framework:

Enabling Zero-Shot Indoor Localization

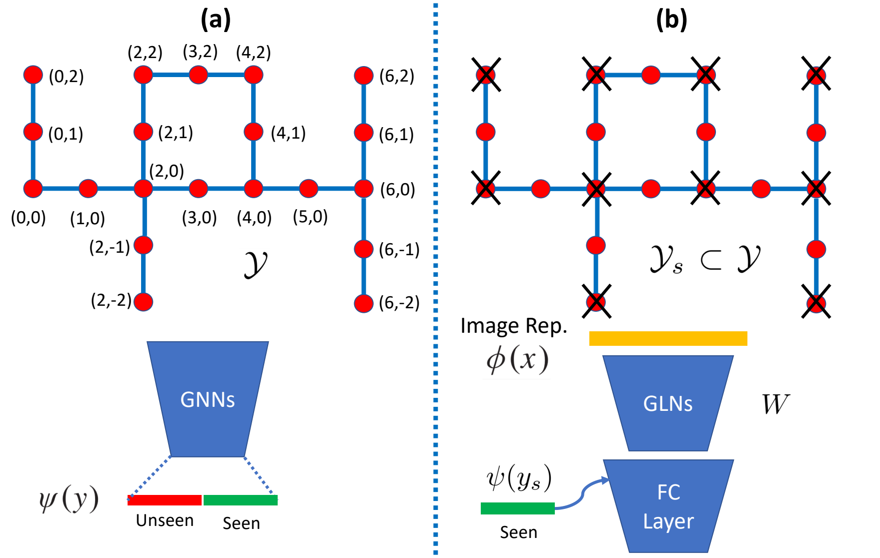
- A Three-step framework:
  1) Train *Map2Vec* location embeddings for both *seen* & *unseen* locations



Output: *Map2Vec* Location Embedding

Enabling Zero-Shot Indoor Localization

- A Three-step framework:
  1) Train *Map2Vec* location embeddings for both *seen* & *unseen* locations
  2) Train an indoor localization architecture (e.g. GLN), where an additional layer (compatibility function) was additionally added, with only *seen* data



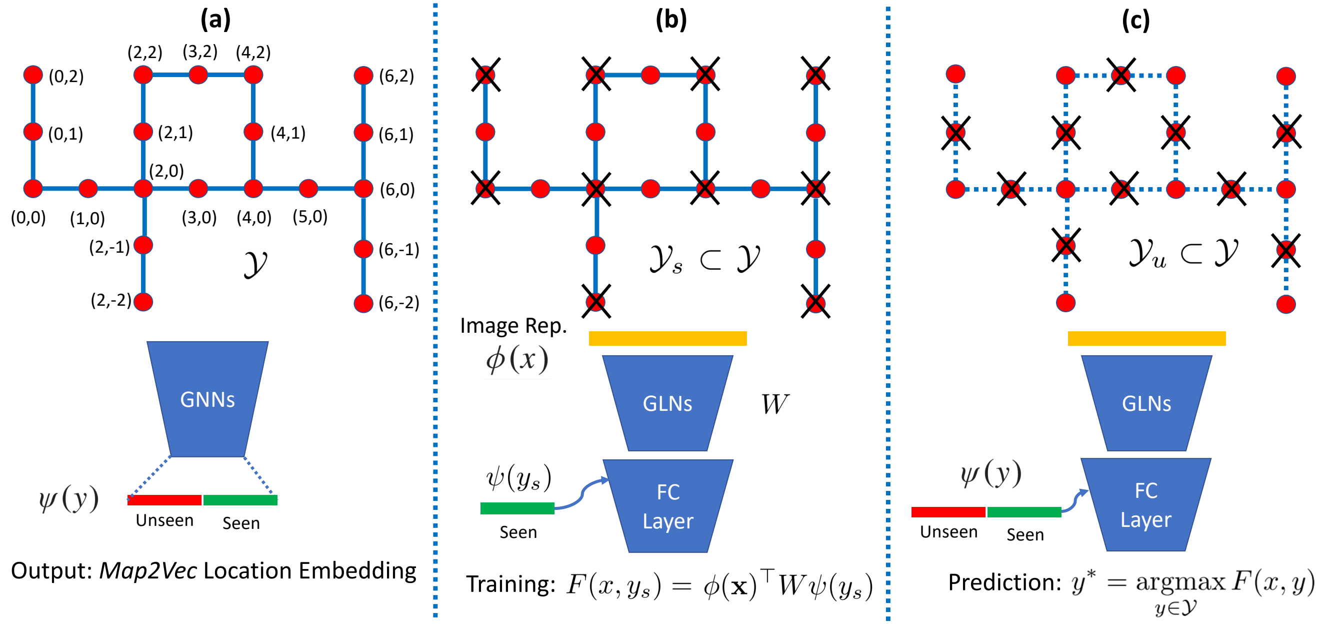
Output: *Map2Vec* Location Embedding

Training: $F(x, y_s) = \phi(\mathbf{x})^\top W \psi(y_s)$

$$F(x, y_s) = \phi(\mathbf{x})^\top W \psi(y_s),$$

Enabling Zero-Shot Indoor Localization

- A Three-step framework:
  1) Train *Map2Vec* location embeddings for both *seen* & *unseen* locations
  2) Train an indoor localization architecture (e.g. GLN), where an additional layer (compatibility function) was additionally added, with only *seen* data
  3) Perform Inference by picking the most probable location



Output: *Map2Vec* Location Embedding

Training: $F(x, y_s) = \phi(\mathbf{x})^\top W \psi(y_s)$

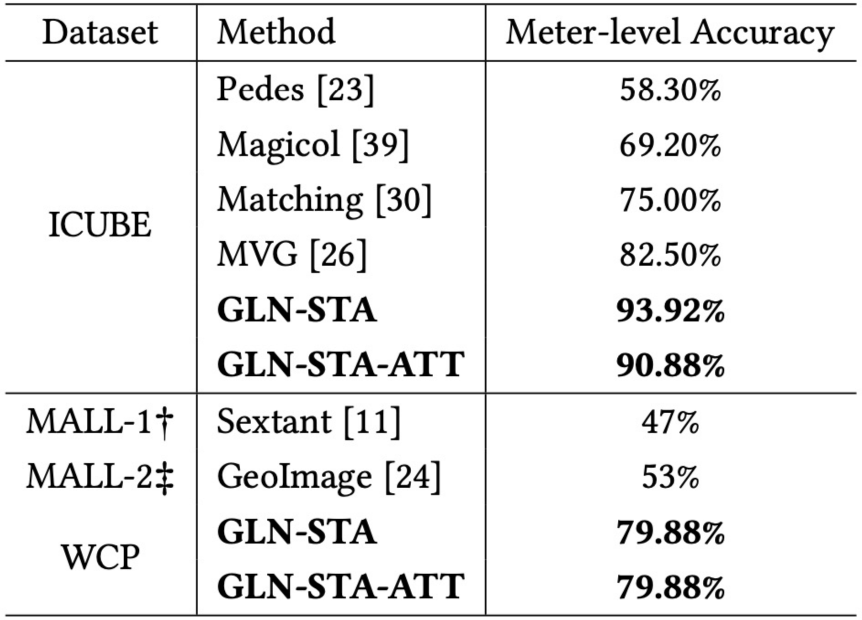
Prediction: $y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} F(x, y)$

Experiments & Results

Datasets

- We experimented our proposed approach on two datasets
- ICUBE: an existing dataset in an office building
  - 2,896 photos of 214 locations
  - Under standard setting, 1,712/1,184 images for training/testing
  - Under zero-shot setting, 1,368/1,528 images for training/testing where 102/112 locations as seen/unseen data
- WCP: collected ourselves in a shopping center
  - 3,280 photos of 394 locations
  - Under standard setting, 2,624/656 images for training/testing
  - Under zero-shot setting, 1,696/1,584 images for training/testing where 204/190 locations as seen/unseen data

Results on Standard Indoor Localization



Dataset	Method	Meter-level Accuracy
ICUBE	Pedes [23]	58.30%
	Magicol [39]	69.20%
	Matching [30]	75.00%
	MVG [26]	82.50%
	GLN-STA	**93.92%**
	GLN-STA-ATT	**90.88%**
MALL-1†	Sextant [11]	47%
MALL-2‡	GeoImage [24]	53%
WCP	**GLN-STA**	**79.88%**
	GLN-STA-ATT	**79.88%**

Table 1: Performance comparison with state-of-the-art models on ICUBE, WCP and the respective MALL datasets. Results of previous approaches on ICUBE are taken from [26], while results on distinct MALL datasets are taken from their respective papers. †MALL-1 consists of 108 locations and 686 images. ‡Mall-2 contains 20,000 images (locations).

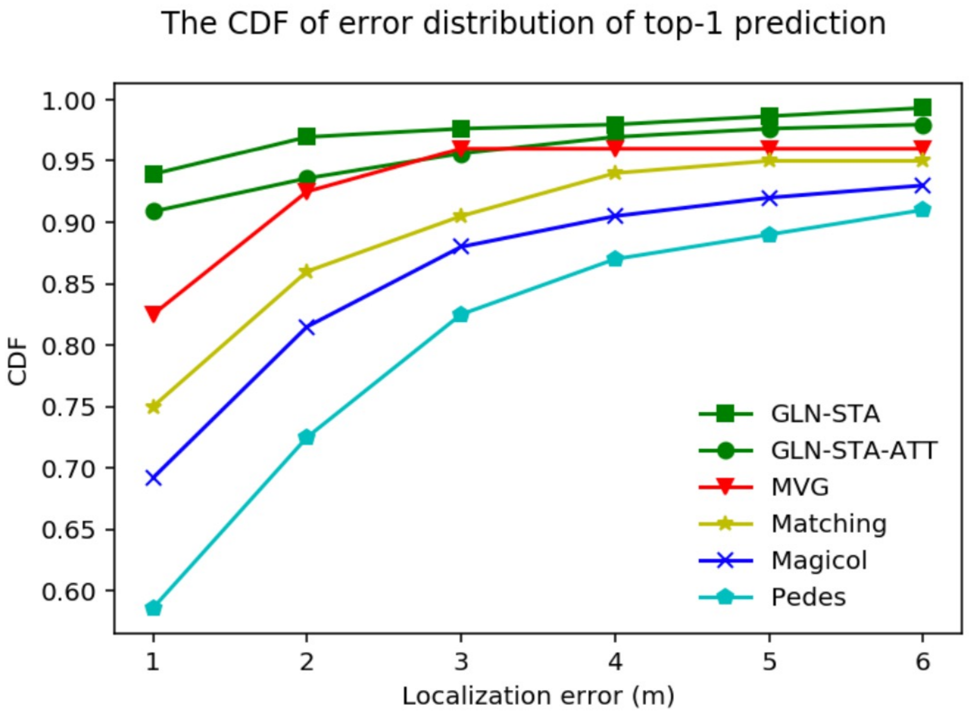
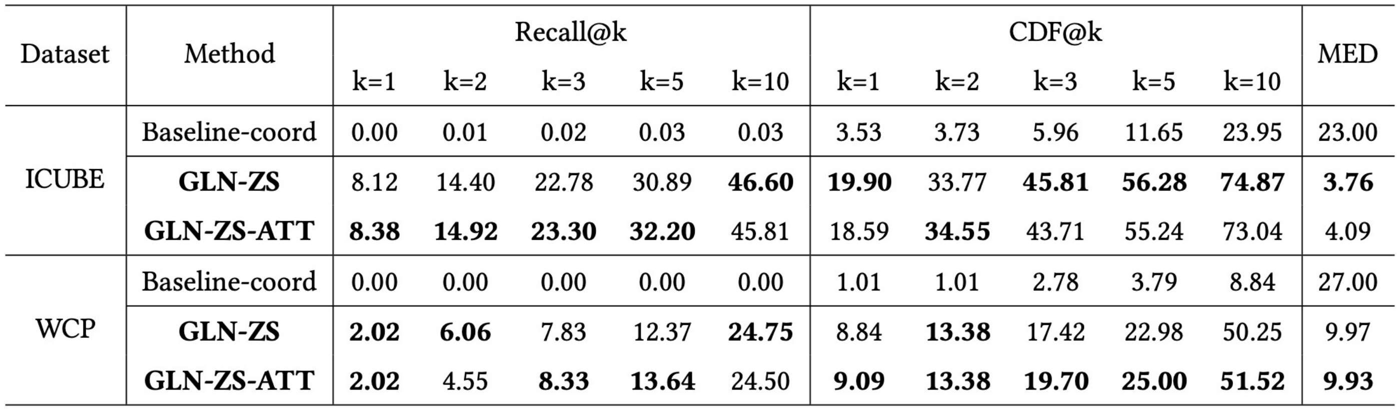


Figure 5: The cumulative distribution function (CDF) curves of the localization error of the previous and our approaches in standard indoor localization setting on ICUBE dataset.

Results on Zero-Shot Indoor Localization



Dataset	Method	Recall@k					CDF@k					MED
		k=1	k=2	k=3	k=5	k=10	k=1	k=2	k=3	k=5	k=10	
ICUBE	Baseline-coord	0.00	0.01	0.02	0.03	0.03	3.53	3.73	5.96	11.65	23.95	23.00
	GLN-ZS	8.12	14.40	22.78	30.89	**46.60**	**19.90**	33.77	**45.81**	**56.28**	**74.87**	**3.76**
	GLN-ZS-ATT	**8.38**	**14.92**	**23.30**	**32.20**	45.81	18.59	**34.55**	43.71	55.24	73.04	4.09
WCP	Baseline-coord	0.00	0.00	0.00	0.00	0.00	1.01	1.01	2.78	3.79	8.84	27.00
	GLN-ZS	**2.02**	**6.06**	7.83	12.37	**24.75**	8.84	**13.38**	17.42	22.98	50.25	9.97
	GLN-ZS-ATT	**2.02**	4.55	**8.33**	**13.64**	24.50	**9.09**	**13.38**	**19.70**	**25.00**	**51.52**	**9.93**

Table 2: Results of zero-shot indoor localization in comparison of *Recall@k*, *CDF@k* and *Median Error Distance* (MED) on ICUBE and WCP datasets. Note that numbers of recall and CDF are in % (the higher the better), while the numbers of median error distance are in meter (the lower the better). MED results are estimated with linear interpolation.

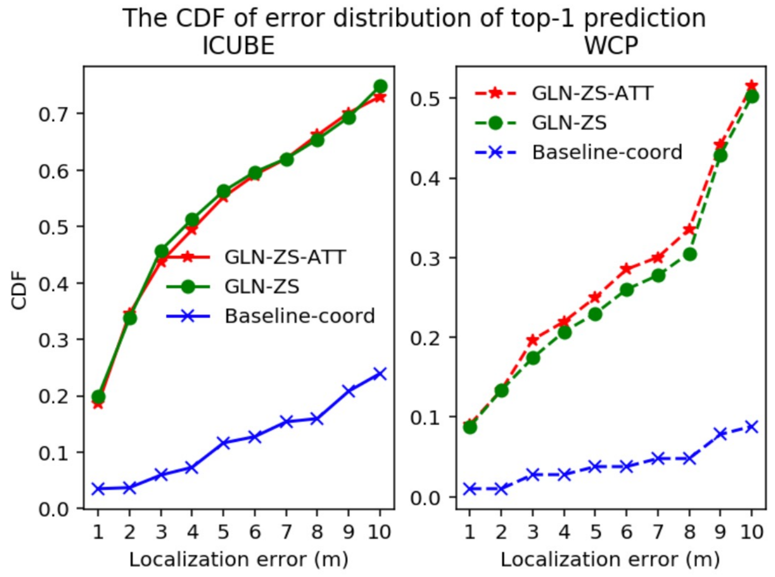


Figure 6: The cumulative distribution function (CDF) curves of the localization error of the zero-shot indoor localization experiments on ICUBE (left) and WCP (right) datasets.

Qualitative Results

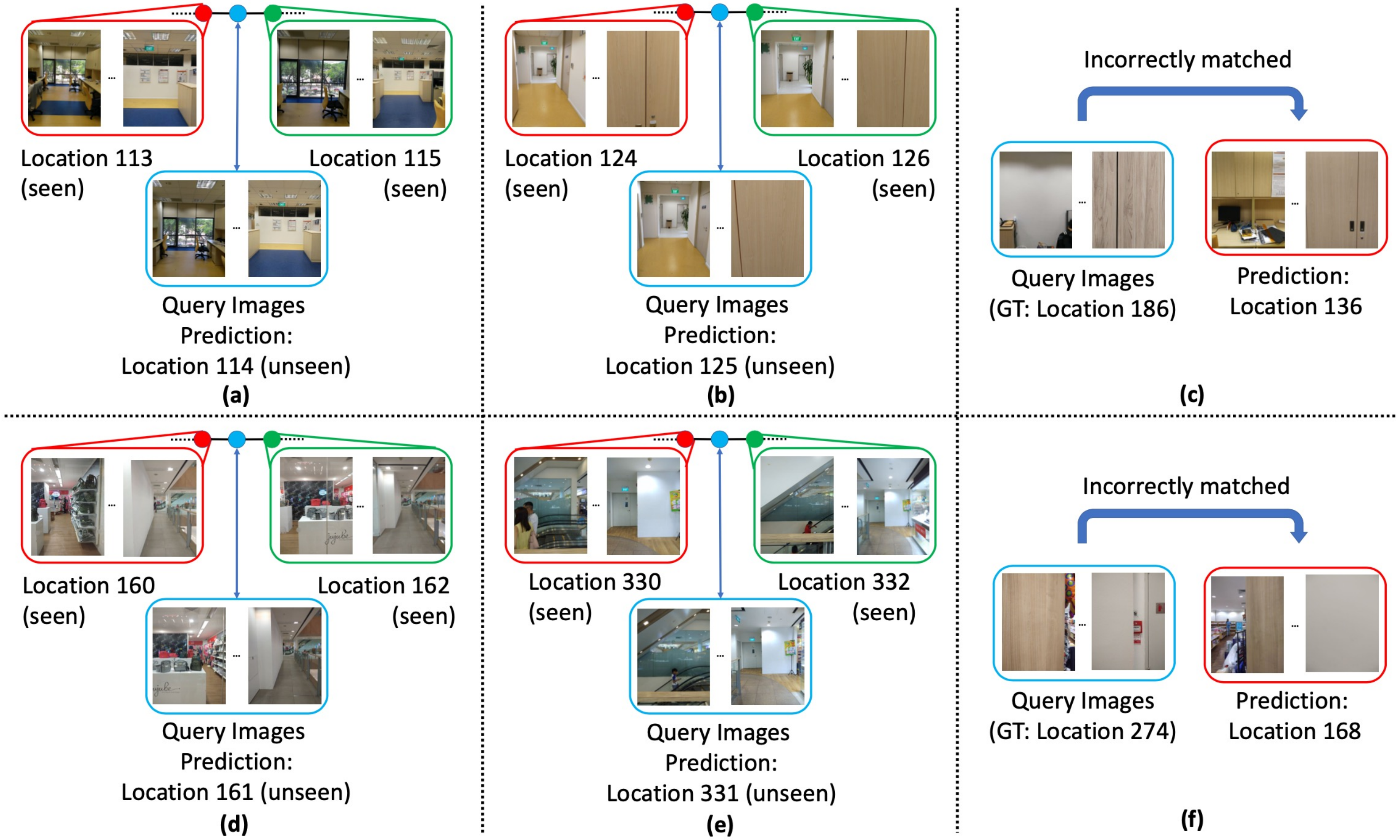


Figure 7: Qualitative results of zero-shot indoor localization on ICUBE (the top row) and WCP (the bottom row) dataset. The first two columns show examples of successful localization cases by utilizing the adjacency of seen classes to unseen classes, where the red, blue and green circles represent three adjacent locations. The last column shows examples of unsuccessful localization cases where our system is misled, especially when there are more query photos lacking distinguishable features.

Summary

- We propose a novel neural network based architecture Graph Location Networks (GLN) to perform multi-view indoor localization. GLN takes in different views and makes location predictions based on robust location representations with message-passing mechanism.
- We propose a novel, three-step zero-shot learning framework for indoor localization that can be applied to any indoor localization approach.
- We additionally contribute an indoor localization dataset, WCP.
- We show through quantitative and qualitative results that our model achieves state-of-the-art under standard setting and produces promising results under zero-shot setting.

Thank you for your attention!