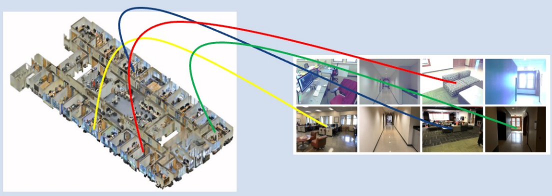
ETH zürich

Indoor Image Retrieval Using Monocular Scene Graphs

Joris Gentinetta, Jinhoo Kim, Vincent van der Brugge, David Zehnder
Advisor: Dr. Zuria Bauer, Mihai Dusmanu

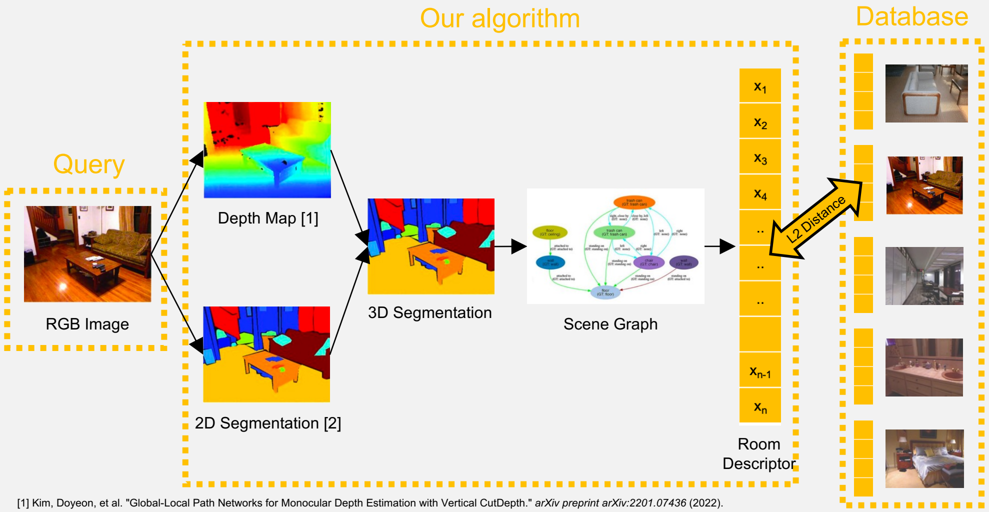
1 Motivation

- Context: Indoor image retrieval from single images
- Usage: Wide area from image search to visual localization for mobile robots or augmented reality applications



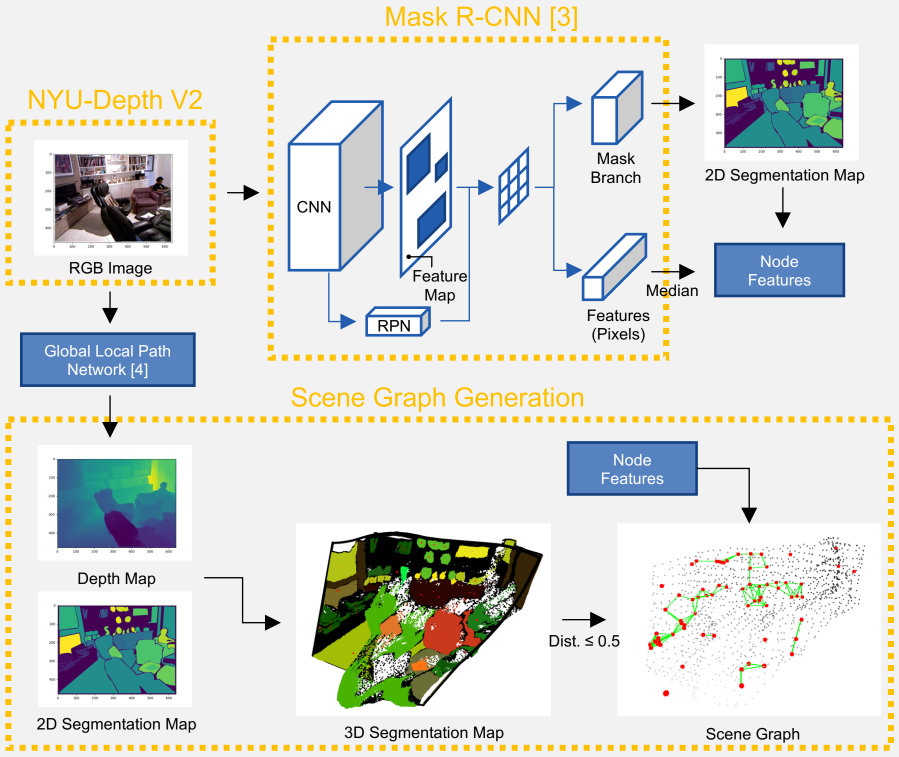
- Hard problem: Self-similarity, textureless areas, dynamic environments

2 Method Overview



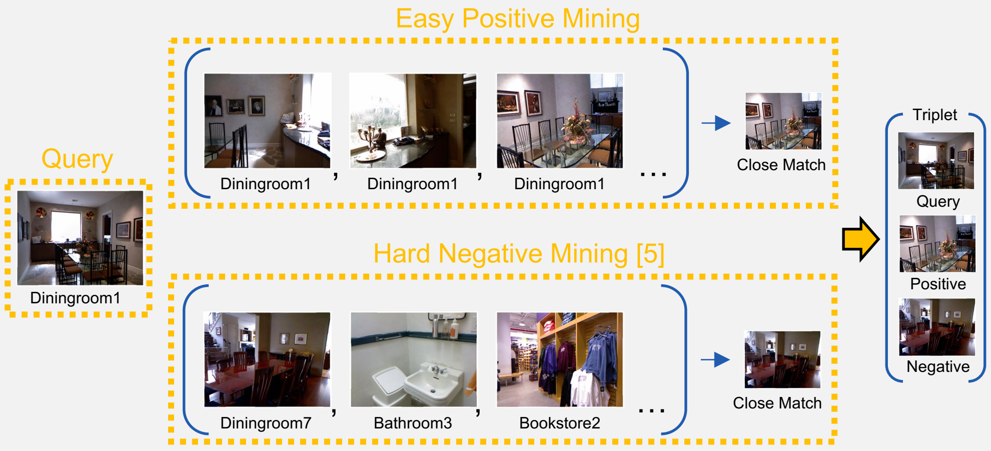
[1] Kim, Doyeon, et al. "Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth." *arXiv preprint arXiv:2201.07436* (2022).
[2] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

3 Scene Graph Generation



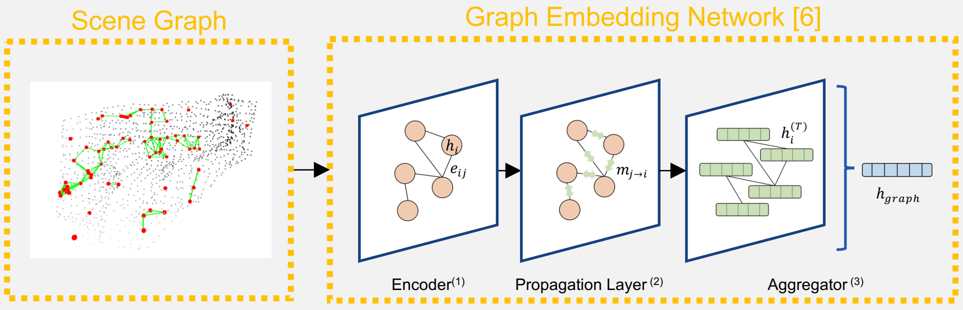
[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, & Ross B. Girshick (2017). Mask R-CNN. CoRR, abs/1703.06870.
[4] Doyeon Kim, Woonghyun Ga, Pyunghwan Ahn, Donggyu Joo, Sehwan Chun, & Junmo Kim (2022). Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth. CoRR, abs/2201.07436.

4 Easy Positive / Hard Negative Mining



[5] Hong Xuan, Abby Stylianou, Xiaotong Liu, & Robert Pless (2020). Hard negative examples are hard, but useful. CoRR, abs/2007.12749.

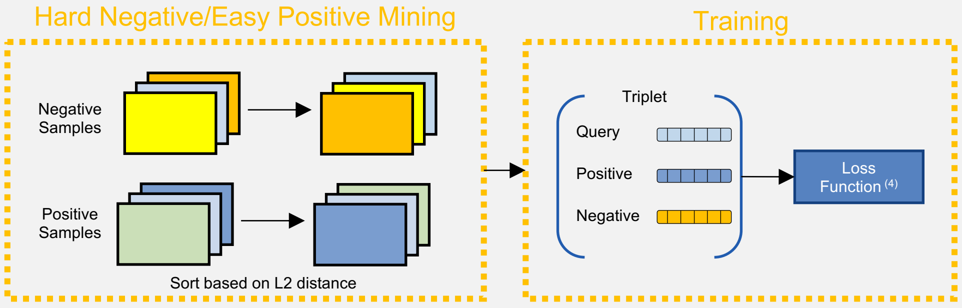
5 Graph Embedding Network



$$(1)\ h_i = MLP_{node}(x_i)$$

$$(2.1)\ m_{j\to i} = MLP_{message}(h_i, h_j, e_{ij}) \qquad (2.2)\ h_i^{(T)} = GRU_{node}(h_i, \sum_{j:(j,i)\in E} m_{j\to i})$$

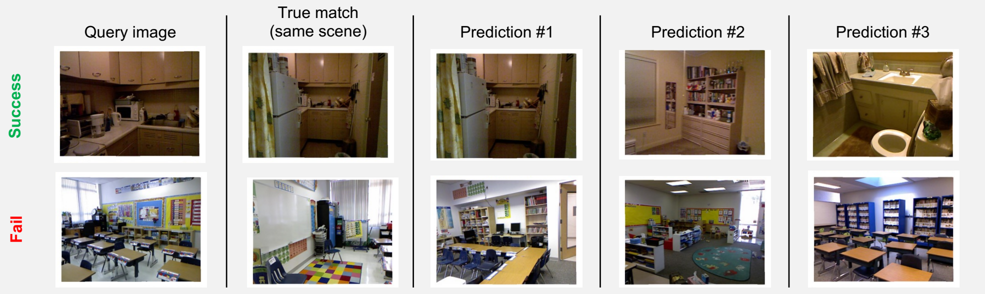
$$(3)\ h_{graph} = MLP_{graph}(\sum_{i\in V}(\sigma\left(MLP_{gate}\left(h_i^{(T)}\right)\right)\odot MLP\left(h_i^{(T)}\right))$$



$$(4)\ L_{triplet} = \mathbb{E}_{(G_Q,G_+,G_-)}[\max\{0, d(G_Q,G_+) - d(G_Q,G_-) + \gamma\}]$$

[6] Li, Y., Gu, C., Dullien, T., Vinyals, O., & Kohli, P.. (2019). Graph Matching Networks for Learning the Similarity of Graph Structured Objects.

6 Results and Discussion

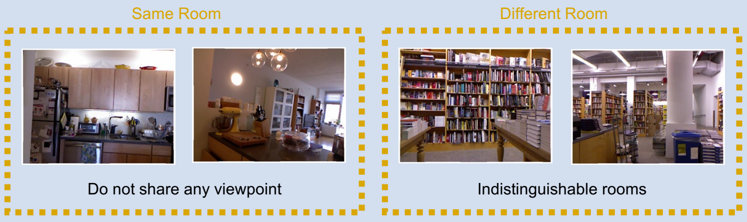


	Hard negative mining	Features	Depth	Segmentation	R@1	R@5	R@10	Triplet accuracy
Graph-based Best	Yes	Resnet	GT	Estimated	6.25%	20.96%	29.04%	79.33%
	Yes	Resnet	Estimated	Estimated	9.12%	26.84%	31.99%	80.22%
	Yes	One-hot	GT	GT	11.03%	28.31%	37.87%	86.76%
	Yes	Resnet	Estimated	GT	17.65%	39.71%	49.26%	91.20%
	No	Resnet	GT	GT	19.85%	41.54%	48.16%	91.40%
	+ Easy Positive	Resnet	GT	GT	17.65%	43.75%	52.57%	90.85%
	Yes	Resnet	GT	GT	23.16%	45.96%	55.88%	91.62%
CNN Best	Baseline 1: Resnet features		-	-	**53.31%**	**75.37%**	**81.25%**	**96.62%**
	Baseline 2: Bag of words		GT	GT	18.75%	33.46%	43.48%	-

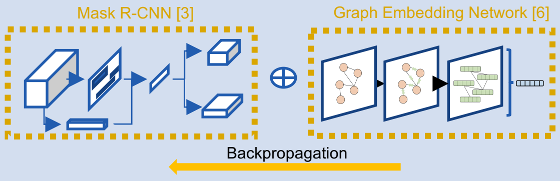
- Worse than Resnet baseline: can use lighting and room style for vastly different viewpoints
- Better than visual bag-of-words: Our model understands the spatial relationships and is more robust to the dynamic indoor scenes
- Segmentation is the bottleneck

7 Limitation and Future Work

- Limitations:



- End-to-end Learning:


Mask R-CNN [3]
⊕
Graph Embedding Network [6]
Backpropagation