Carnegie Mellon University

ARGO AI

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Website arXiv

Pix2Map: Cross-Modal Retrieval for Inferring Street Maps from Images

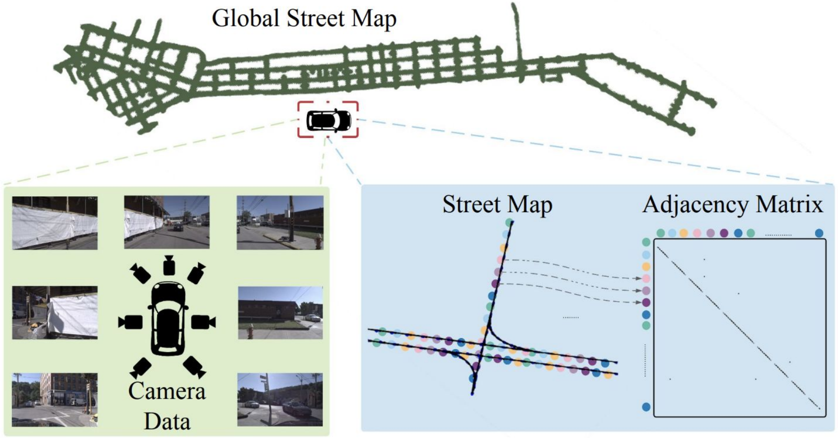
Xindi Wu KwunFung Lau Francesco Ferroni Aljoša Ošep Deva Ramanan

THU-AM-099

Preview

Task: Infer topological *road maps* from *images*.

Prior works: (jointly) learn a non-linear mapping from image pixels to bird's eye view maps, and estimate the road layout by generating a discrete spatial graph from detected lane markings.



Challenges: Learning to map *continuous* pixels to *discrete* graphs (maps) with varying numbers of nodes and topology in BEV is difficult.

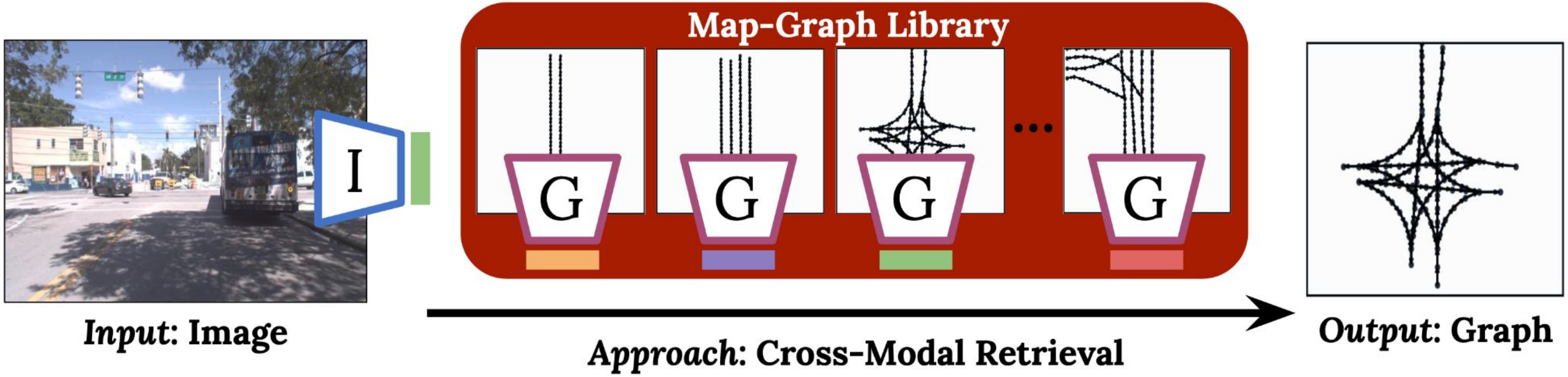
Preview

Key insight: this problem can be posed as cross-modal retrieval by learning a joint, cross-modal embedding space for images and existing maps.

Methods	Chamfer 10^1	RandLoss 10^{-2}	MMD 10^{-1}	U. density 10^{-1}
PINET	4.9244	10.8935	4.2983	2.8194
TOPO-PRNN	7.4811	9.2813	5.7726	3.9371
TOPO-TR	3.0140	**7.1603**	4.6431	2.2467
Pix2Map-Unimodal	4.3967	9.0764	4.1873	1.8391
Pix2Map-Single	2.6819	7.5204	4.0848	2.5339
Pix2Map (ours)	**2.0882**	7.7562	**3.9621**	**1.4354**

Pix2Map improves greatly over several SOTA methods!

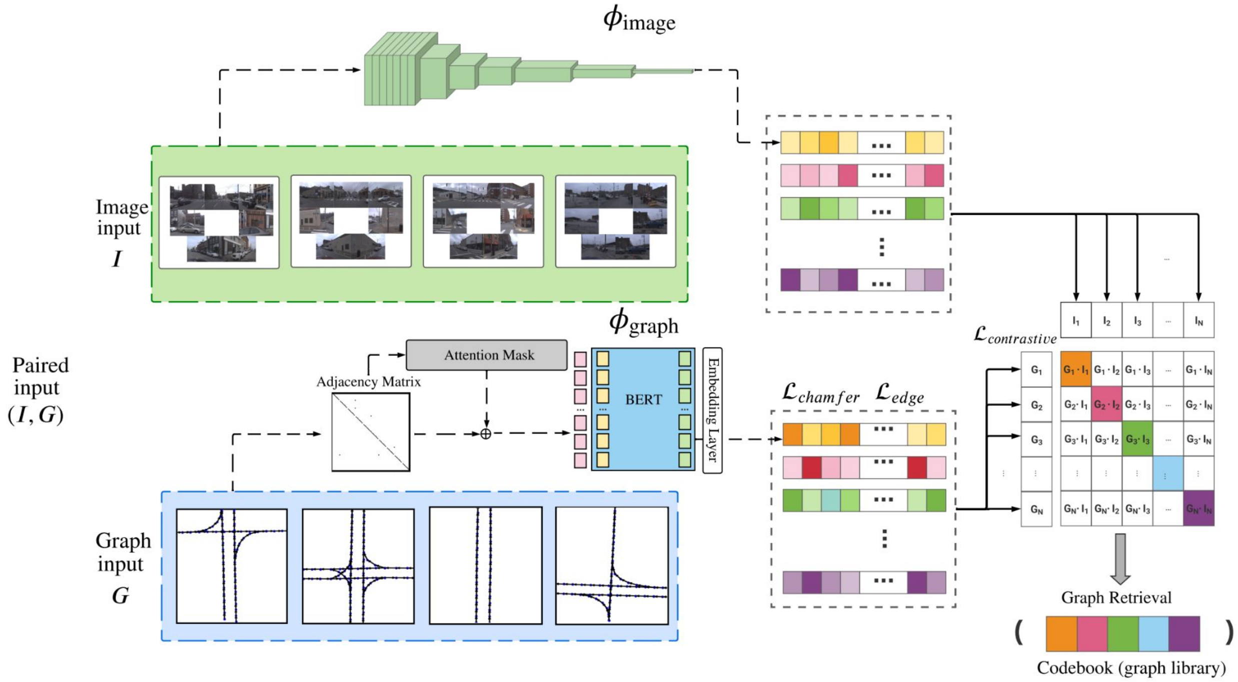
How to map images to discrete graphs?



Pix2Map returns the graph with the embedding most similar to input image via *cross-modal retrieval*!

Approach

Side-step landmark detection, localization, and graph generation by *learning* joint cross-modal embedding space.



Approach

- Mapping then boils down to *cross-modal retrieval* between encoded images and graphs in terms of cosine similarity.

$$\ell = \omega_1 \boxed{\ell_{contrastive}} + \omega_2 \ell_{chamfer} + \omega_3 \ell_{edge}$$

$$\ell_{contrastive} = \frac{1}{2N} \sum_{i=1}^{N} \left(\ell_i^{(I \to G)} + \ell_i^{(G \to I)} \right)$$

$$\ell_i^{(I \to G)} = -\log \frac{\exp \alpha_{ii}}{\sum_j \exp \alpha_{ij}},$$
$$\ell_i^{(G \to I)} = -\log \frac{\exp \alpha_{ii}}{\sum_j \exp \alpha_{ji}}.$$

$$\alpha_{ij} = \frac{\langle \phi_{\text{image}}(I_i), \phi_{\text{graph}}(G_j) \rangle}{||\phi_{\text{image}}(I_i)|| \, ||\phi_{\text{graph}}(G_j)||}.$$

Approach

- Mapping then boils down to *cross-modal retrieval* between encoded images and graphs in terms of cosine similarity.

$$\ell = \omega_1 \ell_{contrastive} + \omega_2 \boxed{\ell_{chamfer}} + \omega_3 \boxed{\ell_{edge}}$$

$$\ell_{chamfer} = \sum_{v \in V_0} \sum_i \alpha_i \text{Distance}(v, \pi_i(v)),$$

$$\ell_{edge} = \sum_{v,w \in V_0} \text{BCE}(\sum_i \alpha_i E_i(\pi_i(v), \pi_i(w)) + \epsilon, E_0(v, w))$$

$$\alpha_i = \text{softmax}_i\, \alpha_{i0}$$

Evaluation

Datasets: Argoverse: **camera ring + street maps** that capture the **geometry and connectivity** of road lanes for Pittsburgh and Miami.

Methods	Chamfer 10^1	RandLoss 10^{-2}	MMD 10^{-1}	U. density 10^{-1}	U. reach 10^{-1}	U. conn. 10^{-1}
PINET [31]	4.9244	10.8935	4.2983	2.8194	7.4194	2.9231
TOPO-PRNN [11]	7.4811	9.2813	5.7726	3.9371	6.8297	1.3934
TOPO-TR [11]	3.0140	**7.1603**	4.6431	2.2467	3.3091	1.1530
Pix2Map-Unimodal	4.3967	9.0764	4.1873	1.8391	3.2746	1.7734
Pix2Map-Single	2.6819	7.5204	4.0848	2.5339	**3.0134**	**1.0291**
Pix2Map (ours)	**2.0882**	7.7562	**3.9621**	**1.4354**	3.2893	1.5532

Key Results: Pix2Map outperforms baselines by a large margin.
Our method is especially strong in terms of preserving the spatial point discrepancy.

Evaluation

Cross-modal > Unimodal:
Cross-modal retrieval can exploit graph embedding space, it regularizes retrieval.

Methods	Chamfer 10^1	RandLoss 10^{-2}	MMD 10^{-1}	U. density 10^{-1}	U. reach 10^{-1}	U. conn. 10^{-1}
Unimodal	3.2168	9.7596	7.7671	0.7365	3.9452	1.3661
Ours	1.5908	7.3283	3.0888	0.7593	3.2997	0.8397
Ours++	1.5208	6.1504	3.0944	0.7407	3.2610	0.8089

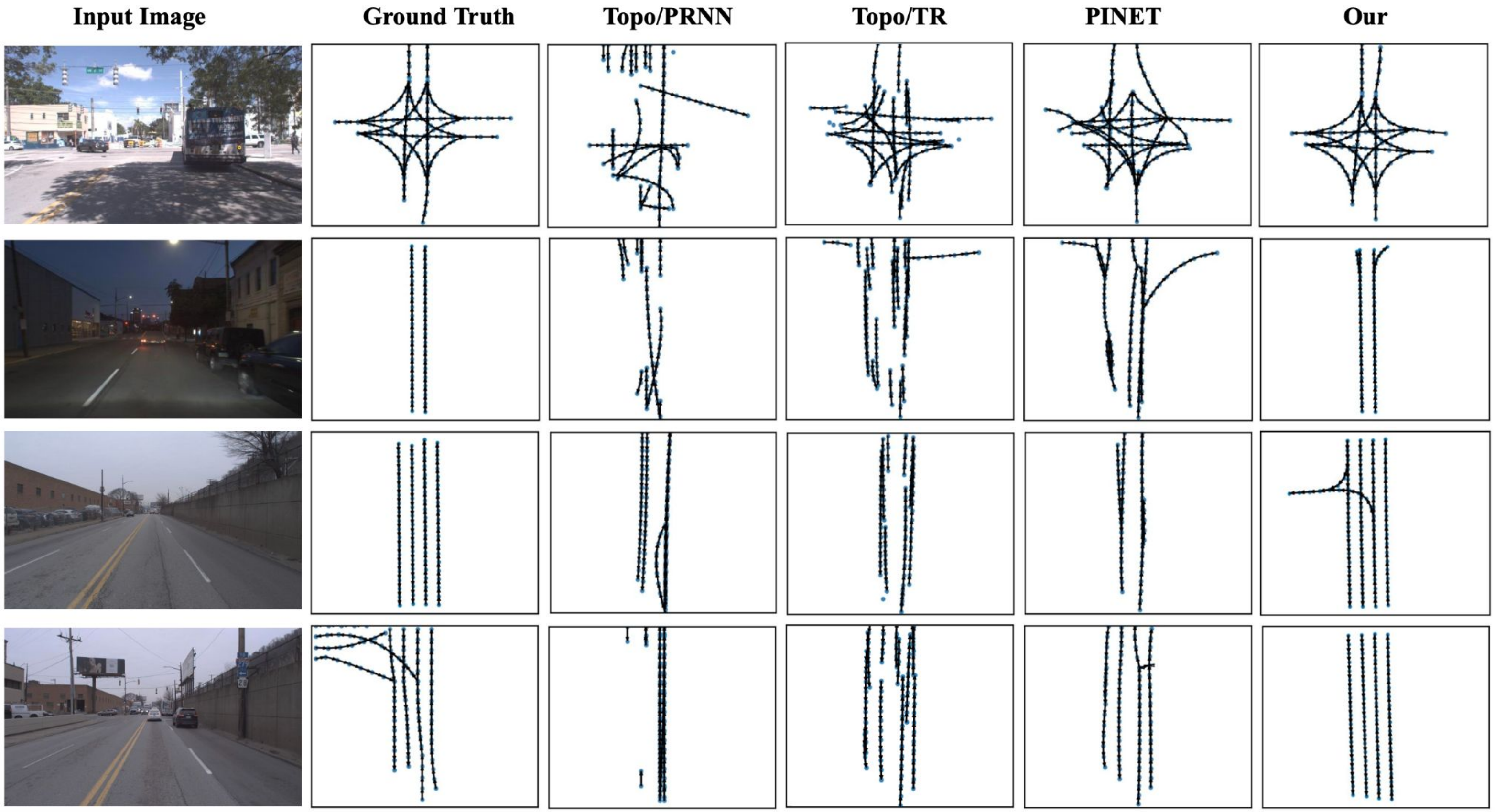
Evaluation

Augmenting Map-Graph Library:

By simply expanding our graph library we consistently improve the performance.

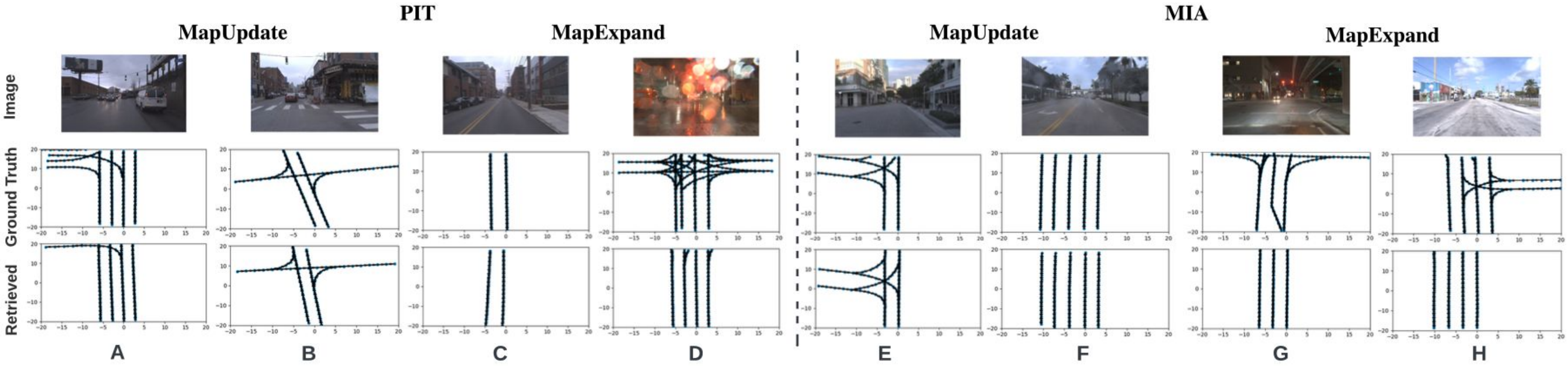
City	Library Size 10^1	Chamfer 10^{-2}	RandLoss 10^{-1}	MMD 10^{-1}	U. density 10^{-1}	U. reach 10^{-1}	U. conn.
PIT	5.7k	1.5908	7.3283	3.0888	0.7593	3.2997	0.8397
	10k	1.6457	7.6247	3.2848	**0.7264**	4.5891	1.6364
	20k	1.5369	6.5373	3.1883	0.7581	3.2902	1.0602
	30k	1.5239	6.6553	**3.0253**	0.8586	4.0642	0.9615
	40k	**1.5208**	**6.1504**	3.0944	0.7407	**3.2610**	**0.8089**
MIA	7.4k	1.4747	6.8693	3.4033	1.0948	4.6253	1.1910
	10k	1.4991	6.2315	3.3118	1.2784	5.5209	1.3679
	20k	1.3878	8.0234	3.3910	1.1290	4.1237	1.3249
	30k	1.4012	7.1898	3.2773	1.2444	4.2471	1.2298
	40k	1.3878	7.6305	3.3351	1.2523	5.3894	1.3385
	60k	1.3080	6.3369	3.18879	1.1972	4.7578	1.1977
	80k	1.2711	6.2852	3.19506	1.0123	4.6827	1.1651
	100k	**1.2462**	**6.2740**	**3.1277**	**0.9884**	**3.8521**	**1.1397**

Qualitative results



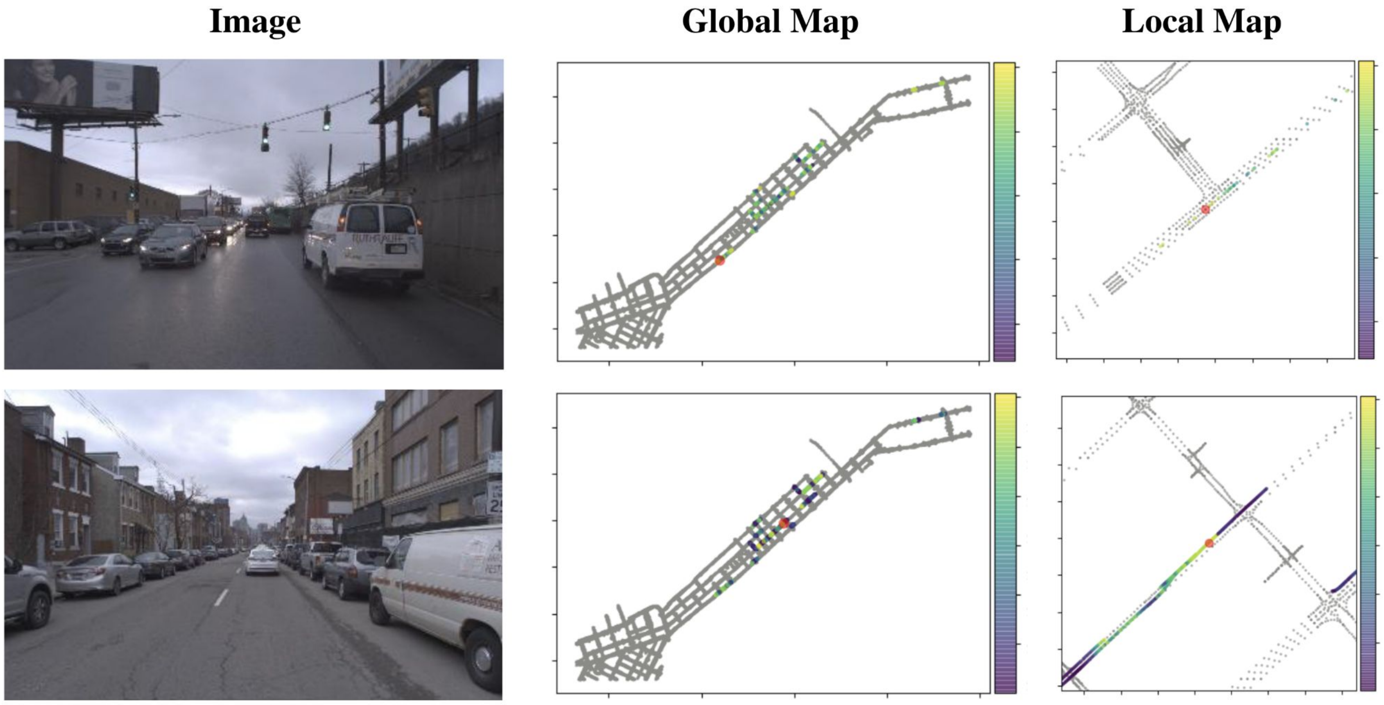
Applications

Map Expansion and Update:



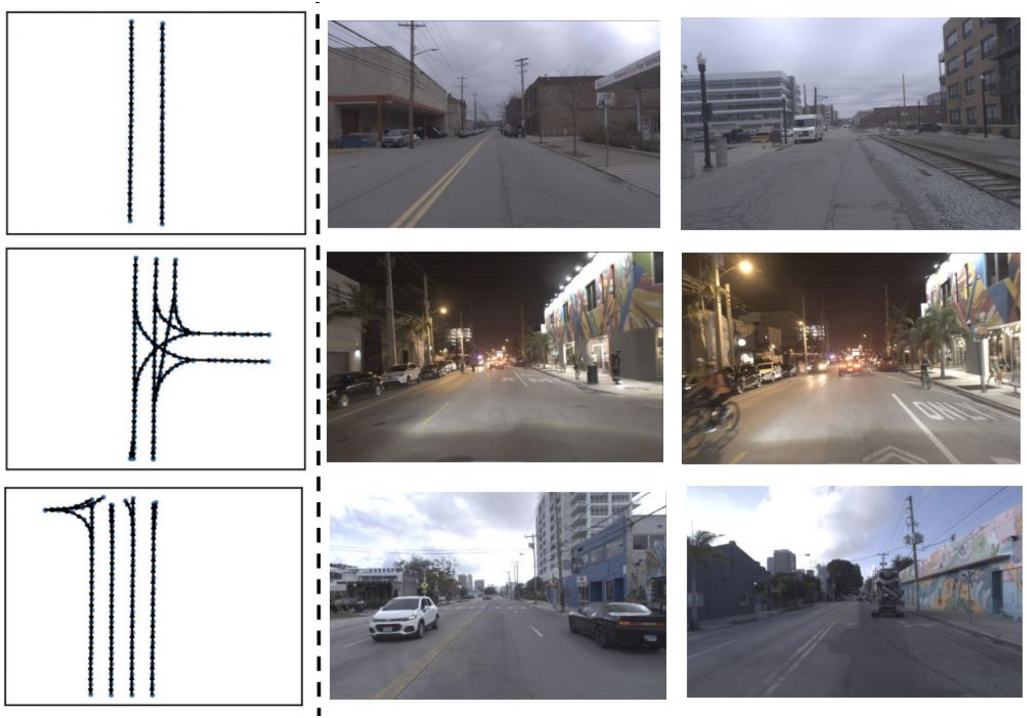
Applications

Visual Localization: Localize image based on graph similarity in the global map.



Applications

Map2Pix: Retrieve image from a graph.



Website arXiv

Thank You!

Carnegie Mellon University

ARGO AI