# Supplementary material for Pix2map:
# Cross-modal Retrieval for Inferring Street Maps from Images

Xindi Wu [1] KwunFung Lau[1] Francesco Ferroni[2] Aljoša Ošep[1] Deva Ramanan[1,2]
[1]Carnegie Mellon University [2]Argo AI
{aosep, deva}@andrew.cmu.edu, xindiw@princeton.edu, kwun.fung.lau@intel.com

## Abstract

*In this supplement, we provide various experiments to illustrate the practical uses of Pix2Map. These experiments include:*

- *Map Expansion and Update, in which we present experiments on expanding and updating existing maps,*
- *Visual Localization, by generating a heatmap of possible locations for the ego-vehicle on a city-level map,*
- *Map2Pix, which is visually demonstrated by retrieving ego-camera images using street maps.*

## 1. Applications

In this section, we discuss how our method can be used for practical purposes, and show that graph library retrieval can greatly improve various downstream applications such as expansion (*MapExpand*) and update (*MapUpdate*) given existing maps, visual image-to-HD map localization and *Map2Pix*.

### 1.1. Map Expansion and Update

We use our graph retrieval method to mimic map expansion (*MapExpand*) and map update (*MapUpdate*) using data splits. For map expansion, we retrieve local graphs corresponding to recordings obtained in a "new traversal" to expand the existing map. For map updates, we similarly retrieve local maps to update the global map.

We qualitatively evaluate the graph retrieval results in Fig. 6 in the main paper. Please see the caption for a detailed description, but generally speaking, we find *Pix2Map* returns reasonable graphs similar to the ground truth. In Tab. 1, we evaluate the performance of map update and map expansion in two cities (Pittsburgh and Miami). We do so by comparing expanded/updated maps with ground-truth maps using metrics. As shown above, map expansion to novel areas is harder than updating previously-seen areas.

| City | Task type | Chamfer<br>$10^1$ | RandLoss<br>$10^{-2}$ | MMD<br>$10^{-1}$ | U. density<br>$10^{-1}$ | U. reach<br>$10^{-1}$ | U. conn.<br>$10^{-1}$ |
|---|---|---|---|---|---|---|---|
| PIT | MapUpdate | 1.5908 | 7.3283 | 3.0888 | 0.7593 | 3.2997 | 0.8397 |
| | MapExpand | 2.6654 | 16.9768 | 8.0468 | 3.9482 | 4.2949 | 3.9699 |
| MIA | MapUpdate | 1.4747 | 6.8693 | 3.4033 | 1.0948 | 5.5333 | 1.1910 |
| | MapExpand | 2.0637 | 11.1354 | 4.2605 | 1.4922 | 4.7318 | 1.5940 |

Table 1. **Map update and expansion evaluation.** As can be seen, map expansion to novel areas can be much harder than updating previously-seen areas.

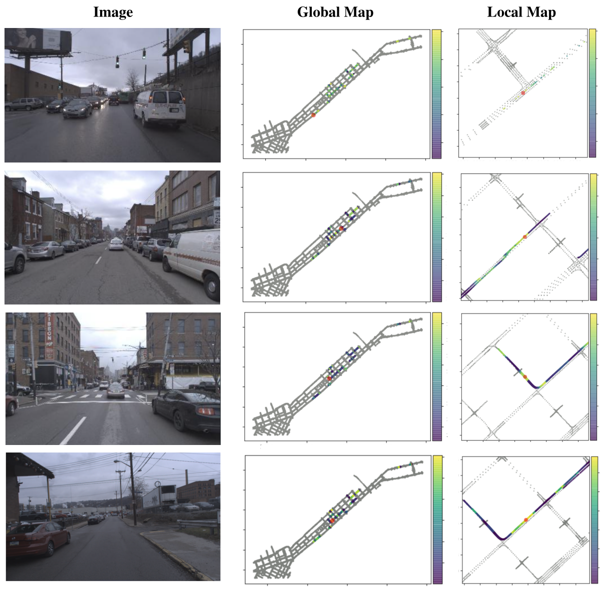

Figure 1. **Visual localization via Pix2Map.** We overlay retrieval scores on the corresponding local graphs from the original city map, generating a graph "heatmap" of possible locations given instantaneous ego-view images. We plot the ground-truth location as a red dot. In general, ground-truth locations tend to lie in high-scoring (yellow) regions. For example, the top ground truth corresponds to an intersection, while other high-scoring regions also tend to be graph intersections as well. Given a sequence of images, one may be able to reduce the ambiguity over time [1]

### 1.2. Localization

Furthermore, our method demonstrates great visual localization ability based on visual and geometric understand-

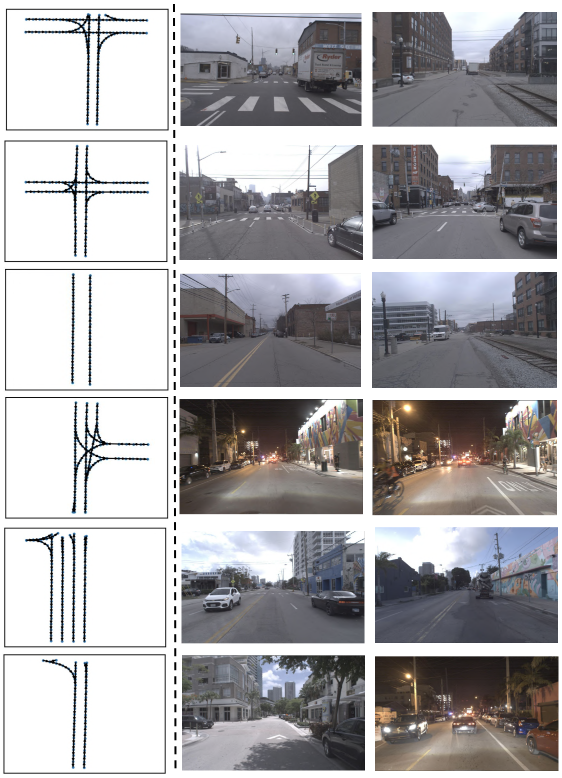

Figure 2. **Qualitative results for Map2Pix.** The goal is to retrieve ego-camera images given a street map. Such image retrieval may be useful for simulator-based training and validation of autonomous stacks. A single street geometry might retrieve multiple consistent, realistic imagery.

ing. We use the cosine similarities of retrieved graphs to generate a heatmap of possible ego-vehicle locations over a city-level map, showing the locations where their corresponding graphs are assigned a high likelihood, relative to the ground truth location shown as a red dot. While the ground truth is usually assigned a high likelihood, which indicates the promising performance of localization ability, the distribution becomes less sharp with respect to position when farther away from intersections. See Fig. 1 for more details.

### 1.3. Map2Pix

We further show that it is also possible to retrieve egocentric camera data using street maps. Such techniques could be used in the future to synthesize virtual worlds consistent with the query road geometry. We provide a few example image retrievals in Figure 2, visualizing the front views of the top $K = 2$ images for each street map. As can be seen, the retrieved images correspond to rough geometric layouts encoded in the query graphs.

## References

[1] Marcus A Brubaker, Andreas Geiger, and Raquel Urtasun. Map-based probabilistic visual self-localization. *IEEE TPAMI*, 2015. 1