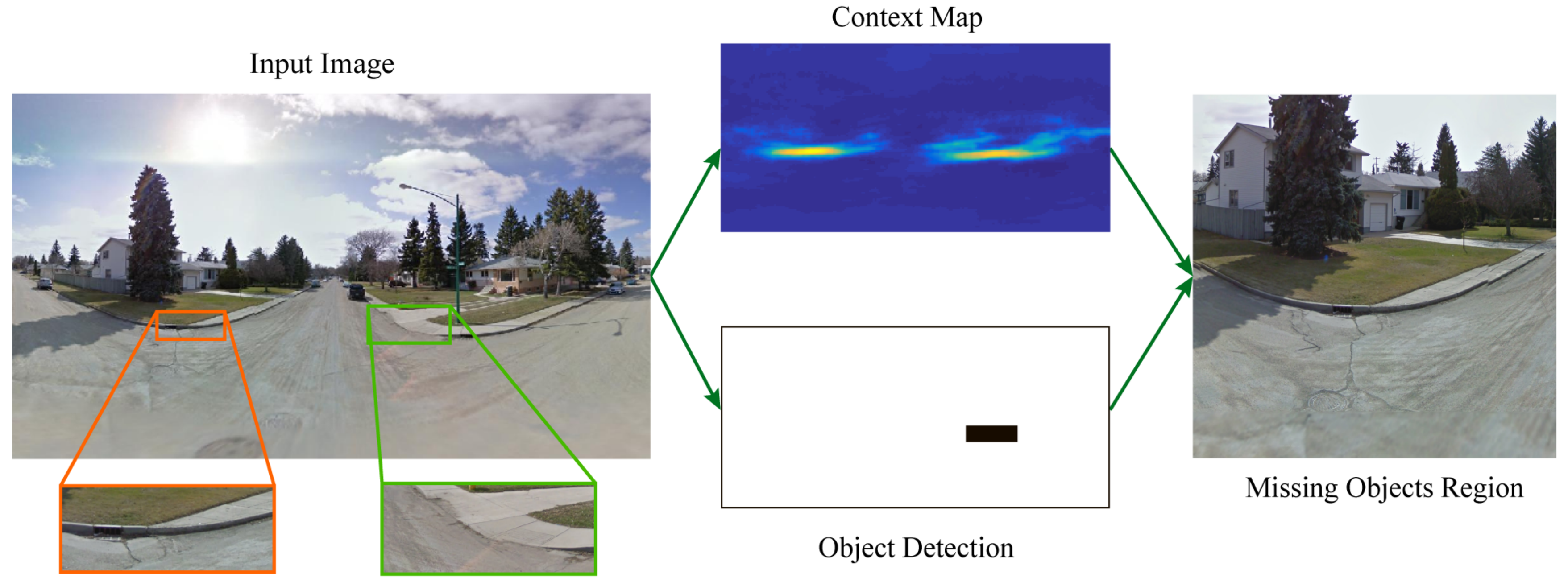
Seeing What is Not There: Learning Context to Determine Where Objects Are Missing

Jin Sun and David W. Jacobs
University of Maryland

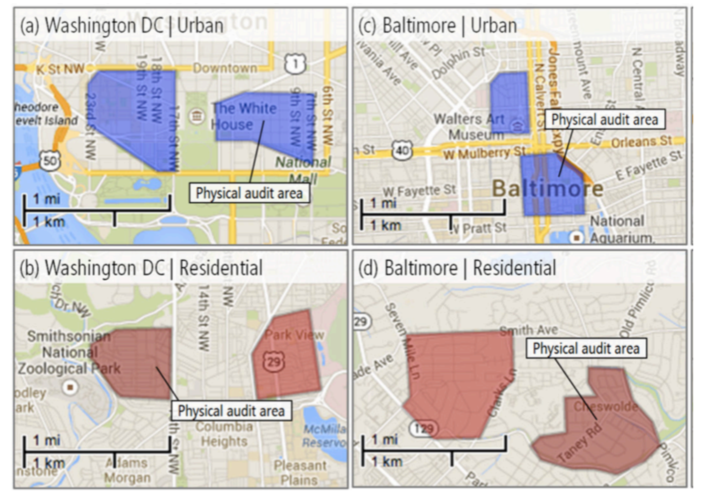
IEEE 2017 Conference on Computer Vision and Pattern Recognition

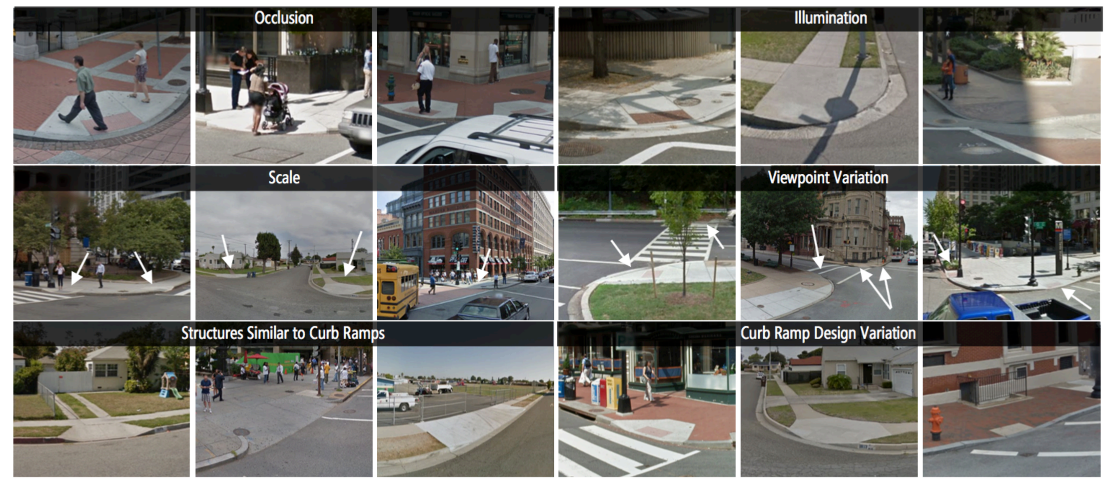
Introduction:

- Most of computer vision focuses on what is in an image.
- We train a standalone object-centric context representation to perform the *opposite* task: seeing what is not there.
- Missing objects region: where context says yes but the detector says no. Our core idea is the object-context decomposition.



- A challenging curb ramp dataset of four cities with large variations.



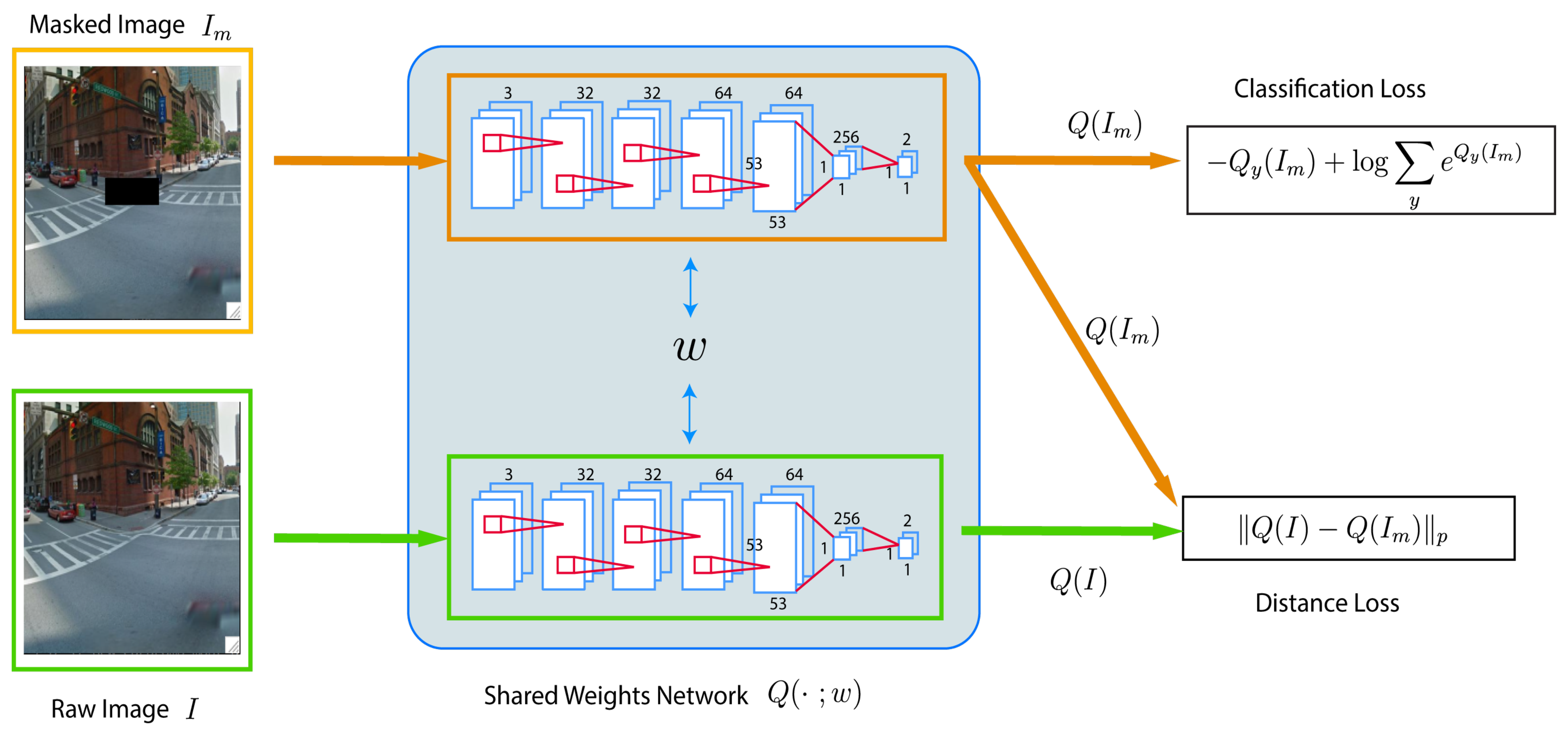


- Context plays a critical role in finding curb ramps: the appearance alone can be ambiguous.

Approach:

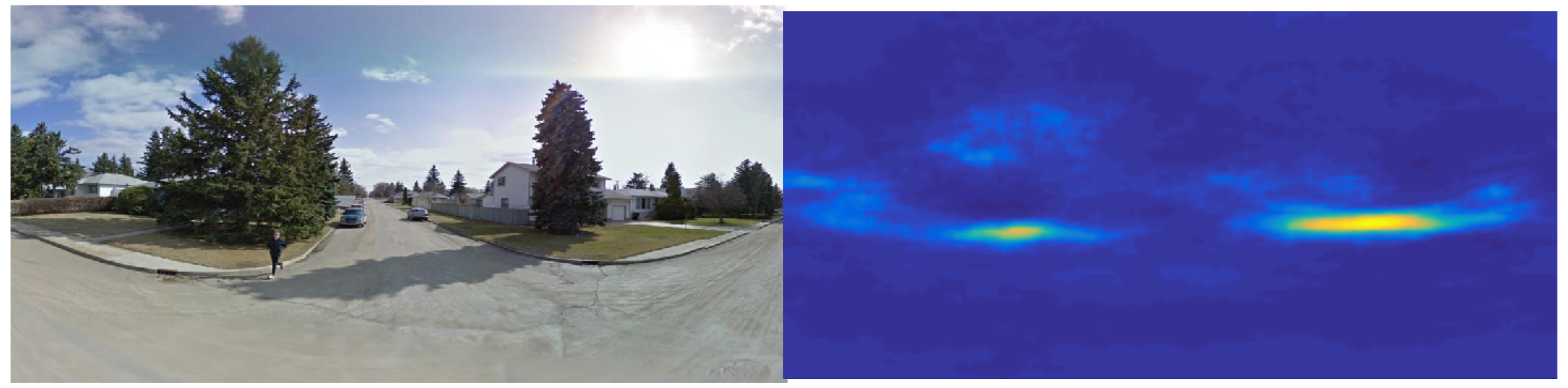
- With a specially designed training strategy, our Convolutional Neural Network model learns to ignore objects and focus on context only. It is fully convolutional thus highly efficient.

Siamese trained Fully convolutional Context Network



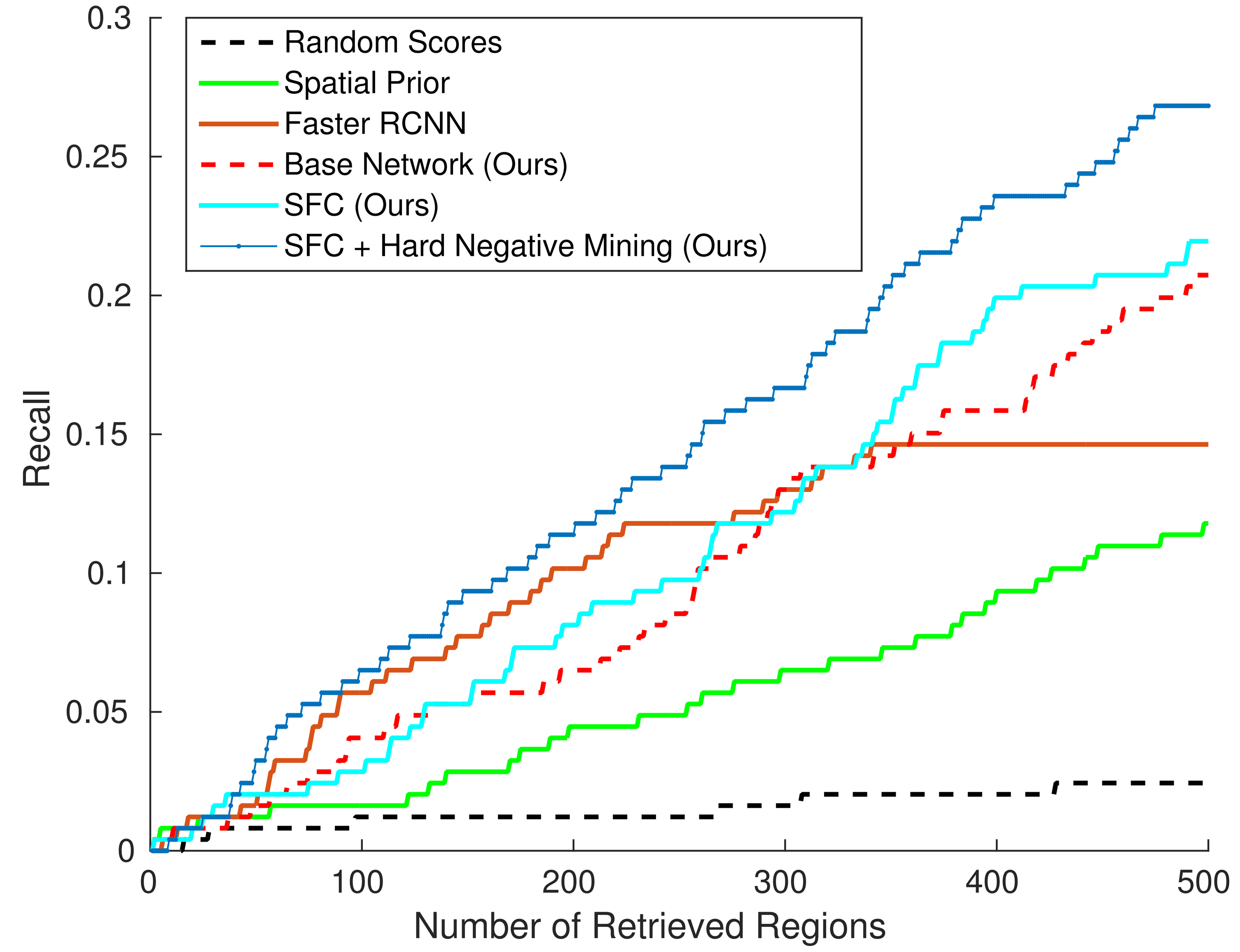
- The classification loss encourages the learning of useful context-only features with objects being masked out.
- The distance loss enforces the network to extract same features regardless of whether an image is masked or not.
- The SFC network is fully convolutional and doesn't require masks during testing: this gives a **speed-up** of a factor of 60, comparing to a naïve approach of directly training a classifier with masked images.
- During training, the efficient fully convolutional structure allows us to do **hard negative mining** for improved performance.

- For a given Google Street View panorama image, the SFC network produces a high resolution probabilistic map of where a curb ramp should exist.



Results:

- Experiments show the effectiveness of the SFC network in finding missing curb ramp regions: our best method (SFC + Hard Negative Mining) can find 27% of true missing curb ramp regions.



- Note that the SFC network doesn't need any missing curb ramp labels! We learn context from normal curb ramp labels, which are more abundant and much easier to collect.

- Efficiency:
  - 20 mins human time → 500 regions → 27% recall.
  - 2,820 intersections in Manhattan (1.6 million population) → a few hours.
- Examples of found missing curb ramp regions: