Trends in 3D:
3D Perception Through Exclusive Priors and Deep Learning

Amir R. Zamir
Stanford University

OMITTED (unpublished)

3D From a Single Image Using Exclusive Priors

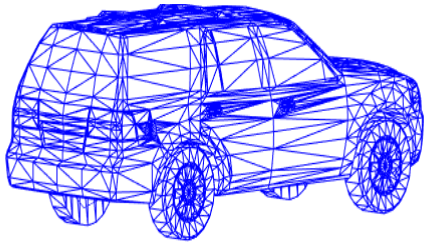
Multi-View 3D

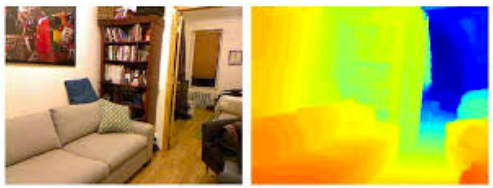
- No inherent ambiguity

Single-View 3D

Inherently ambiguous ←→ prior information

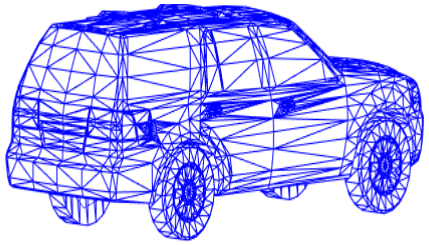
Priors!

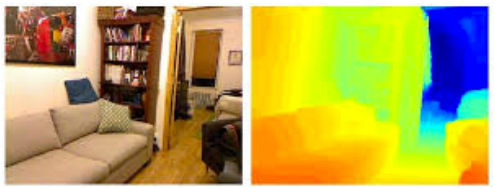


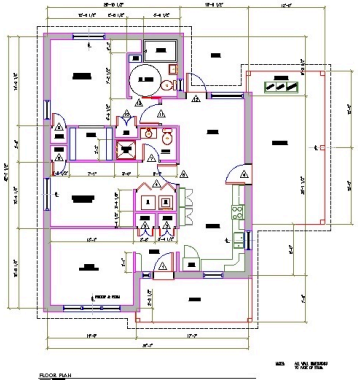


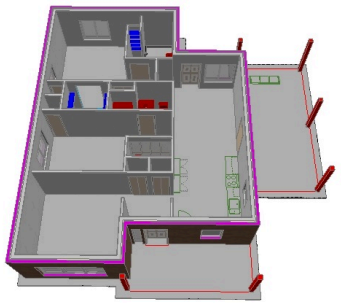
1. Generic

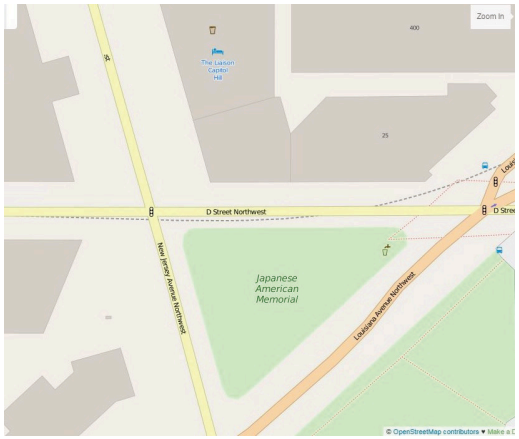
Priors!

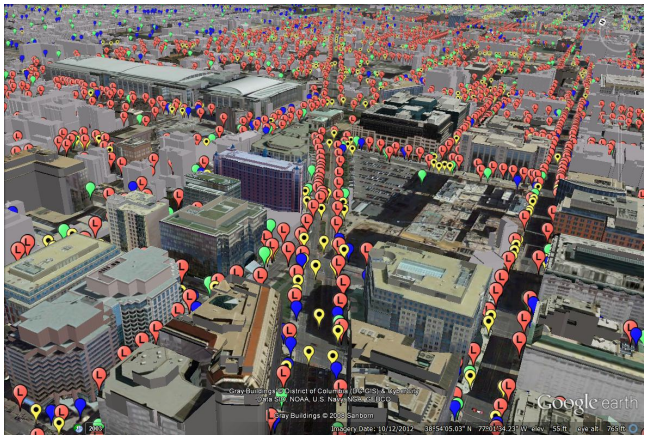








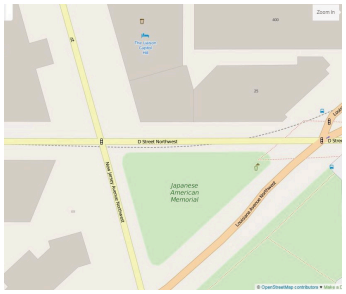




1. Generic

2. Exclusive and specific

Exclusive Priors

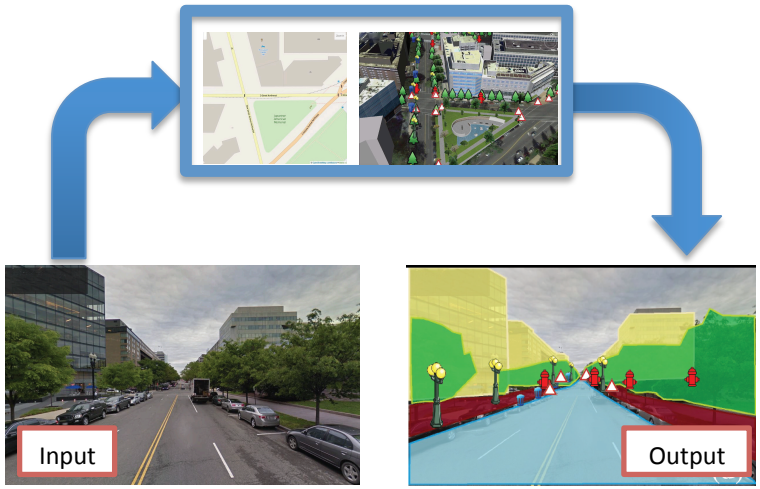


Exclusive Priors

- Holistic 3D Scene Understanding from a Single Geo-tagged Image, In CVPR15.
- Geo-semantic Segmentation, In CVPR15.
- GIS-Assisted Object Detection and Geospatial Localization, In ECCV14.

Exclusive Priors

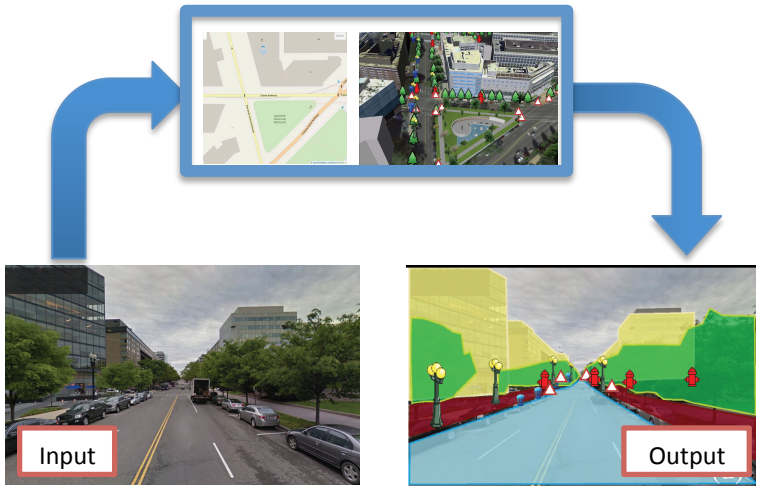
- Reduce “3D from single image” to “alignment” problem.
  - Advantages:
    - Computationally less complex
    - Less reliance on semantics; Better generalization
    - Beyond-Image 3D
  - Disadvantages:
    - More dependency on meta-data
    - Coverage
    - Cross-modality



Exclusive Priors

- Reduce “3D from single image” to “alignment” problem.
  - Advantages:
    - Computationally less complex
    - Less reliance on semantics; Better generalization
    - Beyond-Image 3D
  - Disadvantages:
    - More dependency on meta-data
    - Coverage
    - Cross-modality

- **3D through post-alignment transferring**
- **Robustness**

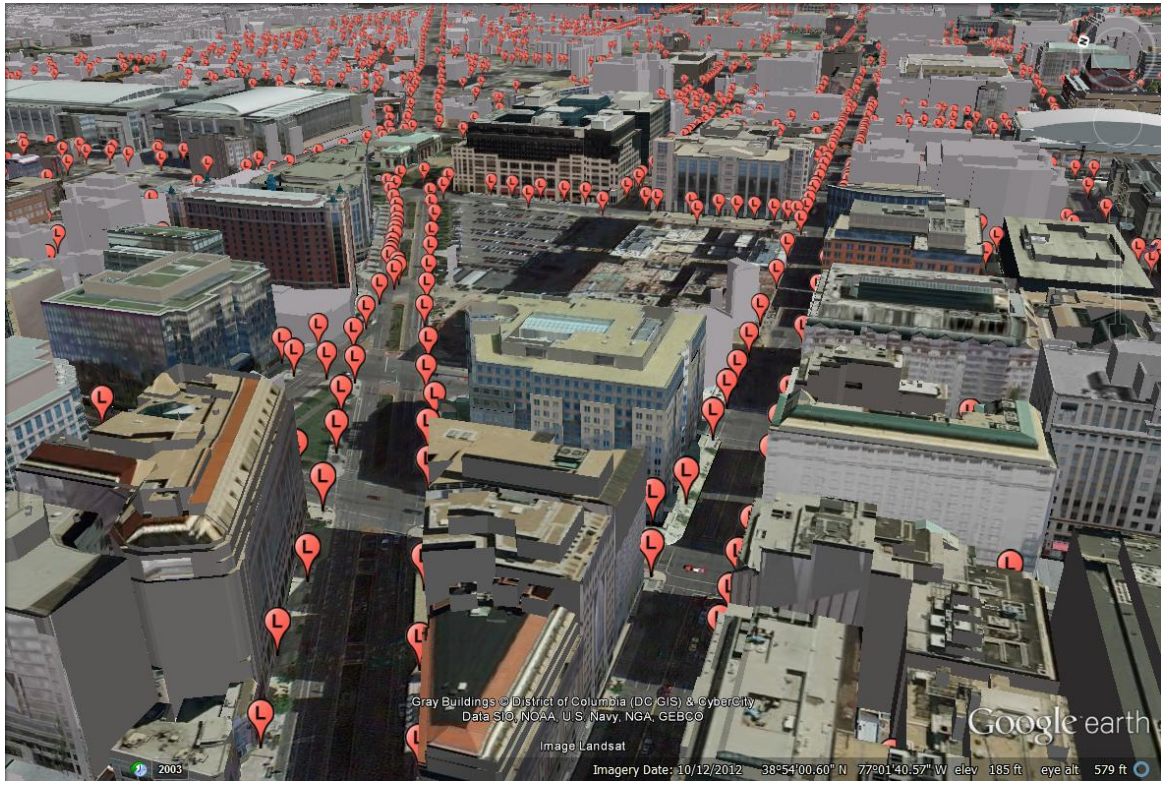


Object Detection and Scene Alignment Using Object Priors

GIS-Assisted Object Detection and Geospatial Localization. Ardeshir, Zamir, Shah. In ***ECCV***, 2014.

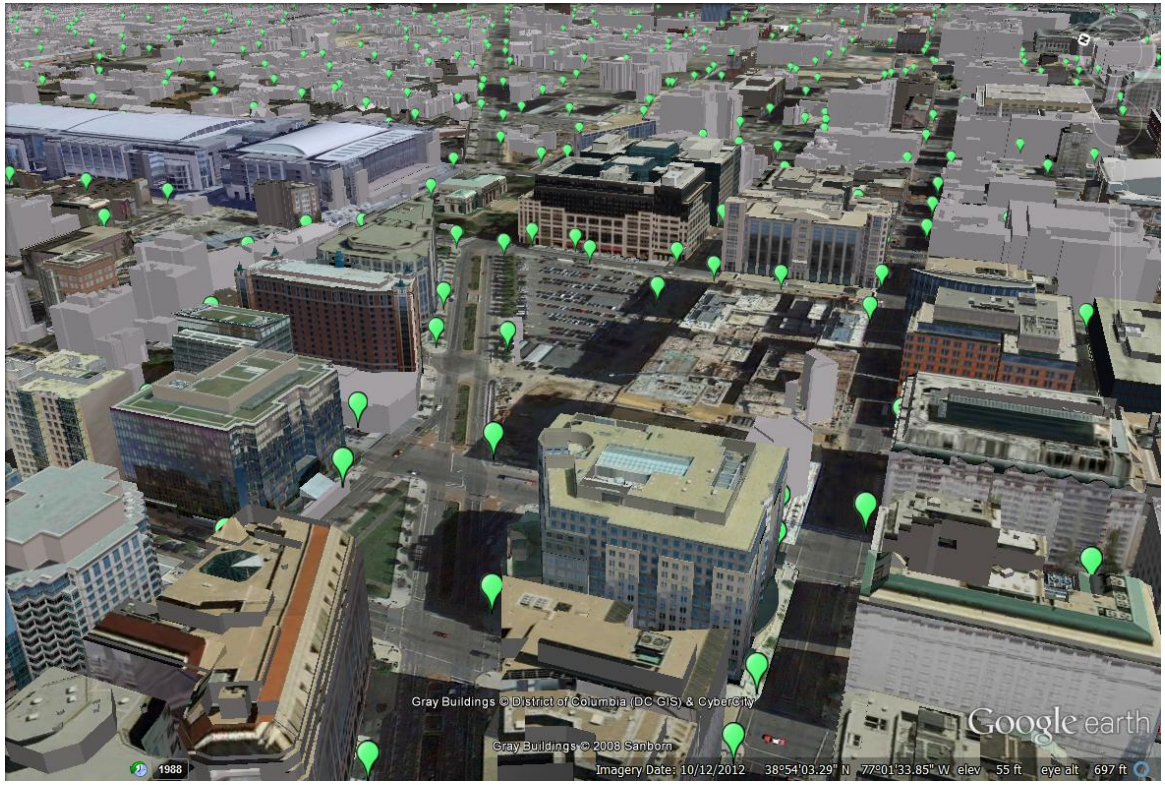
GIS Dataset

- e.g. Washington D.C.
  - Lamp posts



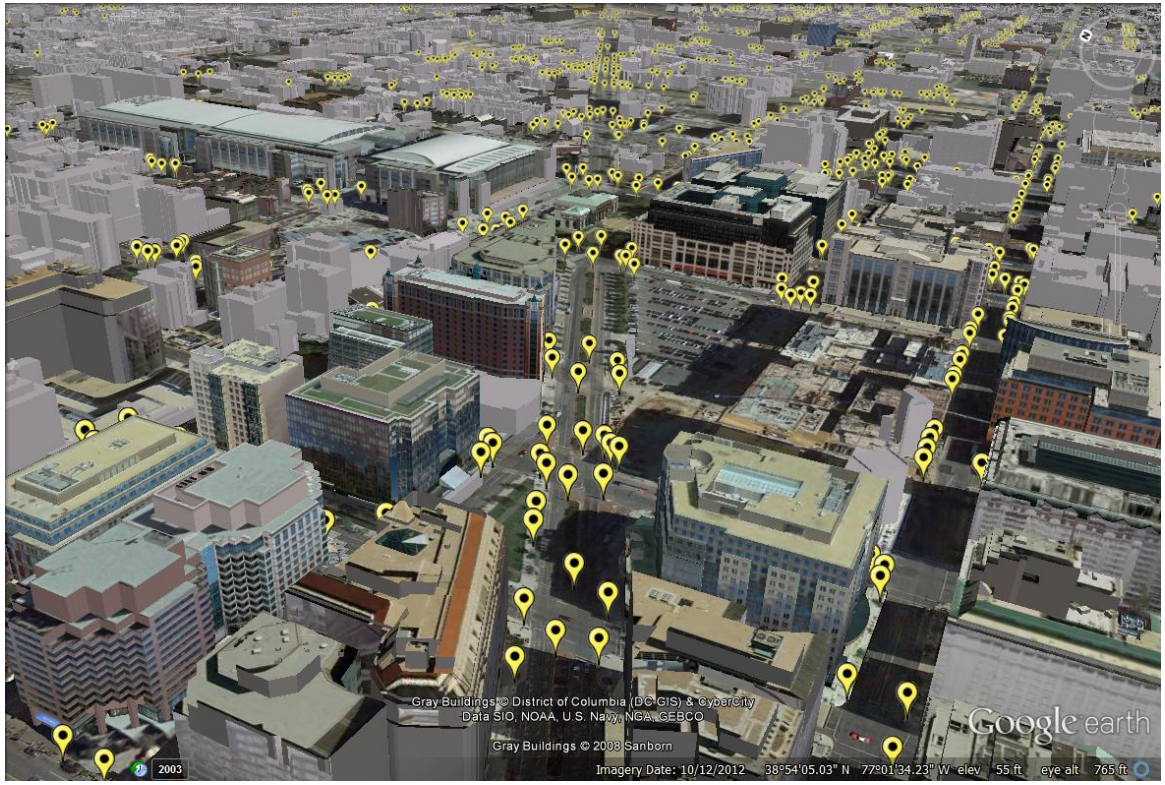
GIS Dataset

- e.g. Washington D.C.
  - Lamp posts
  - Fire Hydrants



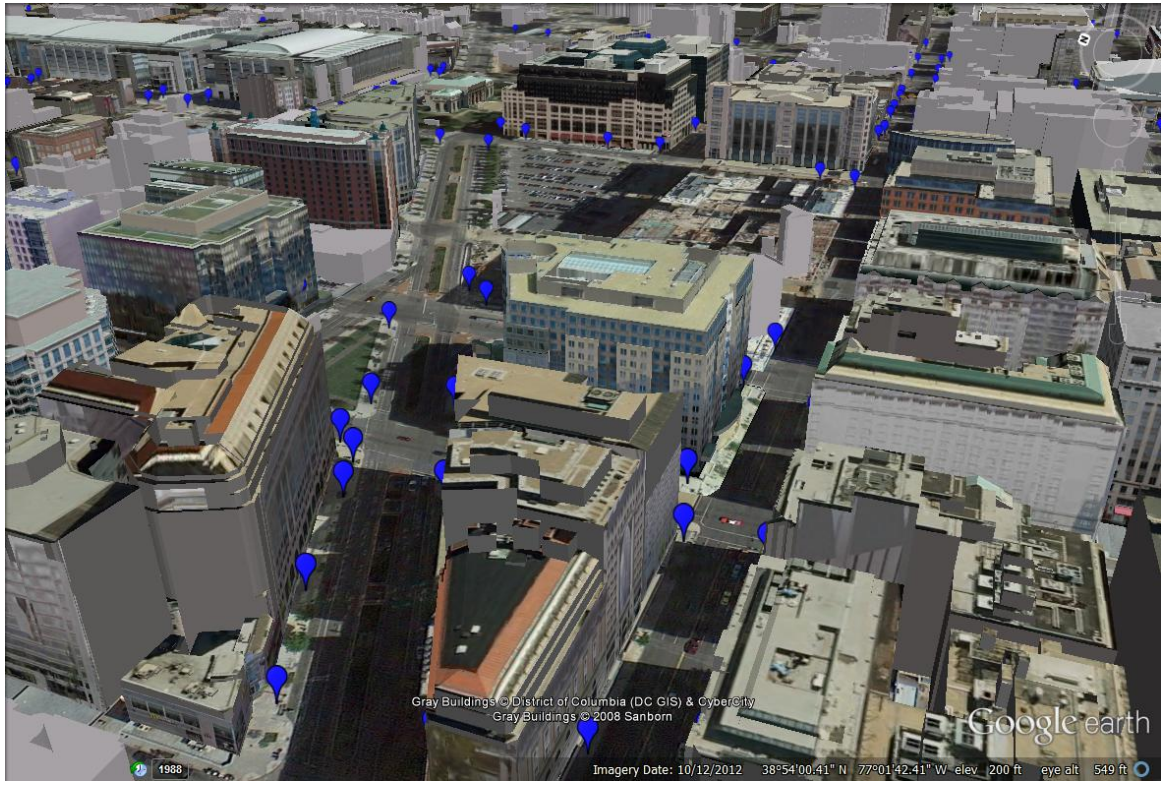
GIS Dataset

- e.g. Washington D.C.
  - Lamp posts
  - Fire Hydrants
  - Road signs



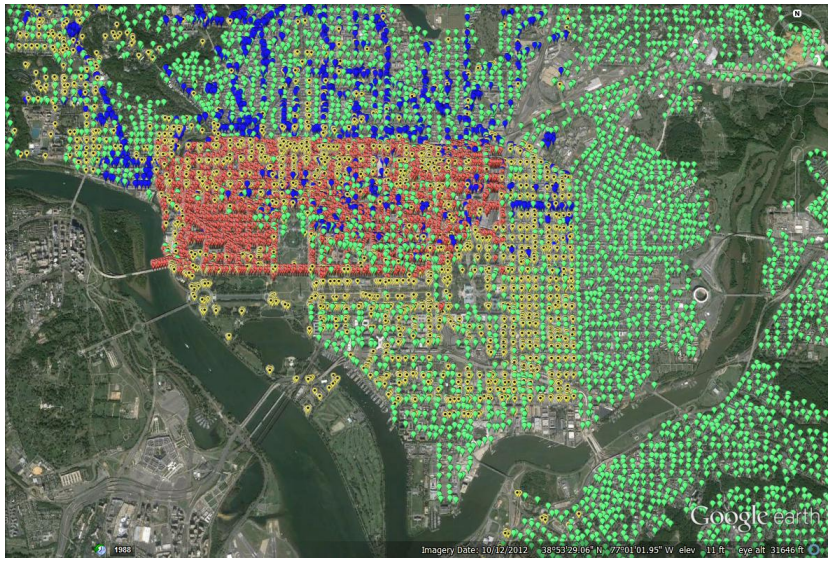
GIS Dataset

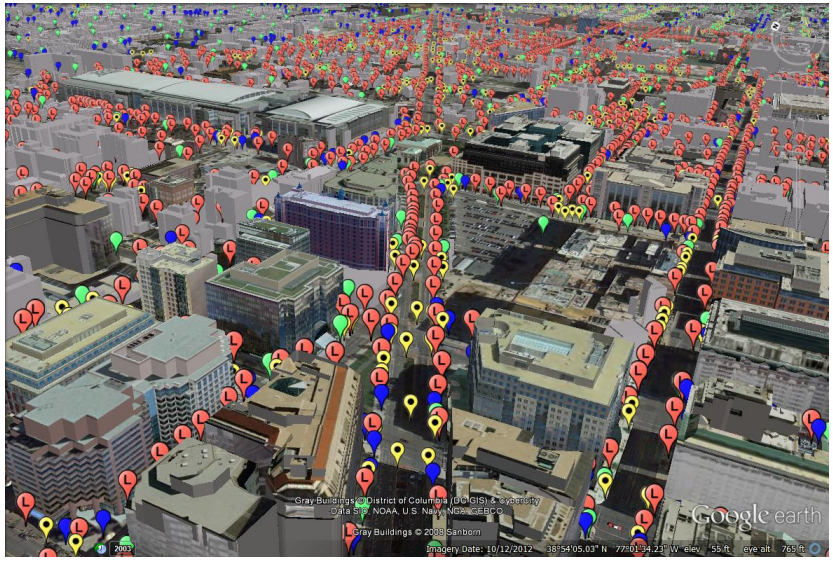
- e.g. Washington D.C.
  - Lamp posts
  - Fire Hydrants
  - Road signs
  - Trash cans



GIS Dataset

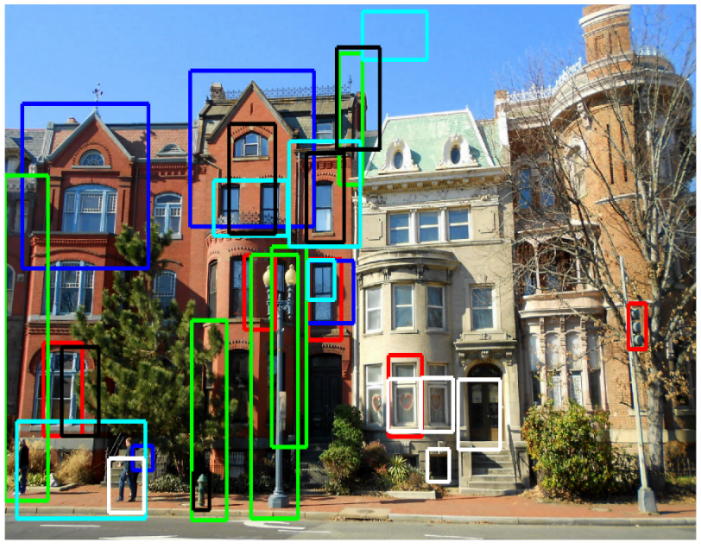
- Locations of most stationary objects are documented!
- e.g. Washington D.C.
  - Buildings, Foliage, Road signs, ATMs, Fire Hydrants, Lamp posts, Cell/ FM towers, Traffic Lights, Bus/subway stations, Trash cans.



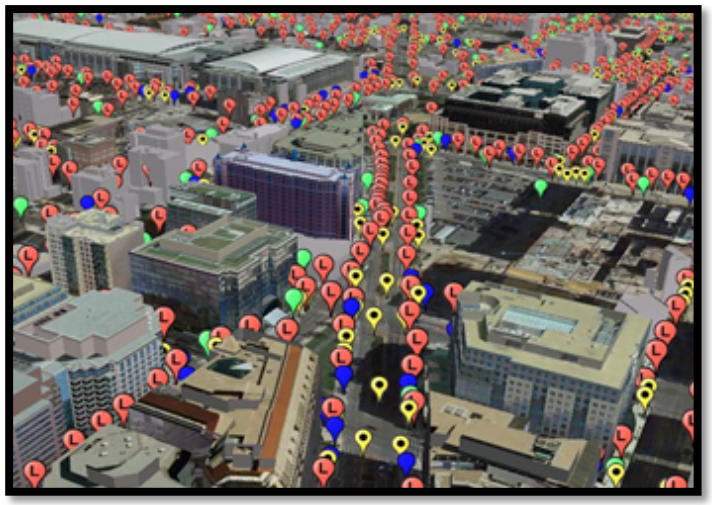


Fusion of Image content and GIS

Object Detectors

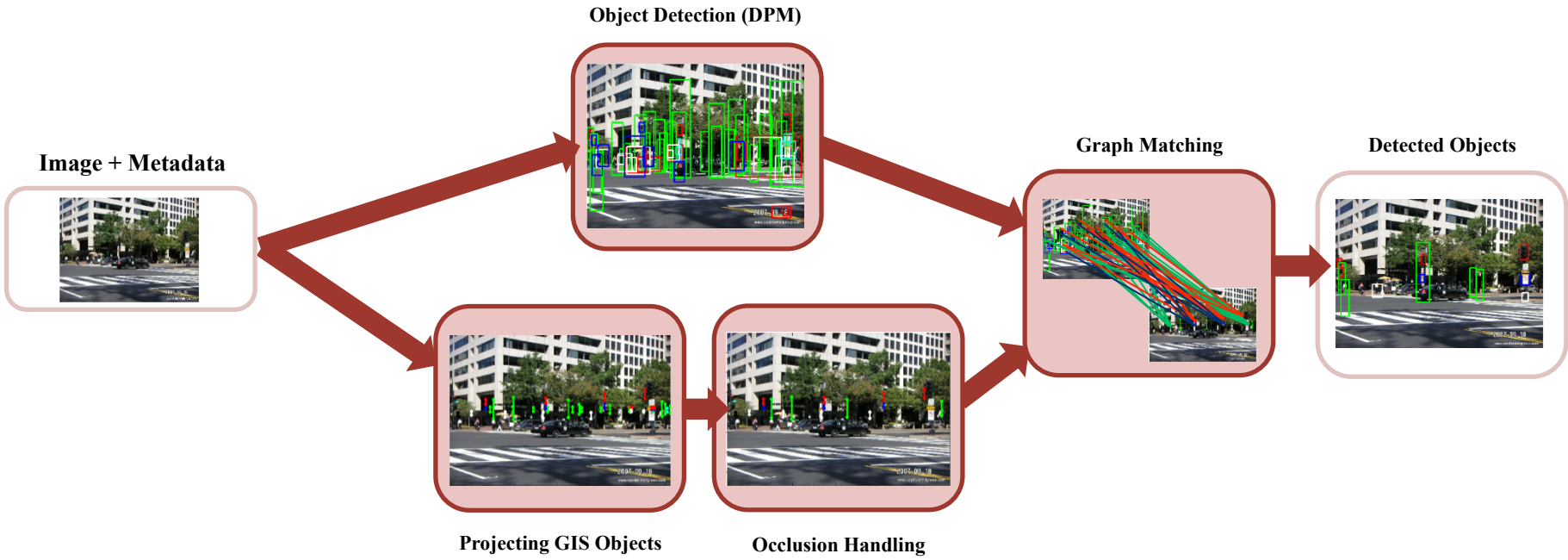


GIS



Lamp Post Road Sign Traffic Light

Location-aware Object Detection



Obtaining Priors from GIS

Camera View

All Projections

Non-occluded Projections

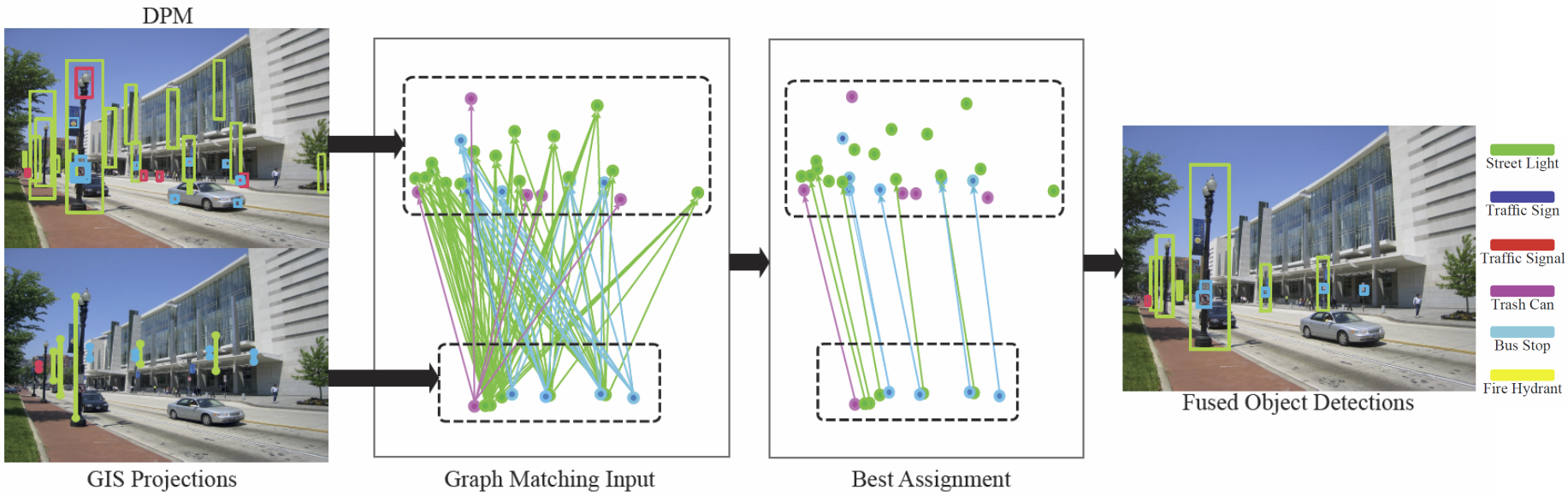
2D projection on the Image

$$\begin{pmatrix} \Phi(i) \\ 1 \end{pmatrix} = \boldsymbol{P}\phi(i)$$

Camera Matrix

GIS priors

Higher Order Graph Matching



Query Image

DPM Results

Loose Threshold

Tuned Threshold

Strict Threshold

Street Light Traffic Sign Traffic Signal Trash Can Bus Stop Fire Hydrant

GIS Projections

DPM Results

Street Light Traffic Sign Traffic Signal Trash Can Bus Stop Fire Hydrant

Non Occluded GIS Projections

DPM Results

Street Light Traffic Sign Traffic Signal Trash Can Bus Stop Fire Hydrant

Non Occluded GIS Projections

Our Results

Street Light Traffic Sign Traffic Signal Trash Can Bus Stop Fire Hydrant

Traffic Signal, **Street Light**, and **Fire Hydrant** are detected successfully.

Non Occluded GIS Projections

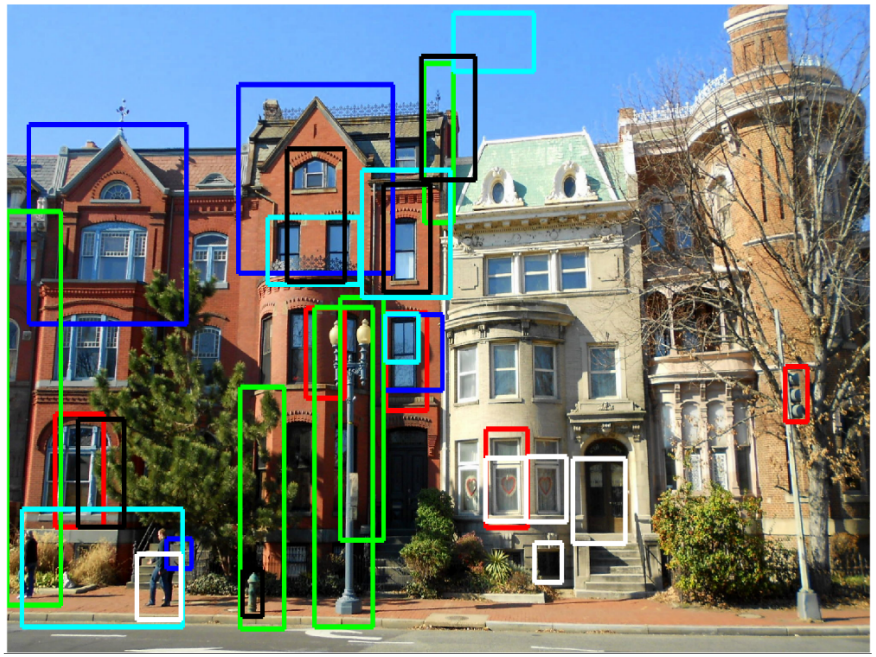
Our Results

Street Light Traffic Sign Traffic Signal Trash Can Bus Stop Fire Hydrant

Object Detection + 3D

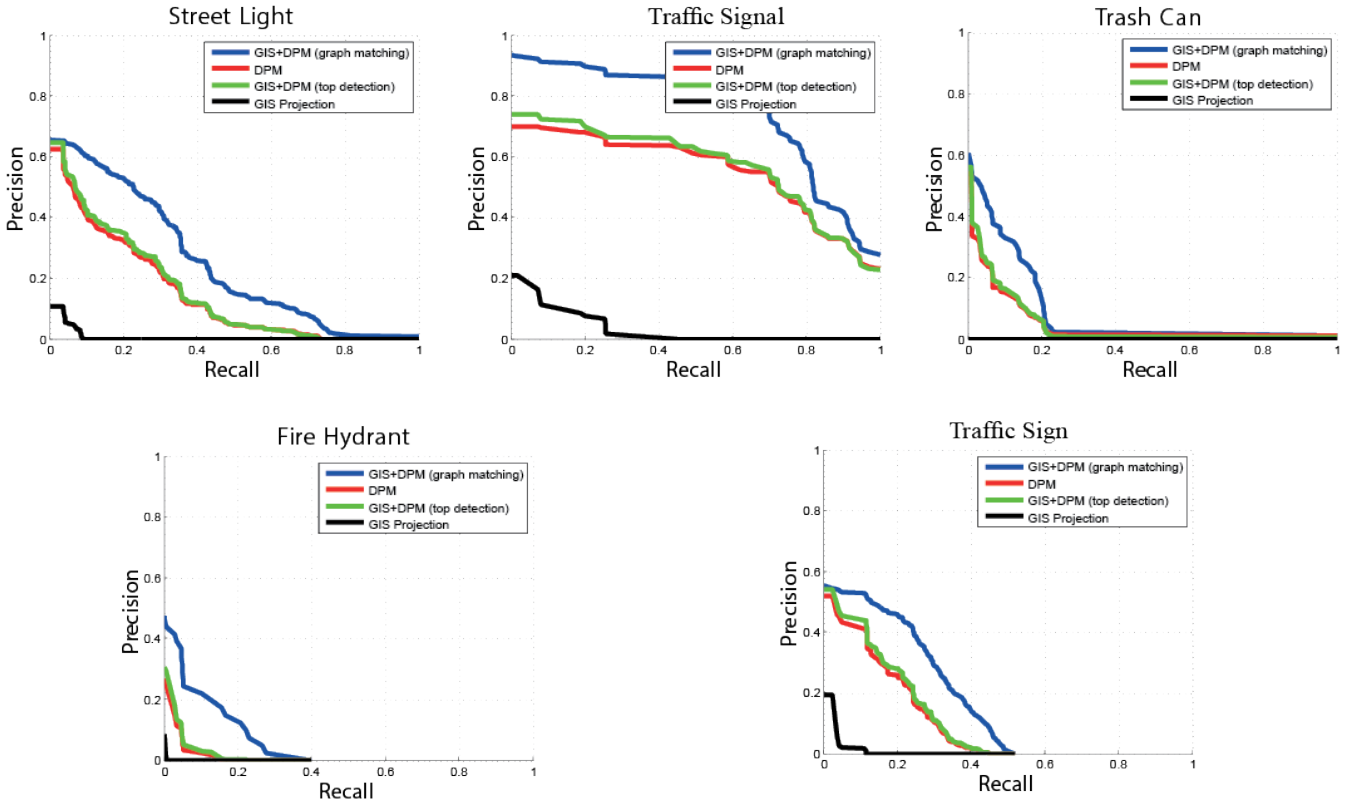
DPM Results (Tuned Threshold)

Our Results



Street Light Traffic Sign Traffic Signal Trash Can Bus Stop Fire Hydrant

Quantitative Object Detection Results

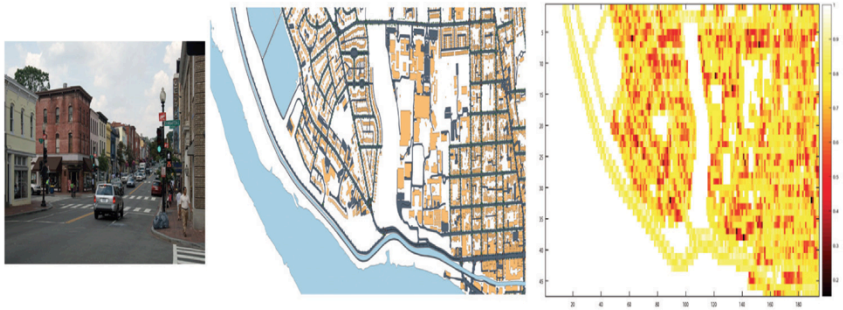


Do we need camera meta-data?

- How far can we get without meta data (i.e., GPS location and compass direction)?

Semantic Cross-View Matching

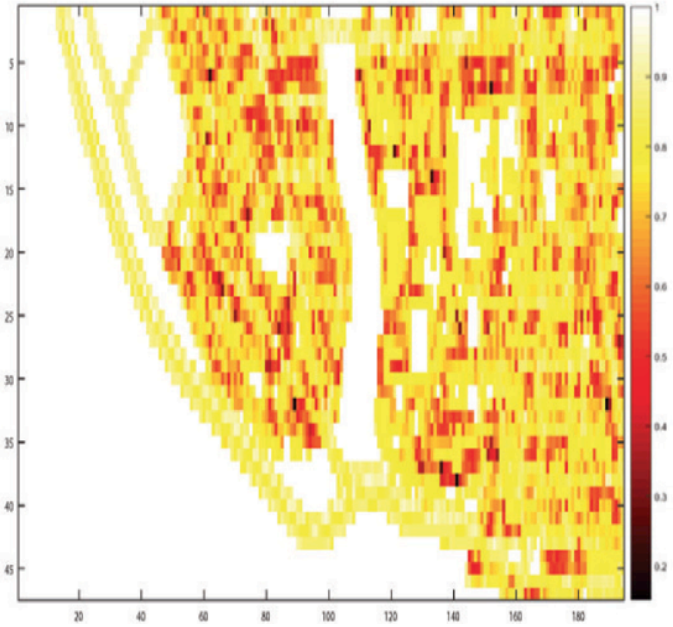
F. Castalo, A. Zamir, R. Angst, F. Palmieri, S. Savarese
In ICCVW 2015



Semantic Cross-View Matching

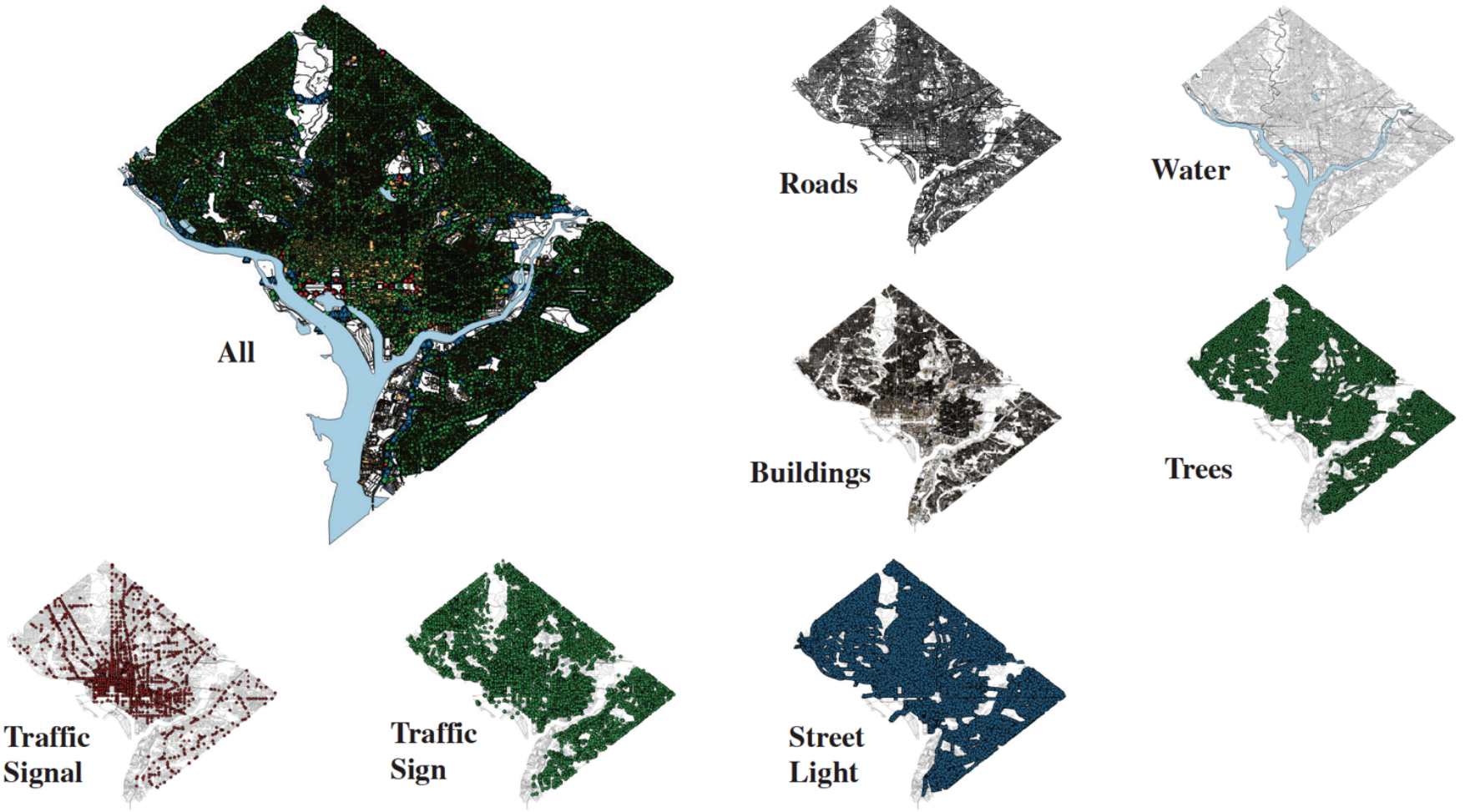
Input

Semantic Priors (map)

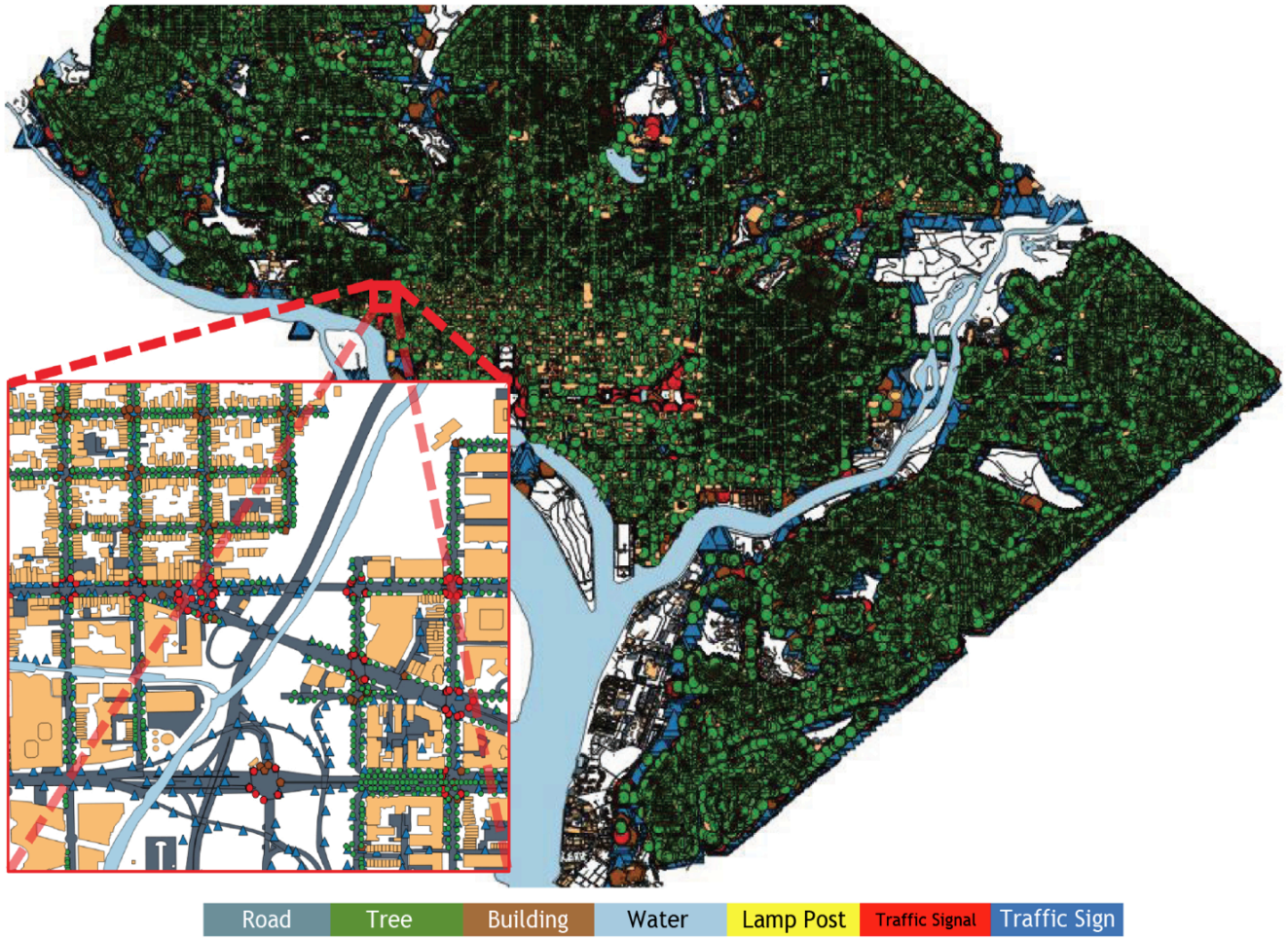


Alignment Heat Map

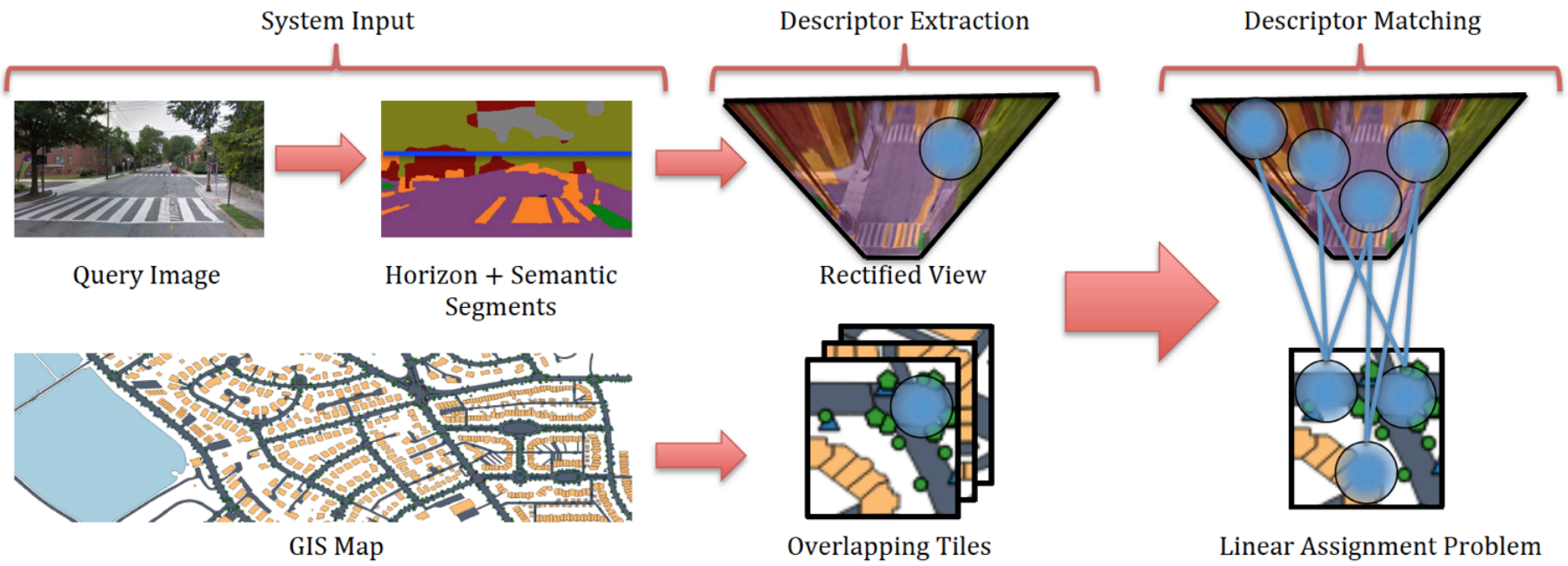
Semantic Map (GIS)



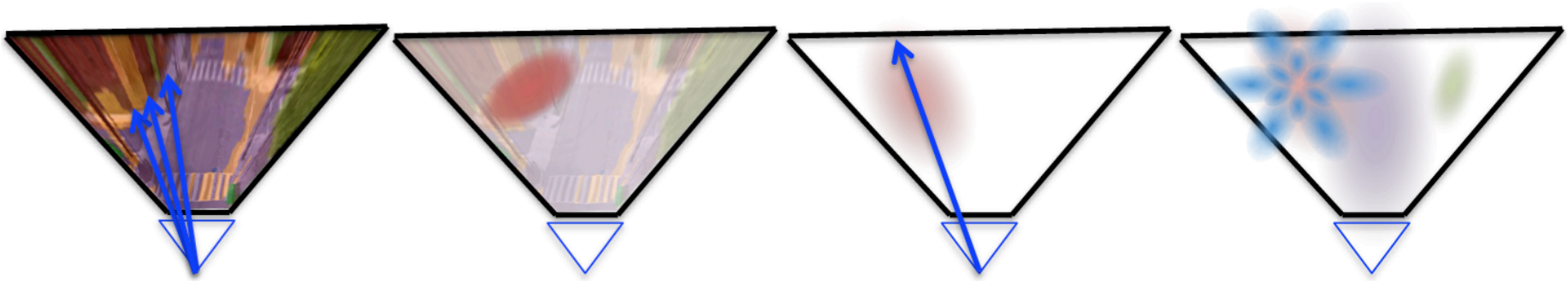
Semantic Map (GIS)



Topological and Semantic Matching



Semantic Segment Layout (SSL) features



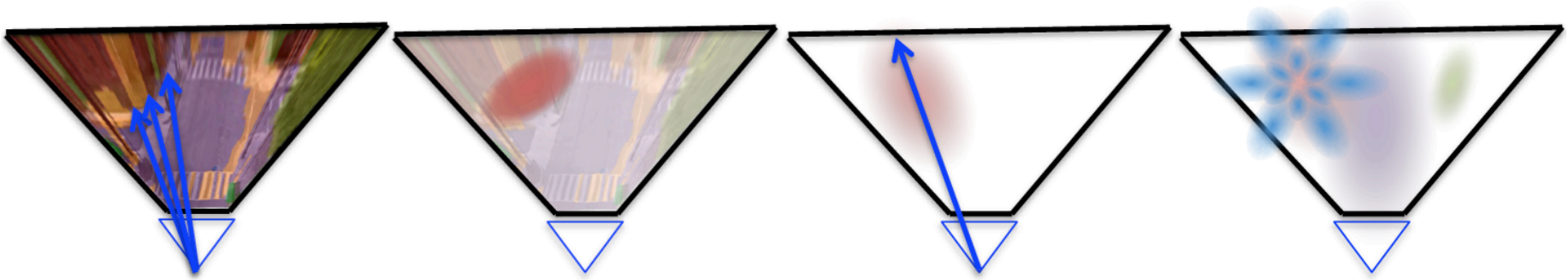
(b) Contact Region Detection

(c) Segment Shape Approximation

(d) Contact Region Uncertainty

(e) Descriptor Extraction

Semantic Segment Layout (SSL) features

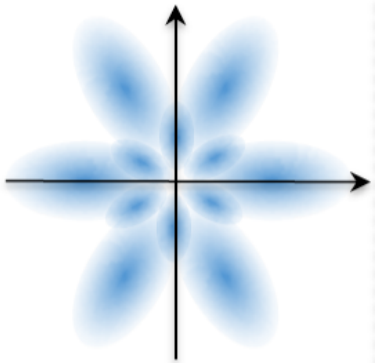


(b) Contact Region Detection

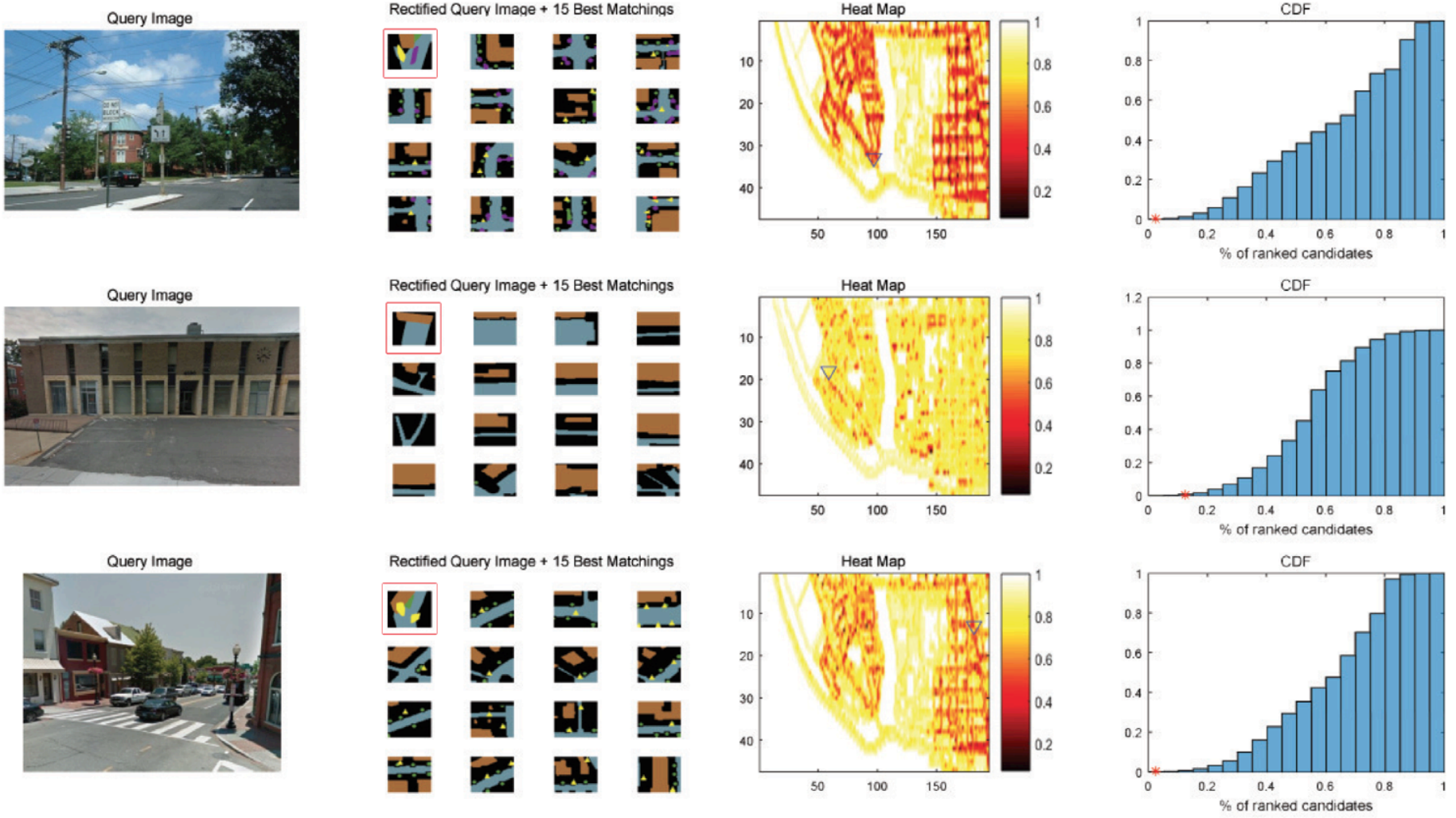
(c) Segment Shape Approximation

(d) Contact Region Uncertainty

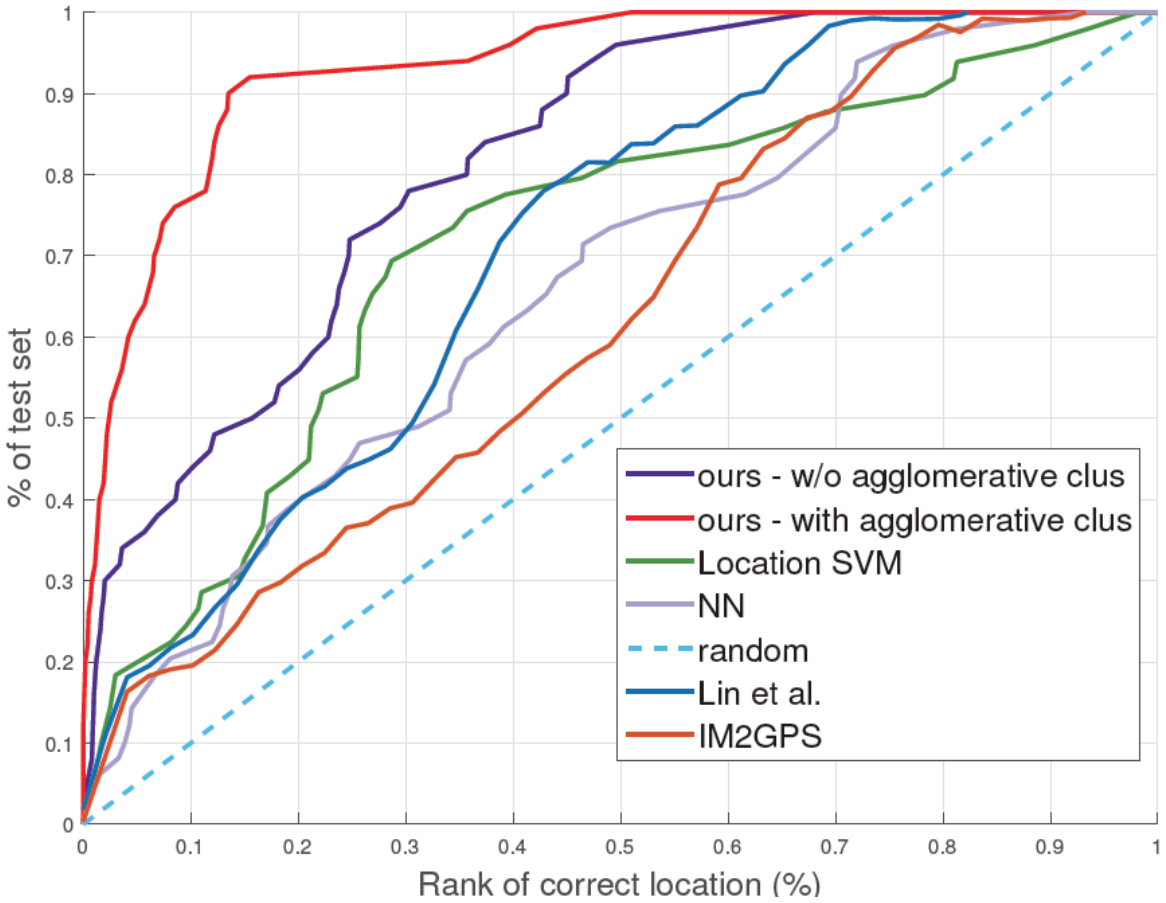
(e) Descriptor Extraction



(a) Gaussian Pooling Regions


Query Image
Rectified Query Image + 15 Best Matchings
Heat Map
CDF
% of ranked candidates
Query Image
Rectified Query Image + 15 Best Matchings
Heat Map
CDF
% of ranked candidates
Query Image
Rectified Query Image + 15 Best Matchings
Heat Map
CDF
% of ranked candidates

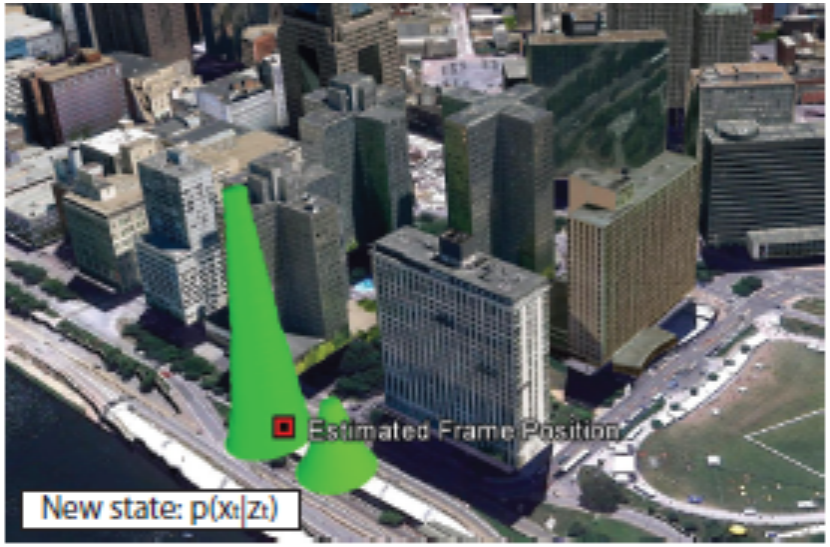
Experimental Results



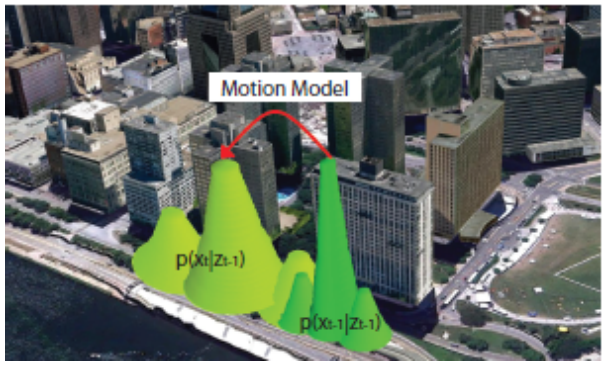
How about Videos?!

City Scale Geo-spatial Trajectory Estimation of a Moving Camera. Vaca, Zamir, Shah. In ***CVPR***, 2012.

Bayesian Recursive Estimation



Likelihood
(Current Segment)



Prediction
(previous Segment)

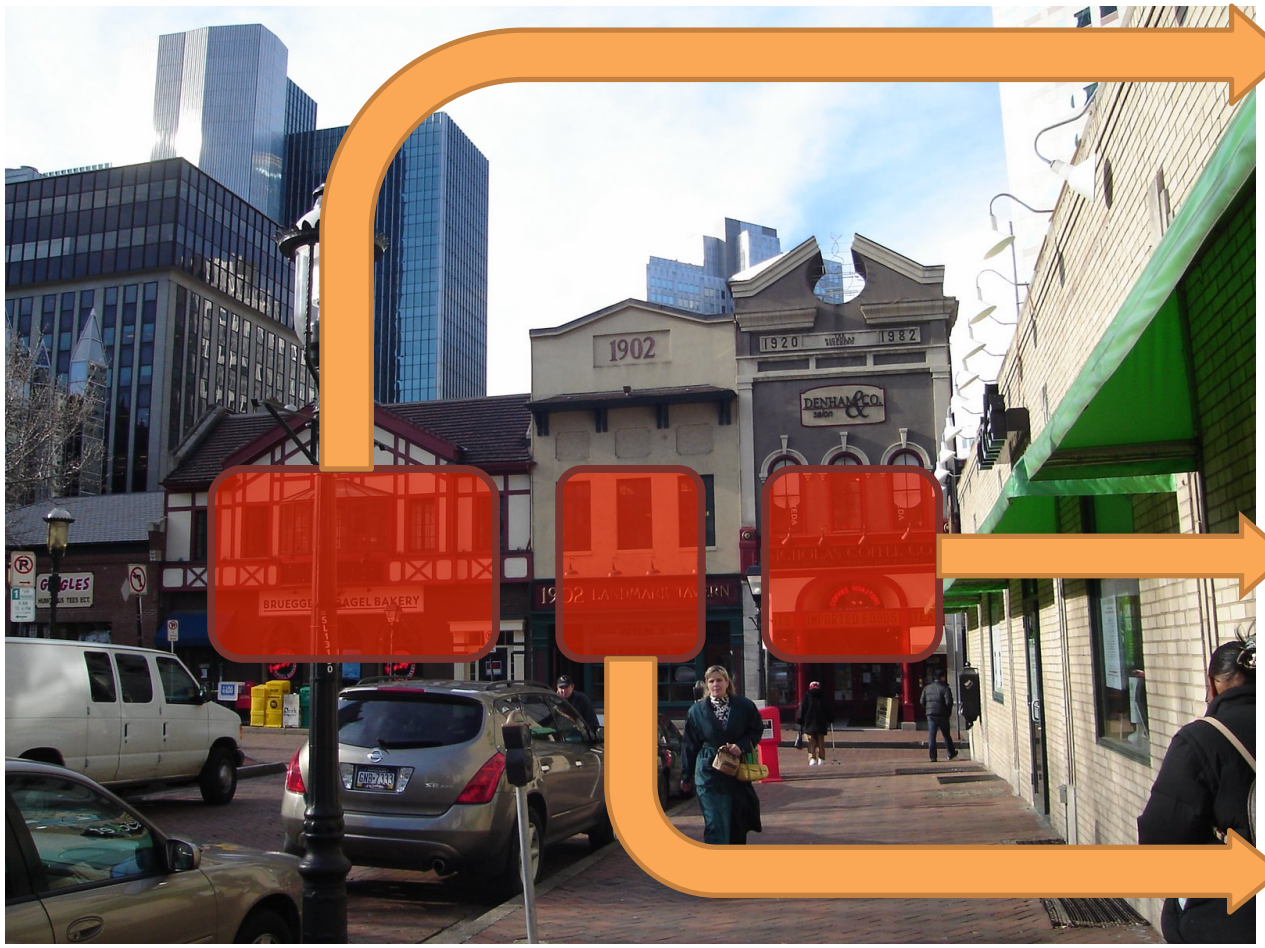
Camera Pose Estimation for a YouTube Video

City Scale Geo-spatial Trajectory Estimation of a Moving Camera. Vaca, Zamir, Shah. In ***CVPR***, 2012.

Simultaneous 3D + Semantics

- **ACM Multimedia'13:** Visual Business Recognition - A Multimodal Approach

3D + Semantics



Bruegger's Bagel

25 Market Square
Pittsburgh, PA 15222

User Rating: 5/5

Nicholas Coffee Co.

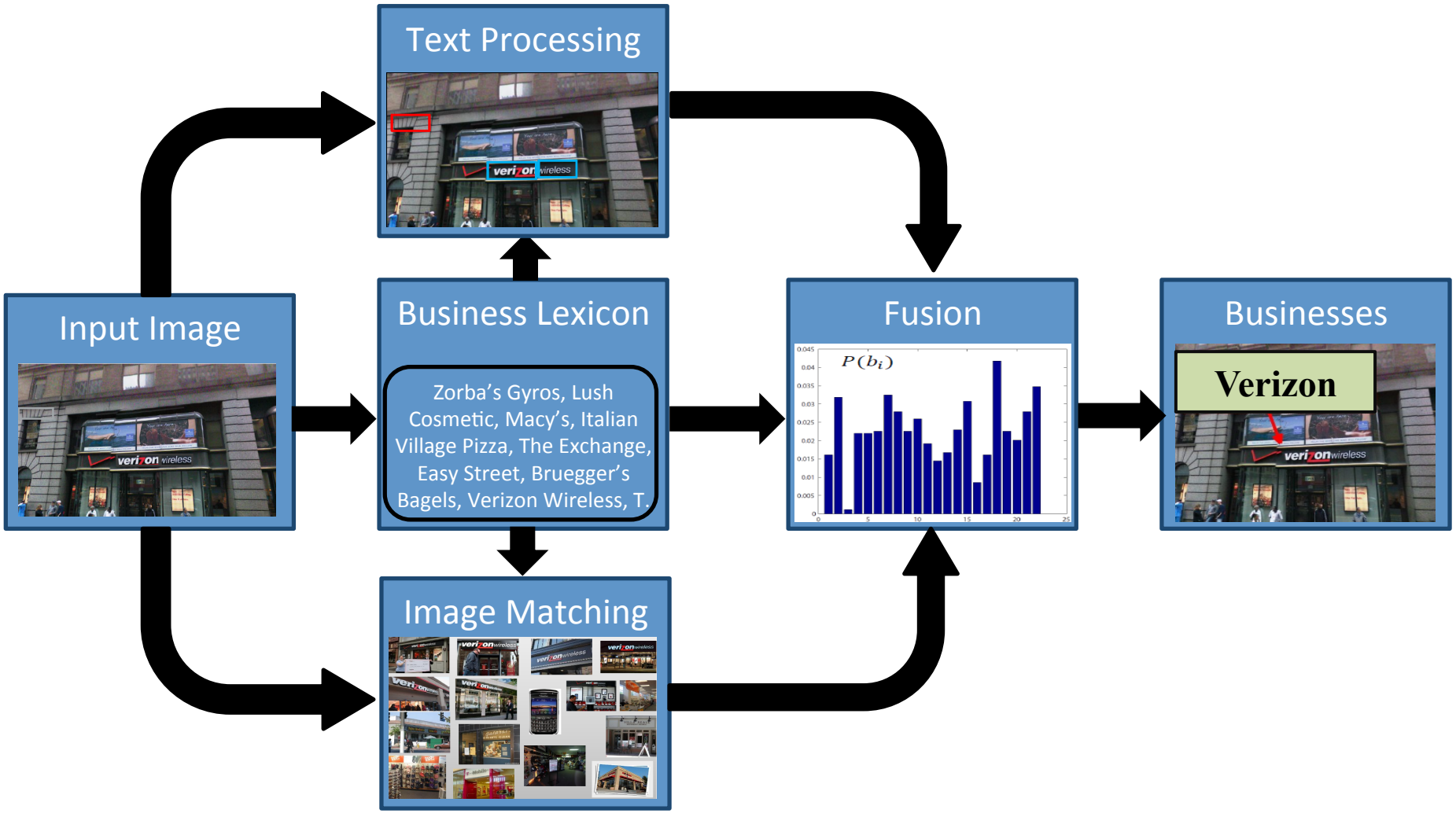
23 Market Square
Pittsburgh, PA 15222

User Rating: 4/5

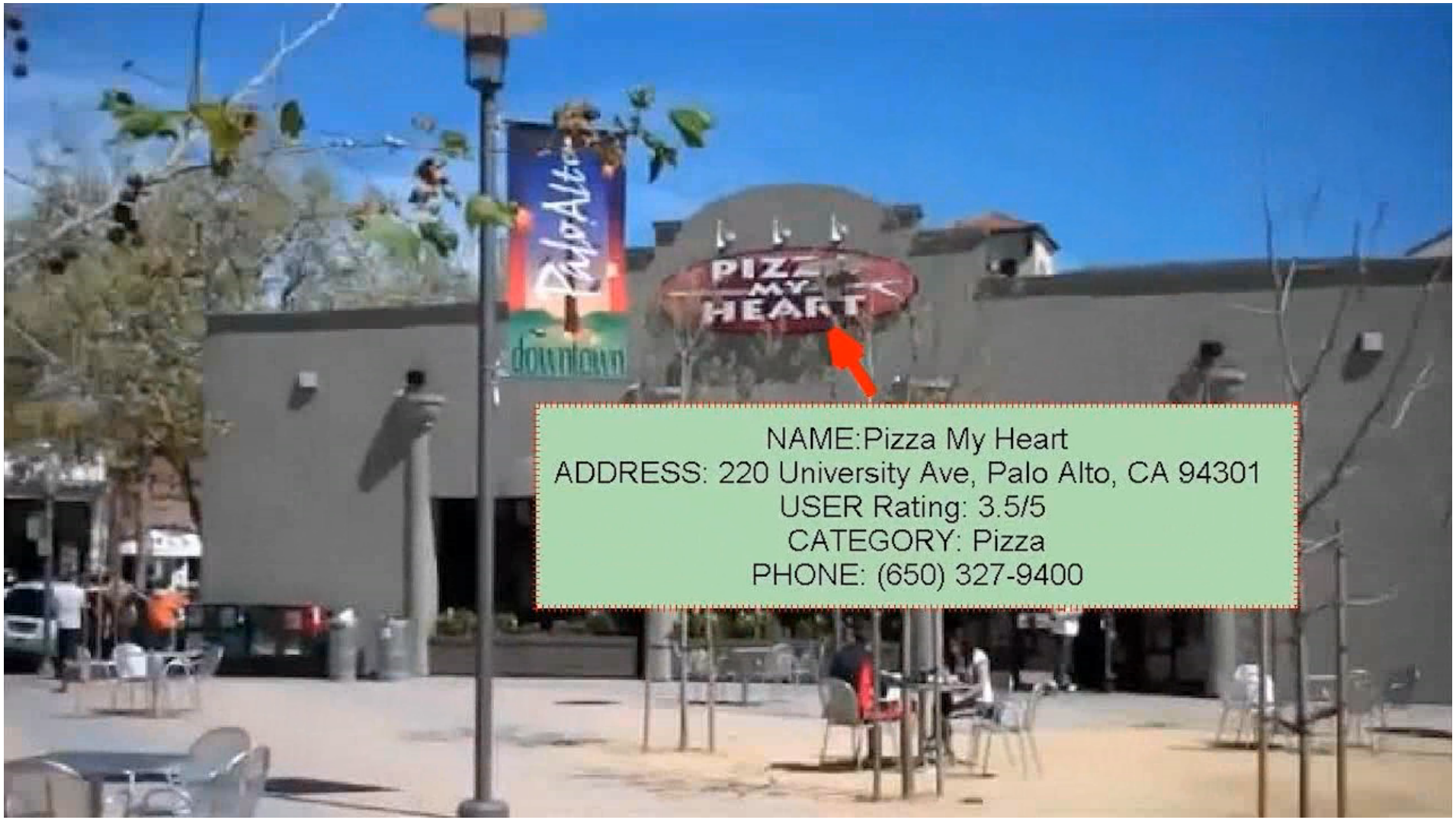
Tavern,

24 Market Square
Pittsburgh, PA 15222

User Rating: 2/5


Text Processing
Input Image
Business Lexicon
Zorba's Gyros, Lush Cosmetic, Macy's, Italian Village Pizza, The Exchange, Easy Street, Bruegger's Bagels, Verizon Wireless, T.
Fusion
$P(b_i)$
Businesses
Verizon
Image Matching

Results


Palo Alto
downtown
PIZZA MY HEART
NAME:Pizza My Heart
ADDRESS: 220 University Ave, Palo Alto, CA 94301
USER Rating: 3.5/5
CATEGORY: Pizza
PHONE: (650) 327-9400

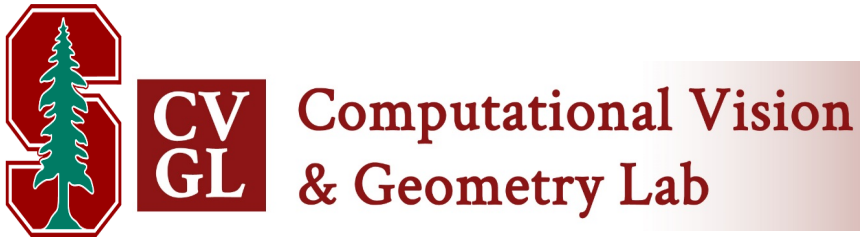
Data Driven 3D Voxel Patterns for Object Category Recognition

Yu Xiang[1,2], Wongun Choi[3], Yuanqing Lin[3], and Silvio Savarese[1]

[1]Stanford University, [2]University of Michigan at Ann Arbor

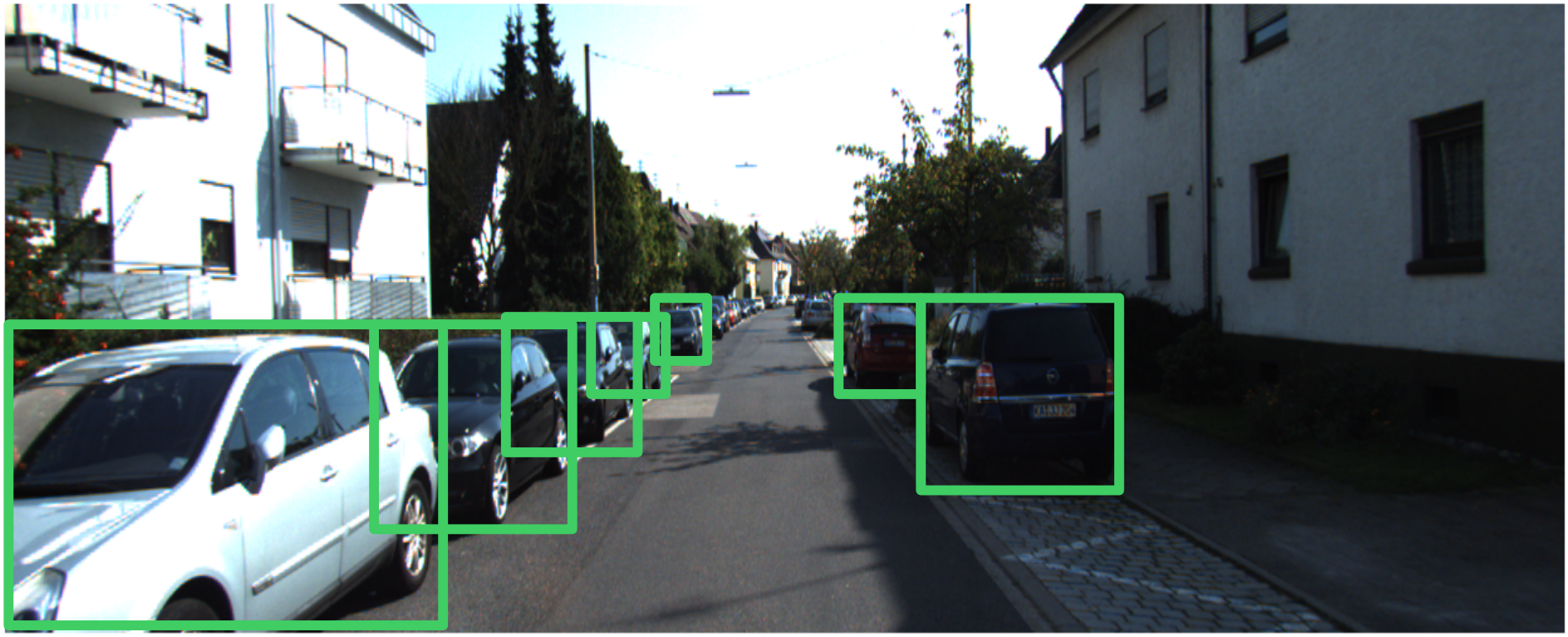
[3]NEC Laboratories America, Inc.

CVPR 2015



The image is from the KITTI detection benchmark (Geiger et al. CVPR'12)

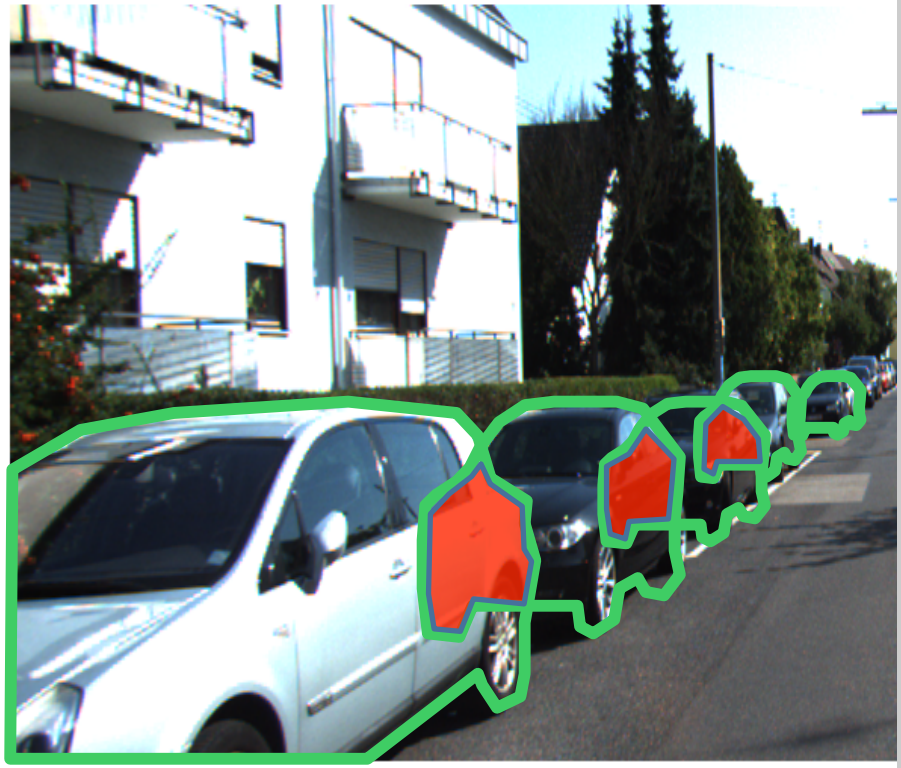
2D Object Detection

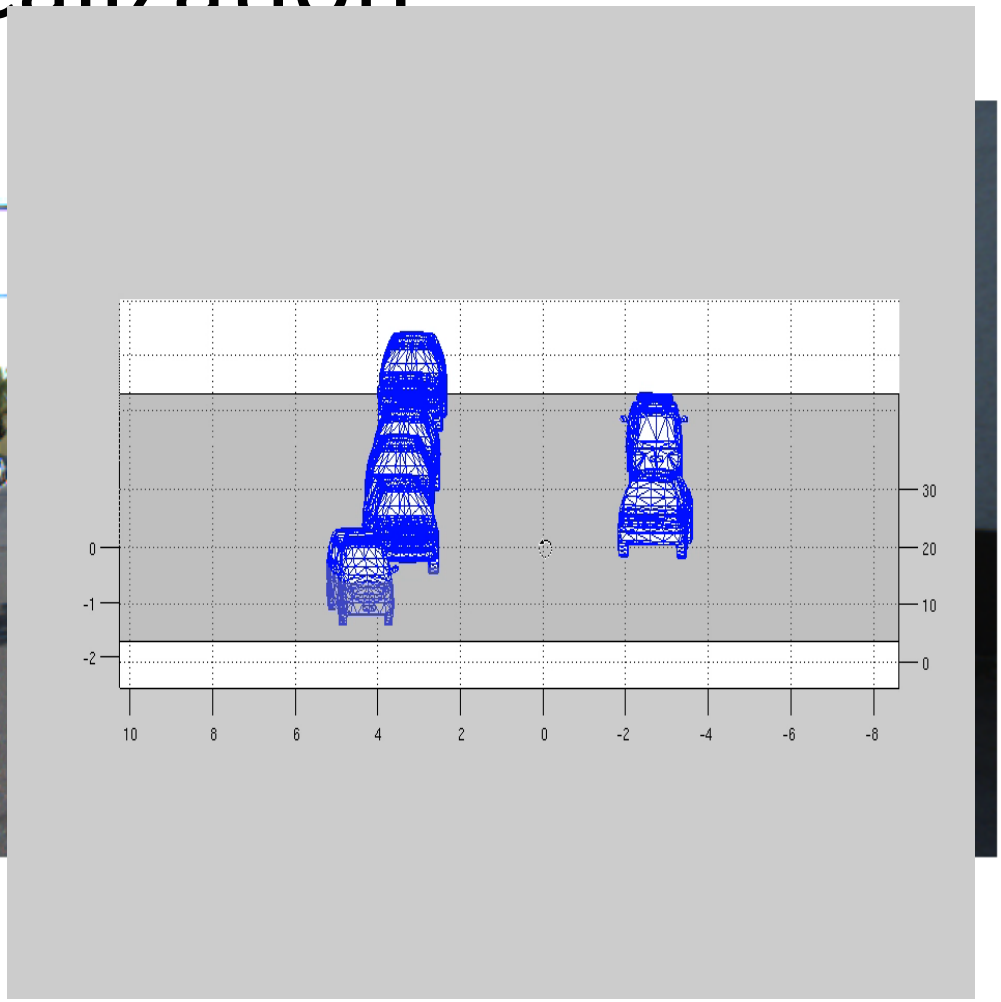


2D Object Segmentation

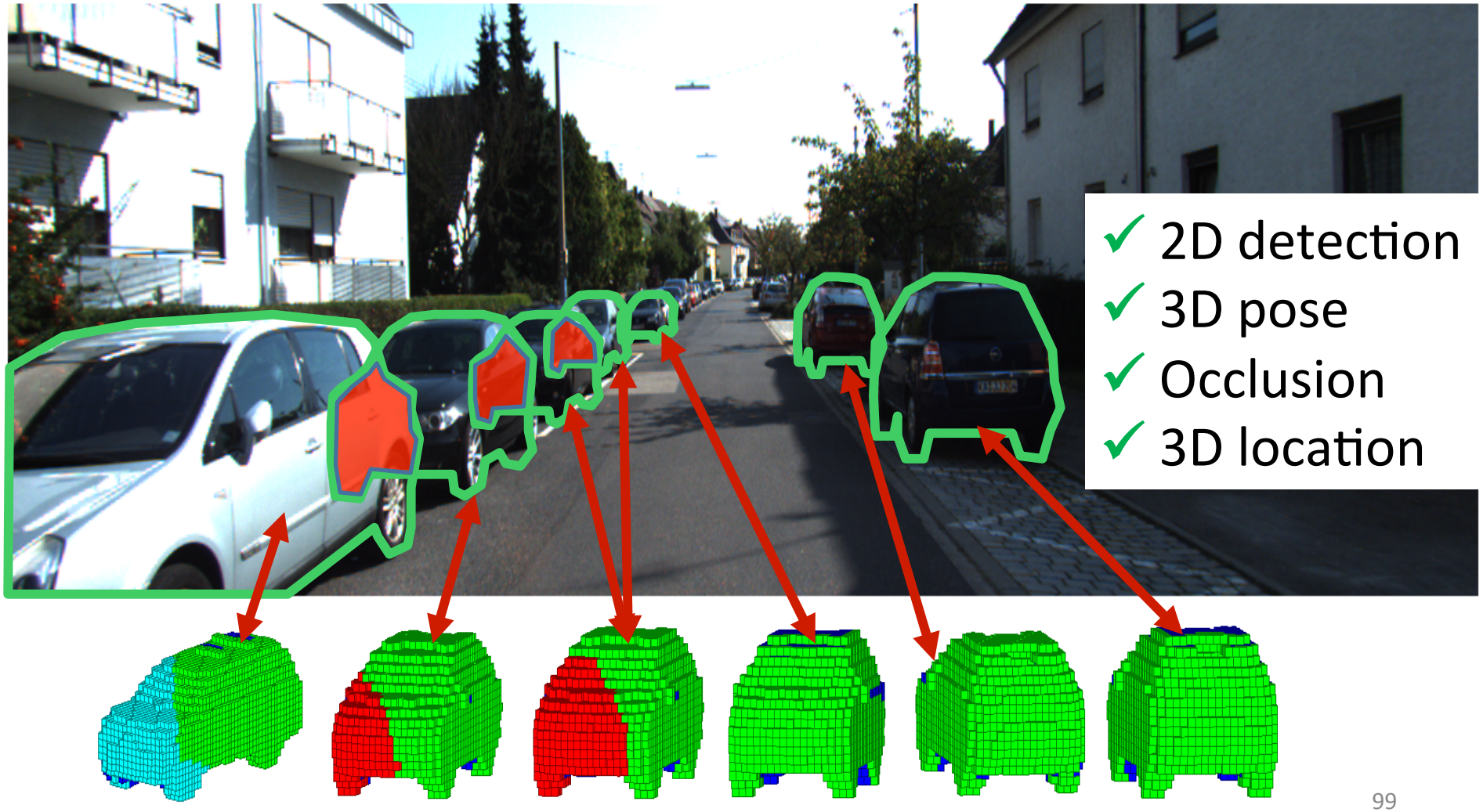
Occlusion Reasoning

3D Localization



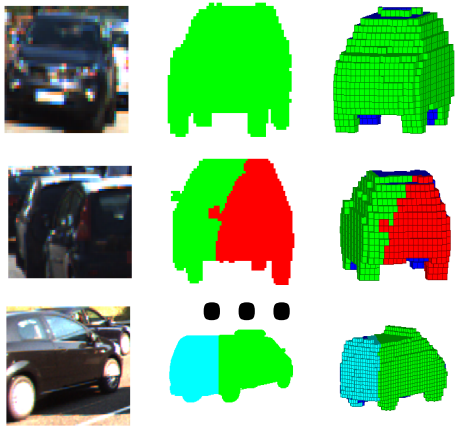


Our Contribution: Data-Driven 3D Voxel Patterns



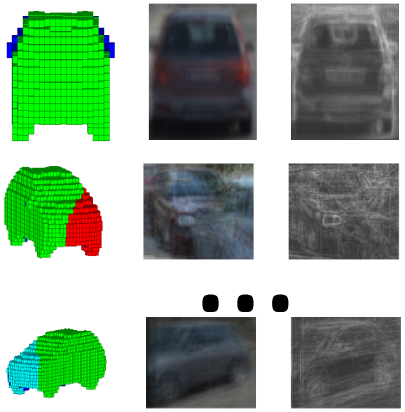
Training Pipeline Overview

1. Align 2D images with 3D CAD models



2. 3D voxel exemplars

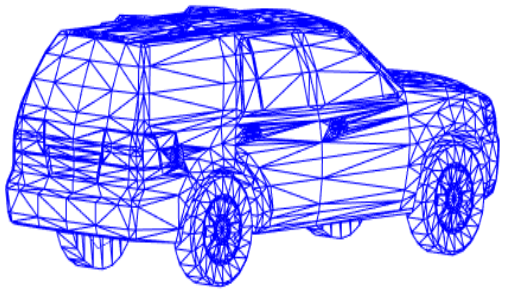
3. 3D voxel patterns



4. Training 3D voxel pattern detectors

Building 3D Voxel Exemplars

Depth ordering



3D CAD model

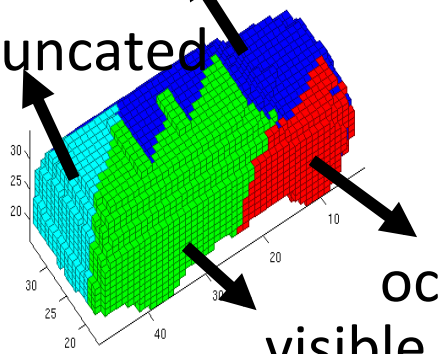
Voxelization

truncated visible occluded

2D mask labeling

self-occluded

truncated

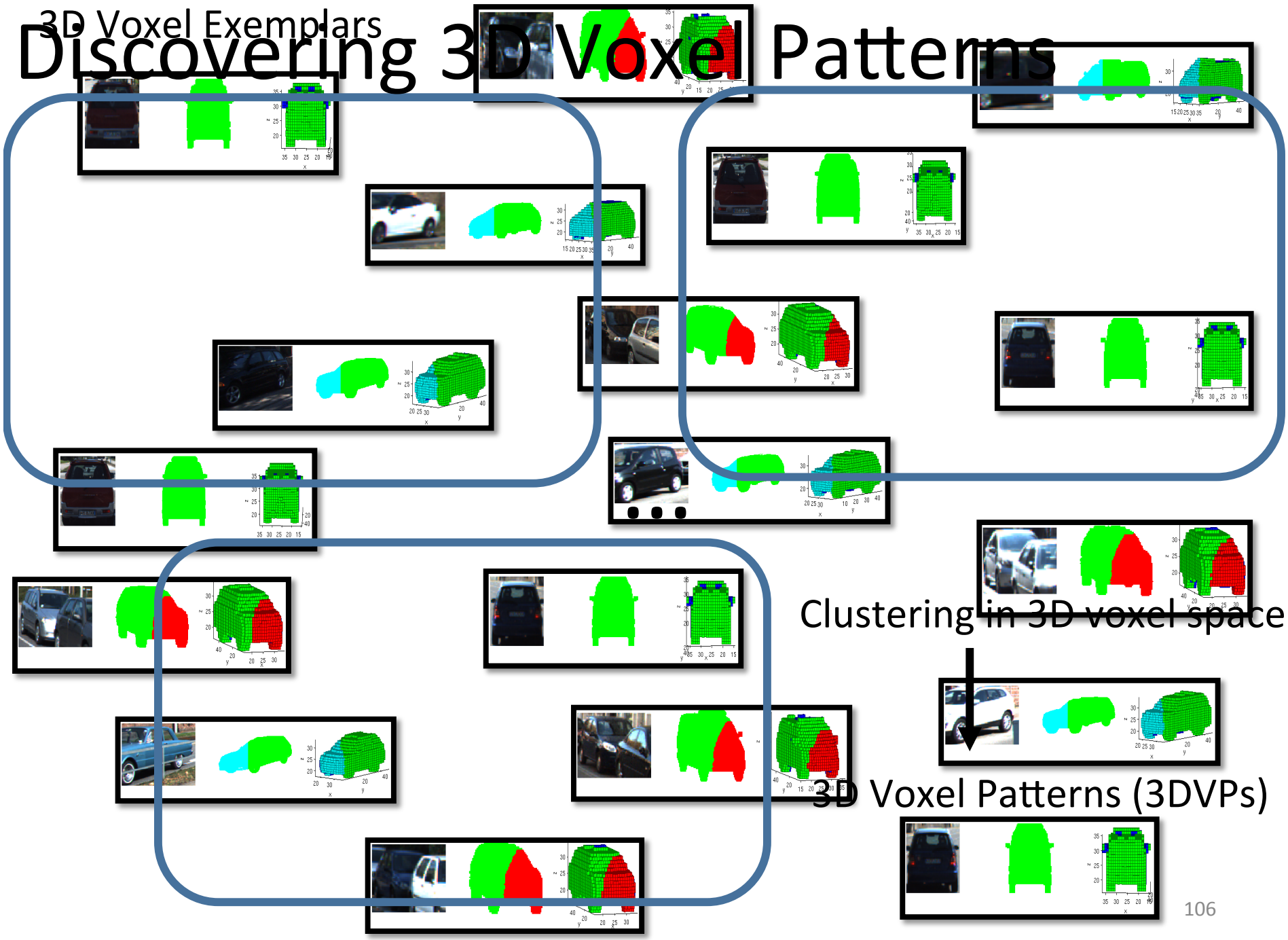


occluded

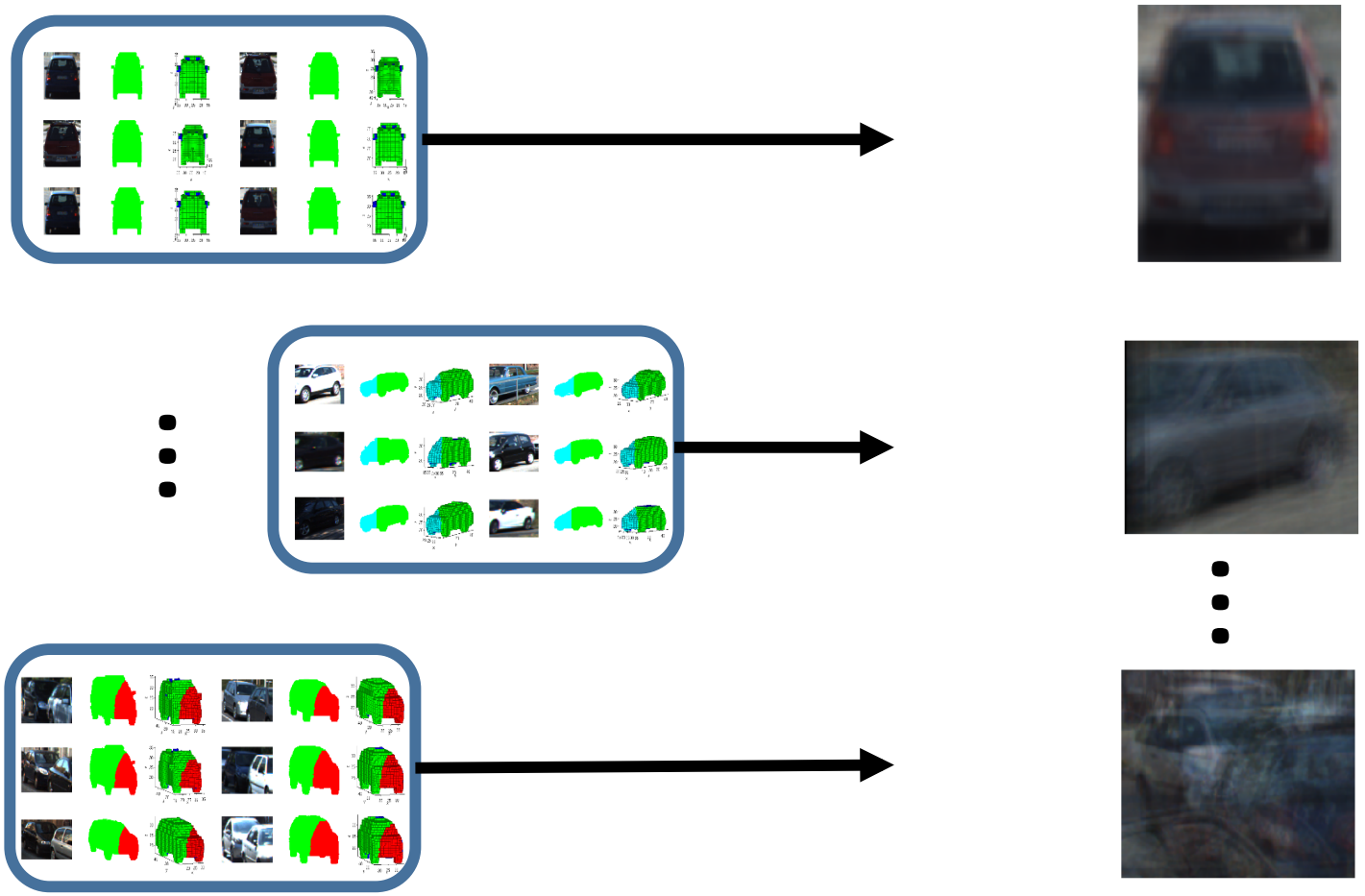
visible

3D voxel model

Discovering 3D Voxel Patterns



Training 3D Voxel Pattern Detectors



- Train a ACF detector for each 3DVP.

P. Dolla´r, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. TPAMI, 2014.

Testing Pipeline Overview

Input 2D image

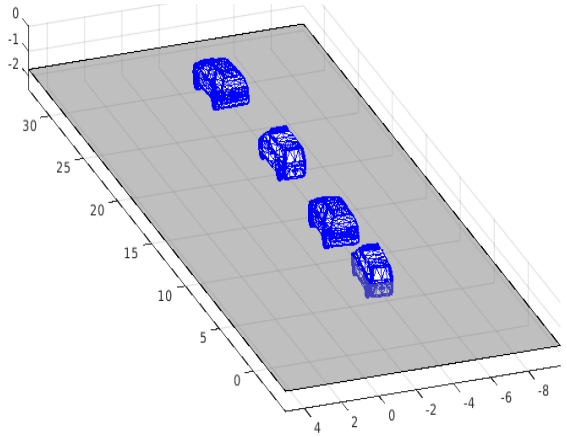
1. Apply 3DVP detectors

2D detection

Transfer meta-data
Occlusion reasoning

2D segmentation

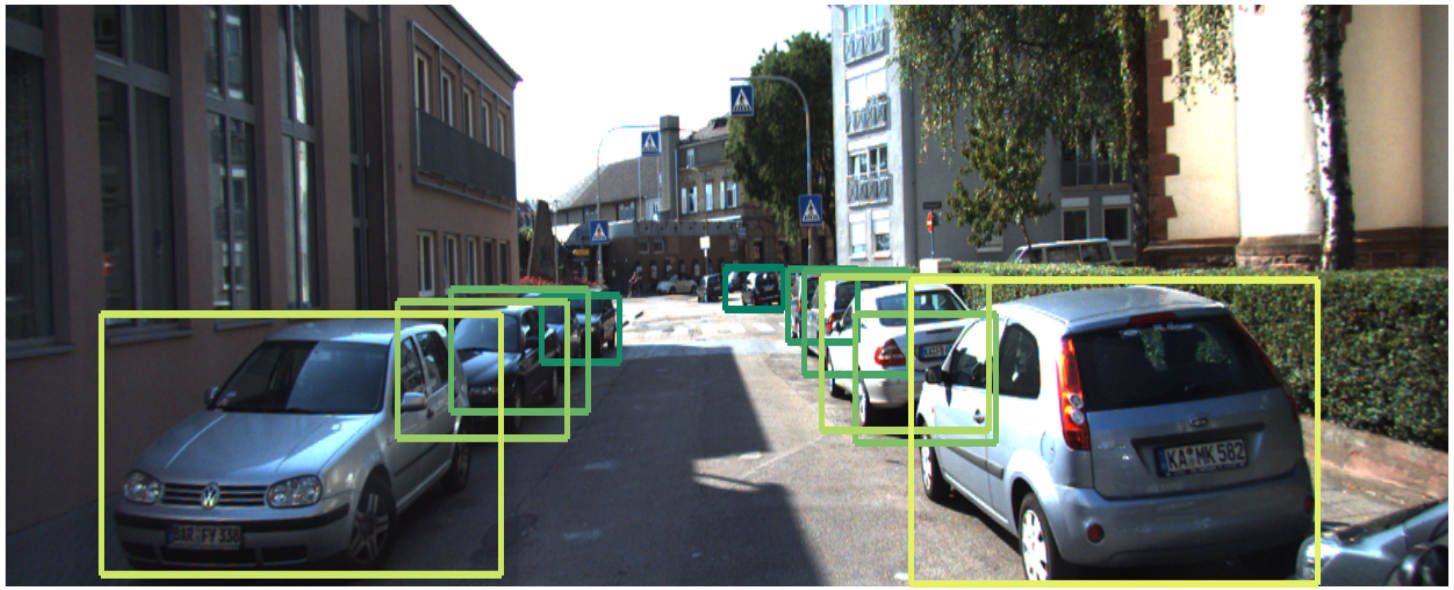
4. Backproject to 3D



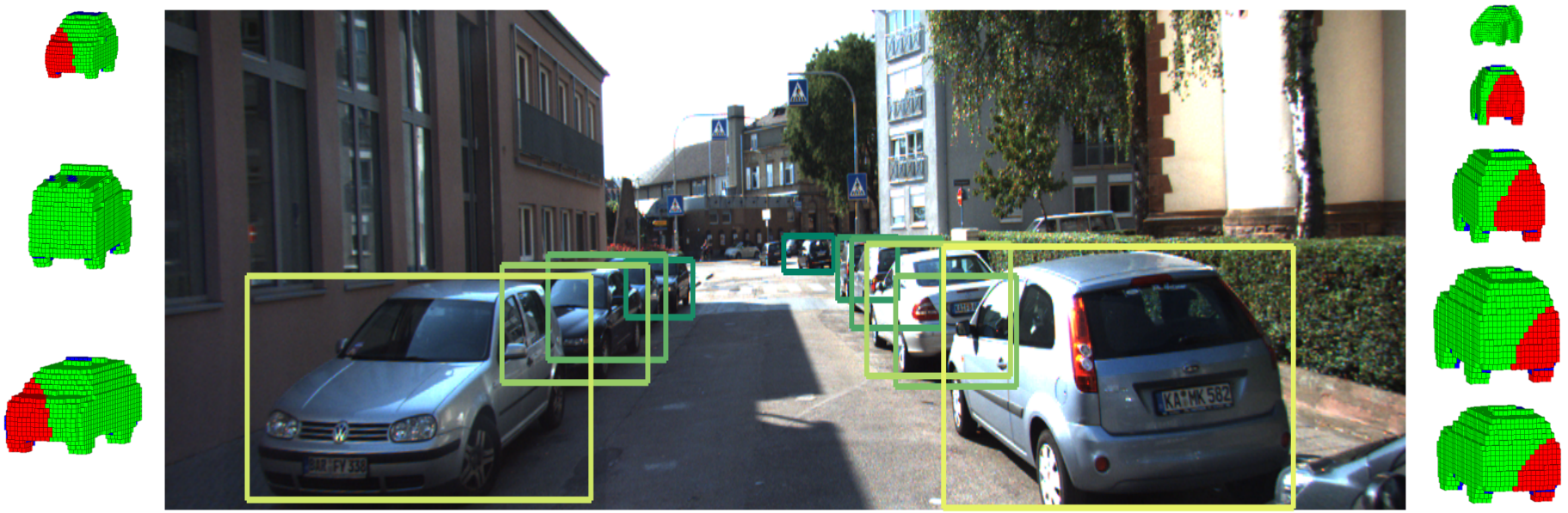
3D localization

1. Apply 3DVP Detectors

1. Apply 3DVP Detectors



2. Transfer Meta-Data



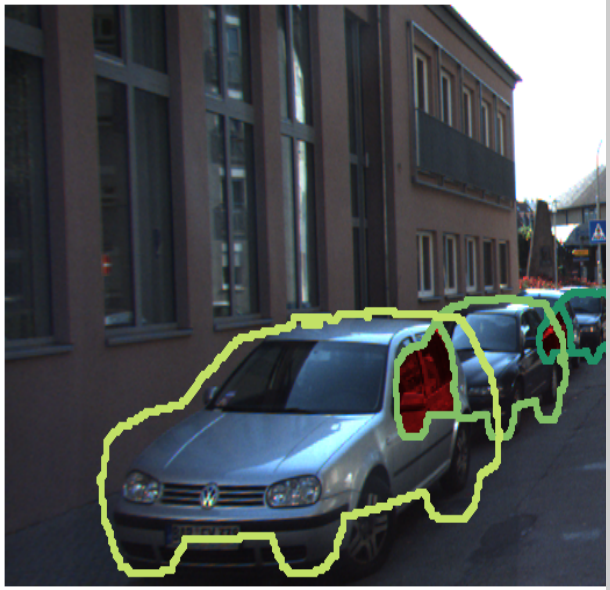
3DVPs

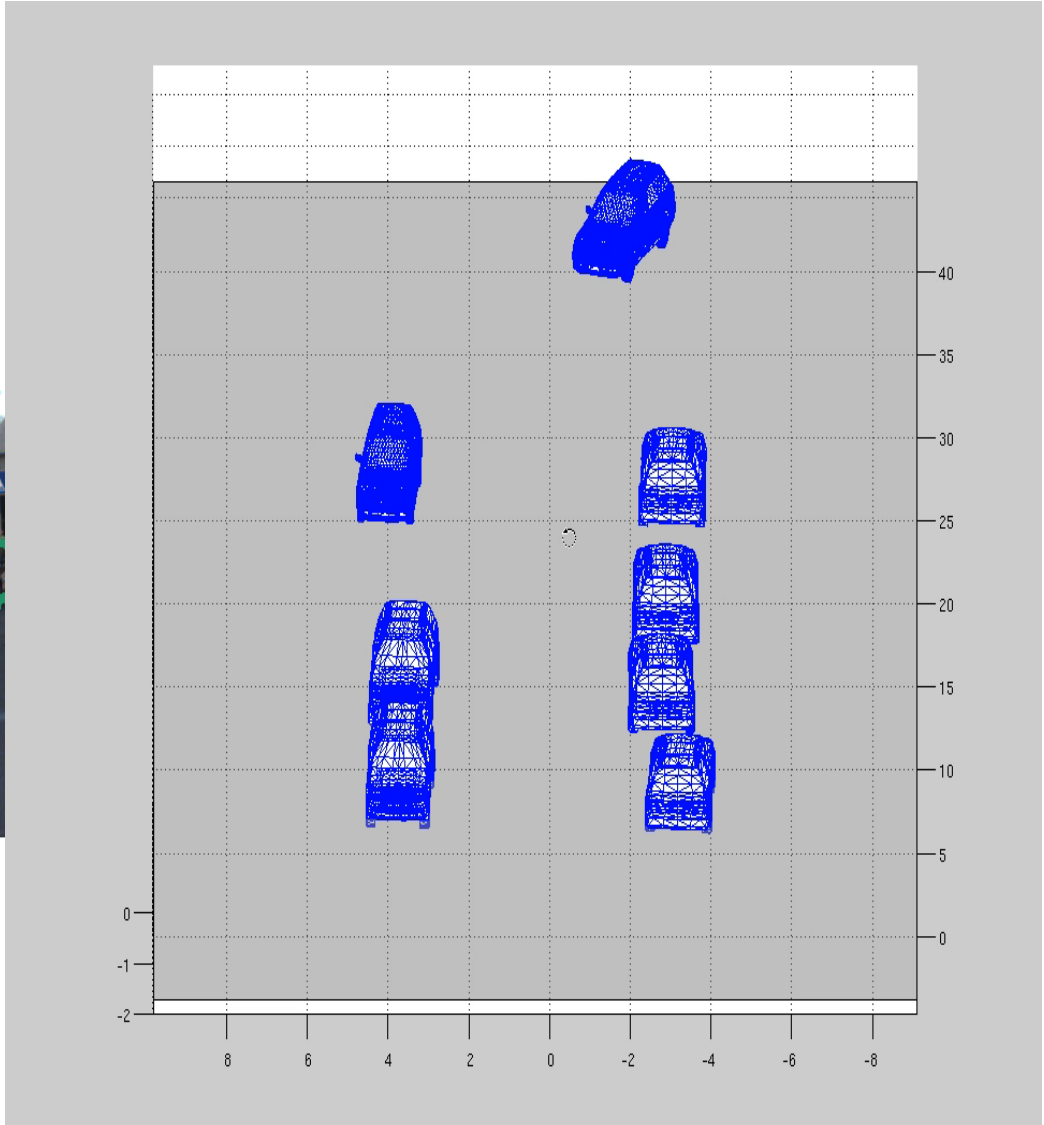
2. Transfer Meta-Data

3. Occlusion Reasoning

Occlusion reasoning: find a set of visibility-compatible detections

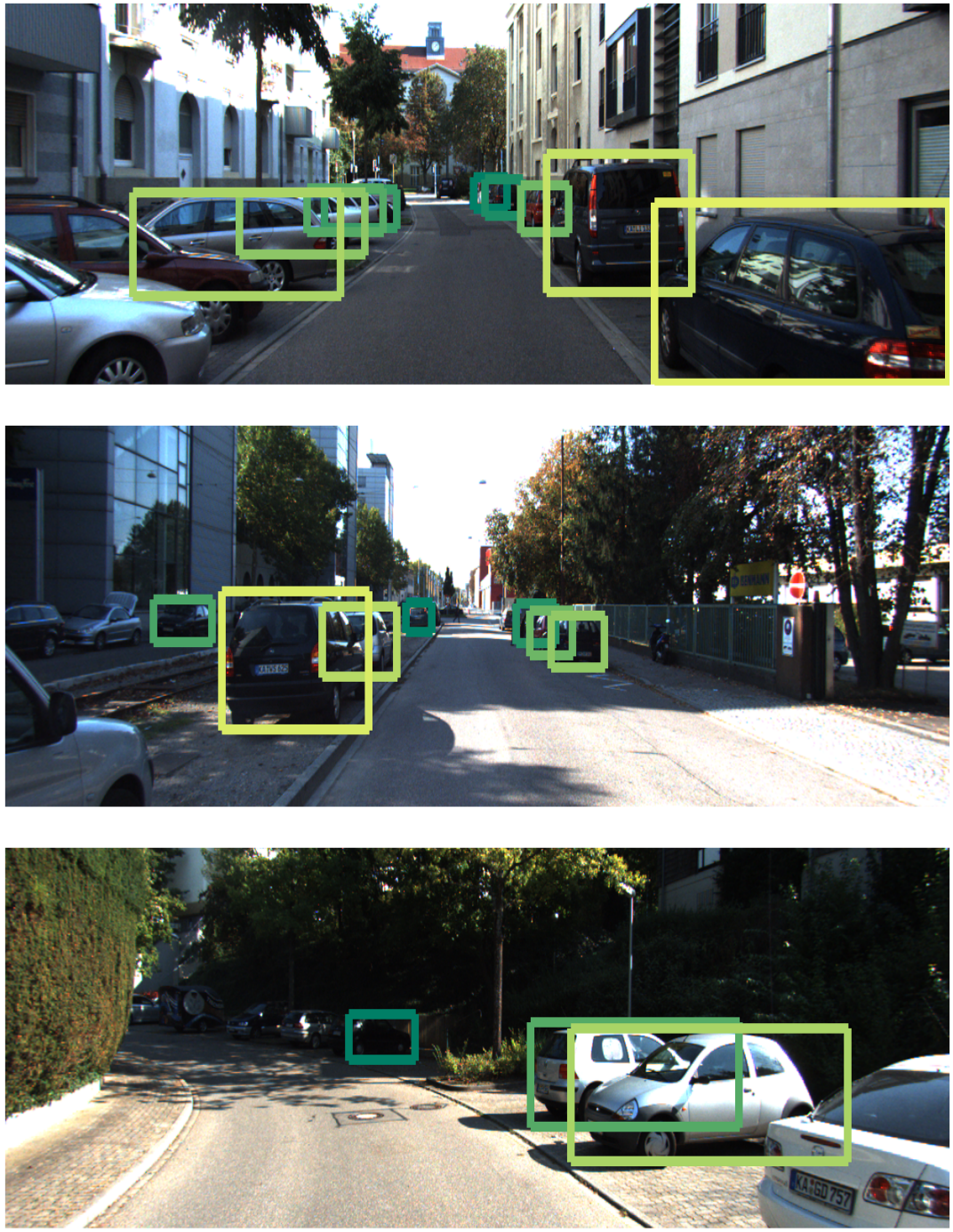
4. 3D Localization

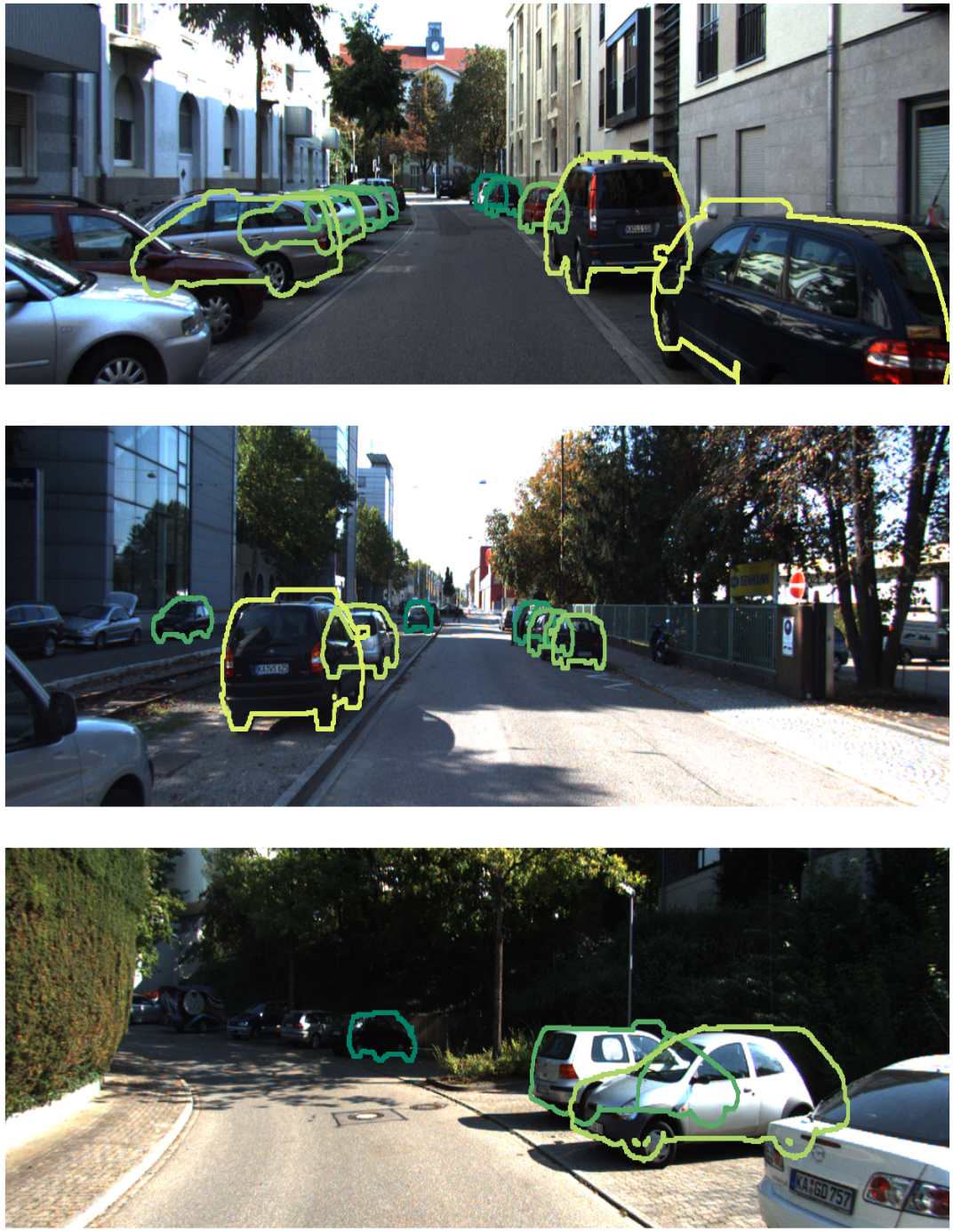


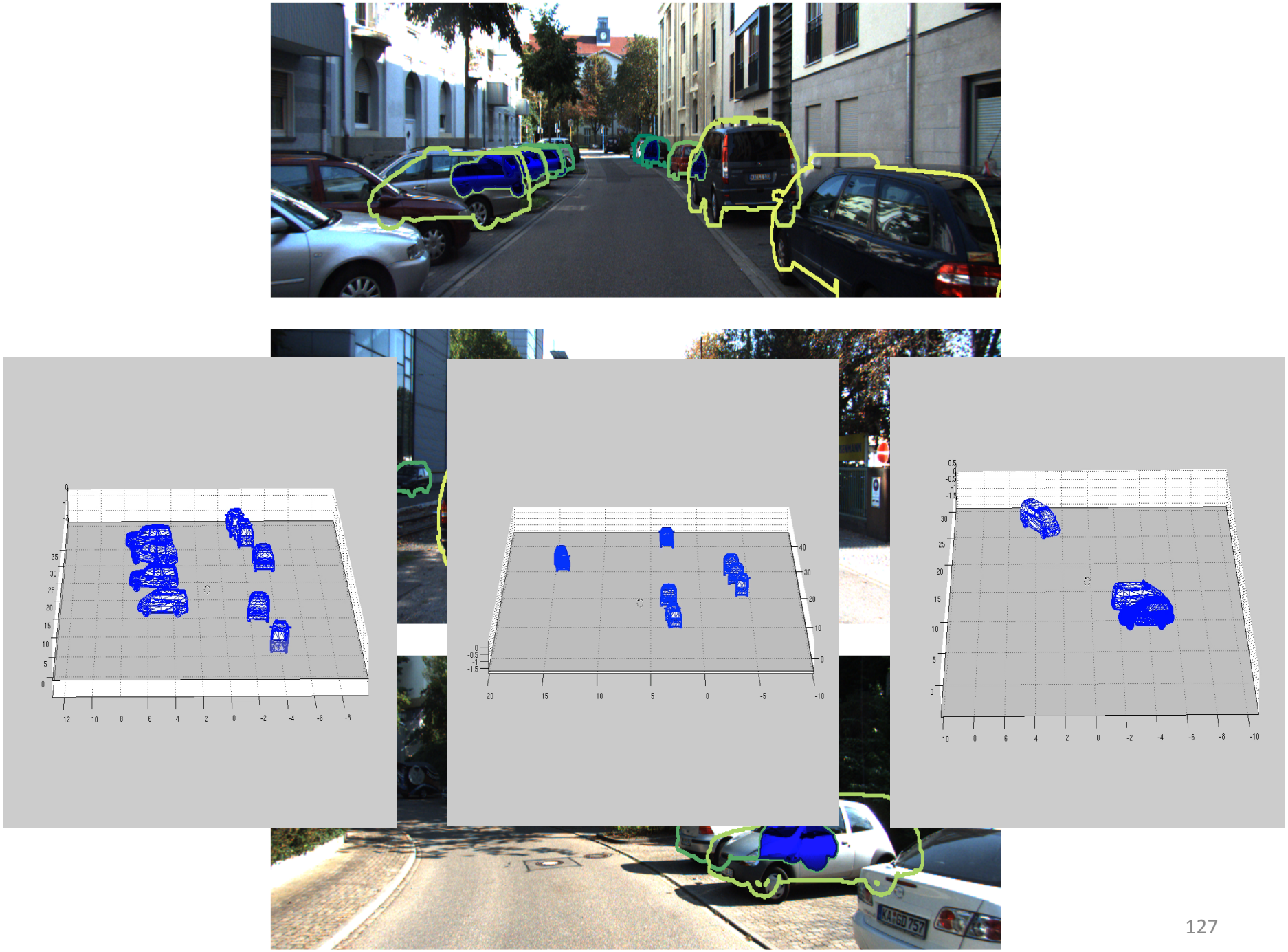


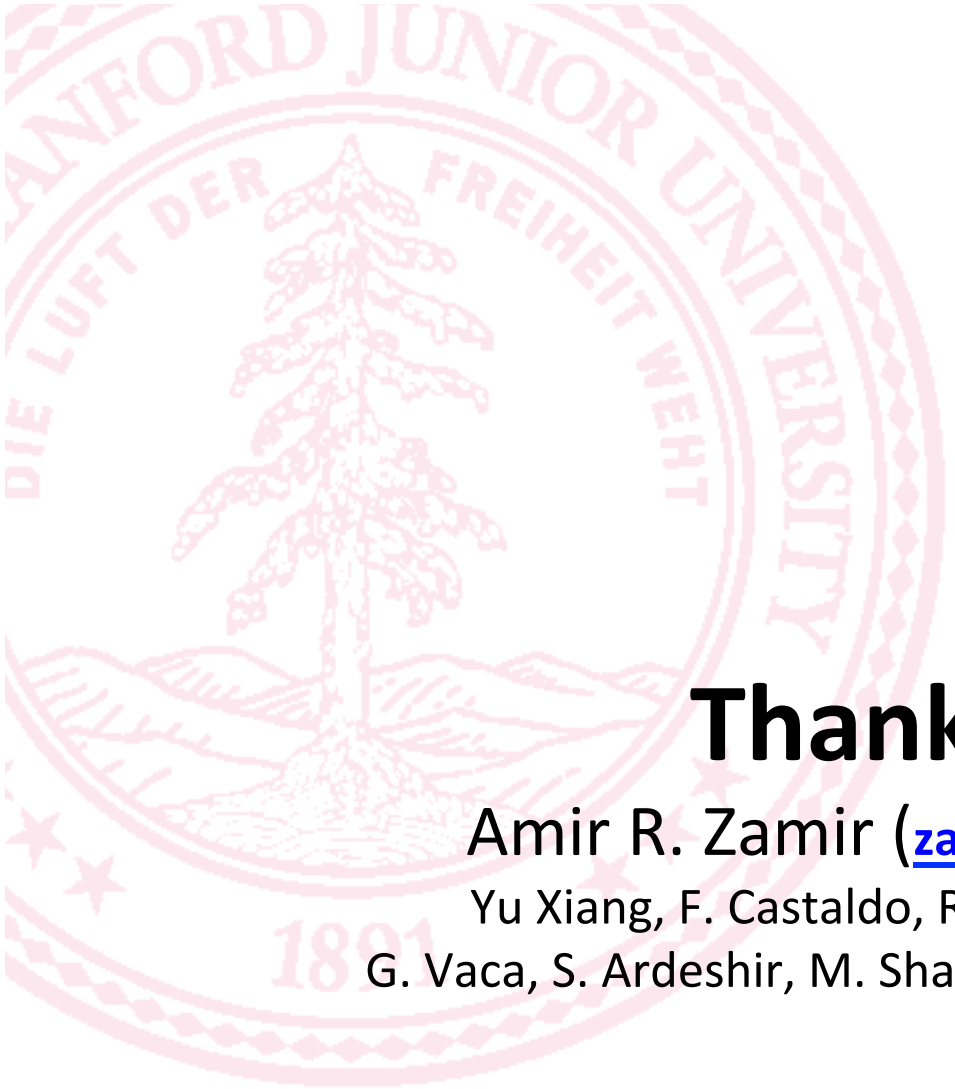
Backprojection

Anecdotal Results on KITTI









Thank You!

Amir R. Zamir (zamir@cs.stanford.edu)
Yu Xiang, F. Castaldo, R. Angst, Wongun Choi,
G. Vaca, S. Ardeshir, M. Shah, Yuanqing Lin, S. Savarese

- **CVPR'15**: Data Driven 3D Voxel Patterns for Object Category Recognition
- **ECCV'14:** GIS-Assisted Object Detection and Geospatial Localization
- **ACM Multimedia'13:** Visual Business Recognition - A Multimodal Approach
- **CVPR'12:** City Scale Geo-spatial Trajectory Estimation of a Moving Camera