Scene Representation II
Simple Maps

D.A. Forsyth, UIUC

Goal: Road Layout Map

- With minimal/no labelling
- In nasty geometries

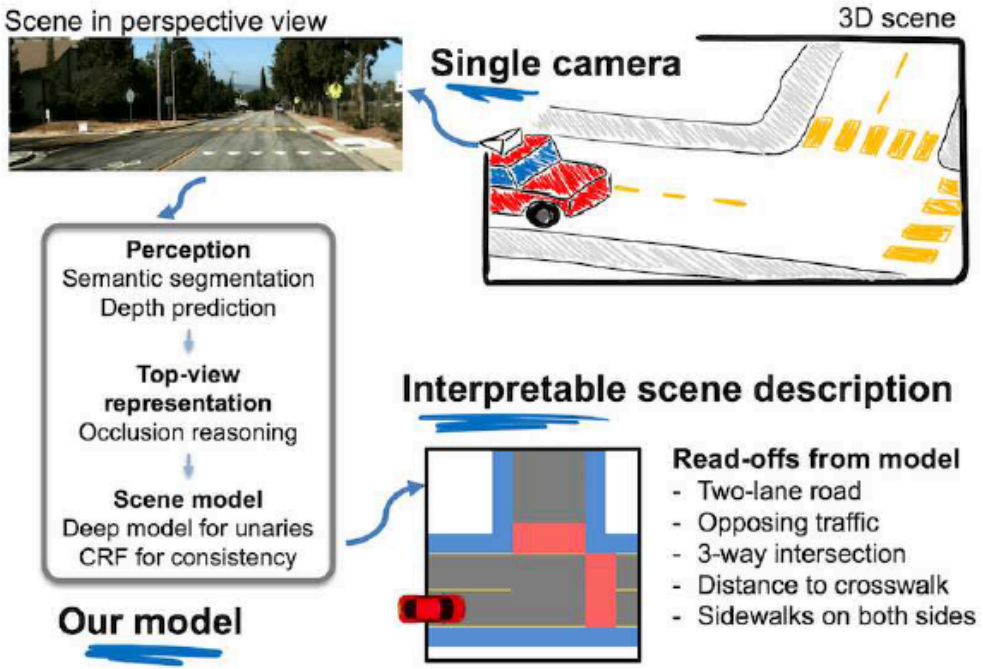


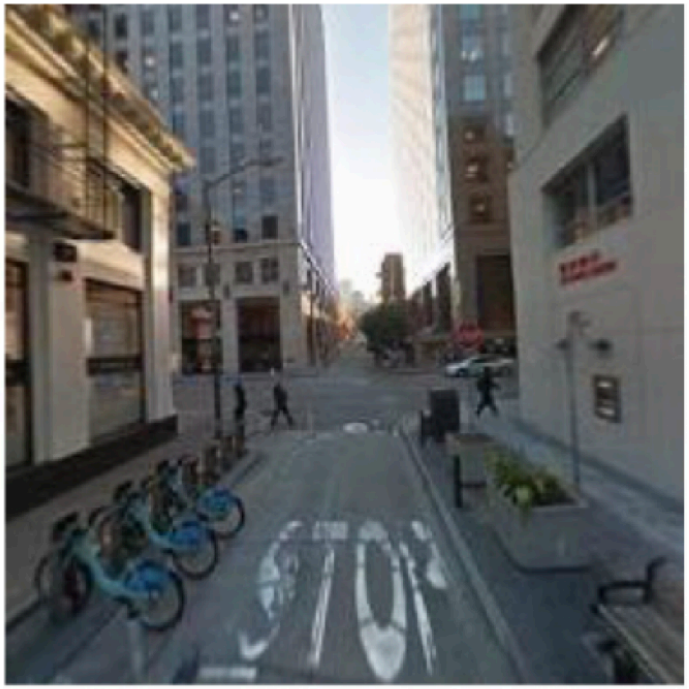
Figure 1: Our goal is to infer the layout of complex driving scenes from a single camera. Given a perspective image (top left) that captures a 3D scene, we predict a rich and interpretable scene description (bottom right), which represents the scene in an occlusion-reasoned semantic top-view.

Wang et al 19

Road layout maps

- A representation of the main scene in front
  - in terms of properties/what you can do
    - how far is the next intersection?
    - what is it like
    - is there a bike lane
  - as an explicit map
    - distinguish between
      - transients (cars, pedestrians, etc)
      - and persistent (road, walkways, bicycle lanes, buildings)
    - including
      - intersections
      - lane boundaries

(Part of) An implicit road layout map



Pred = 18.5 m

Seff+Xiao

An explicit road layout map

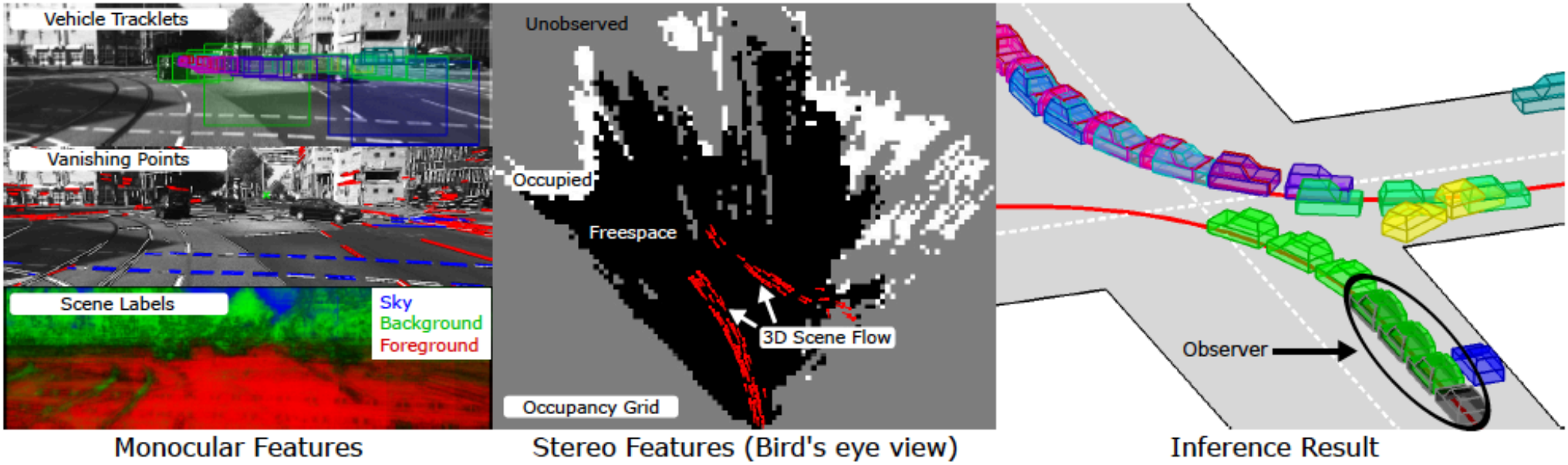


Fig. 1: **3D Intersection Understanding.** Our system makes use of monocular (left) and stereo (middle) feature cues to infer the road layout and the location of traffic participants in the scene (right) from short video sequences. The observer is depicted in black.

Geiger et al

Road layout maps

- Potential cues
  - streetview
  - openmaps
  - labelled segmentation data
  - layout is stylized
  - persistent categories have coherent (but variable) appearance
  - scene flow/photometric consistency

Semantic segmentation

Cues

- Incidental data
  - streetview+openmaps
- layout is stylized
- persistent categories have coherent appearance
- scene flow/photometric consistency

Partially supervised cues

- Open Street Maps (OSM)

Map data: OpenStreetMap is an open-source mapping project covering over 21 million miles of road. Unlike proprietary maps, the underlying road coordinates and metadata are freely available for download. Accuracy and overlap with Google Maps is very high, though some inevitable noise is present as information is contributed by individual volunteers or automatically extracted from users' GPS trajectories. For example, roads in smaller cities may lack detailed annotations (e.g., the number of lanes may be unmarked). These inconsistencies result in varying-sized subsets of the data being applicable for different attributes.

Seff+Xiao

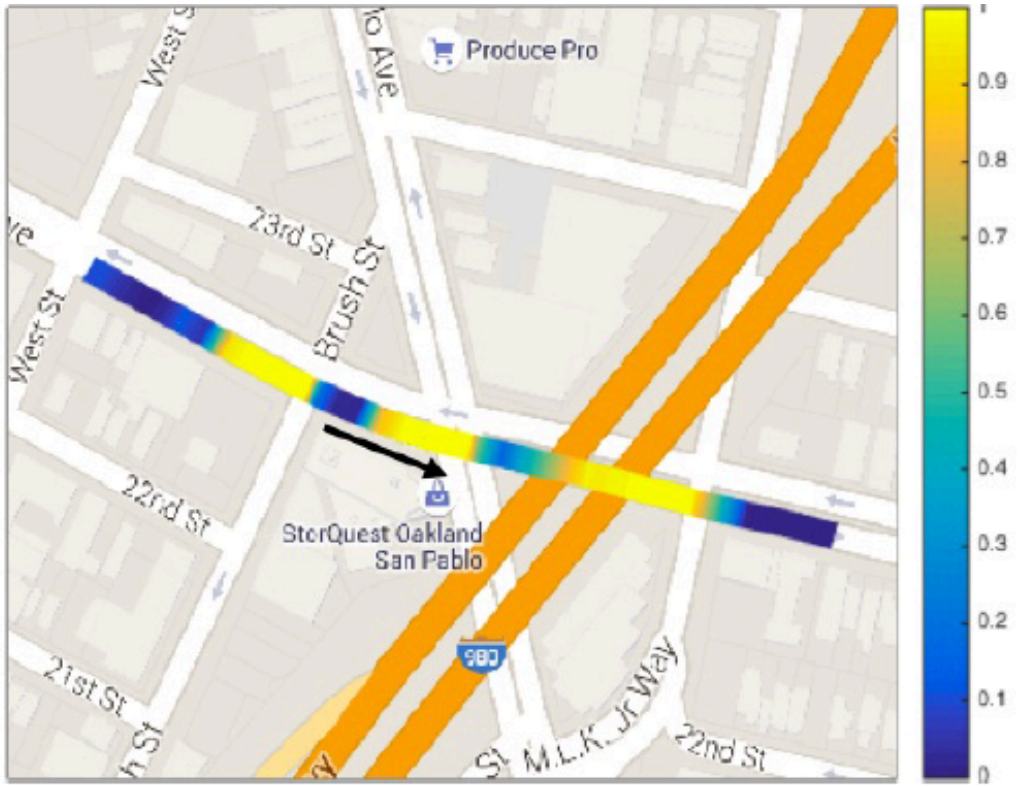


Fig. 3. Intersection detection heatmap. Images are cropped from test set GSV panoramas in the direction of travel indicated by the black arrow. The probabilities of "approaching" an intersection output by the trained ConvNet are overlaid on the road. (The images are from the ground level road, not the bridge.)

Seff+Xiao

Partially supervised cues

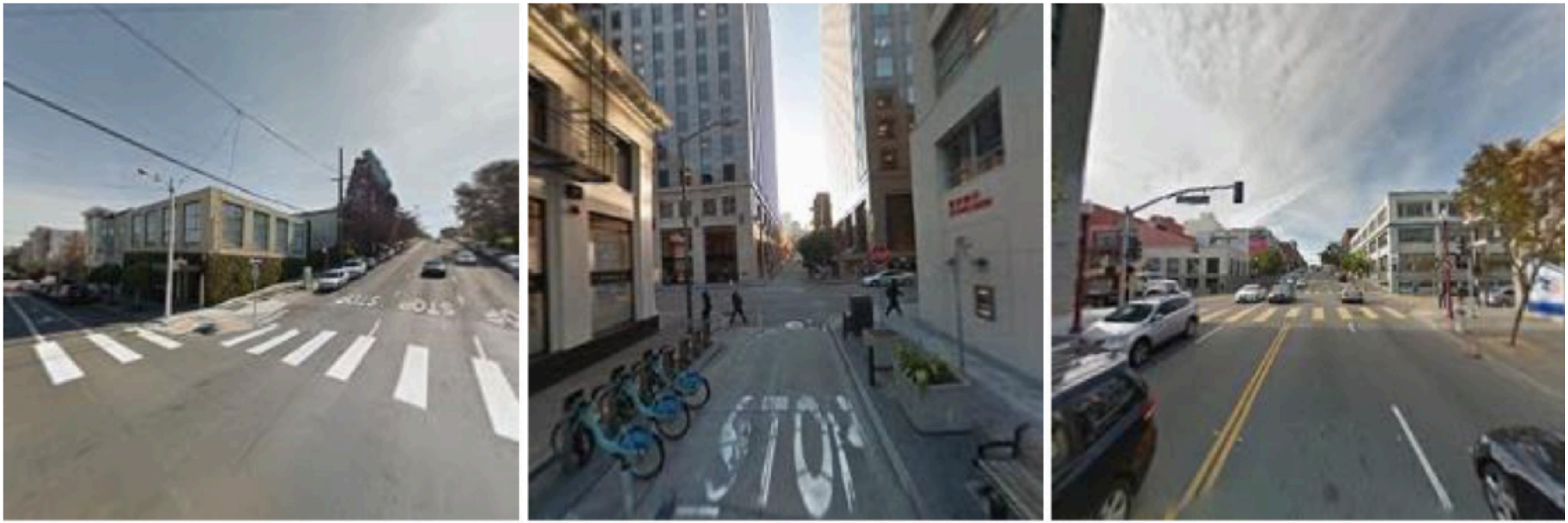
- Google street view

Seff+Xiao

Image collection: Google Street View contains panoramic images of street scenes covering 5 million miles of road across 3,000 cities. Each panorama has a corresponding metadata file storing the panorama's unique "pano_id", geographic location, azimuth orientation, and the pano_ids of adjacent panoramas. Beginning from an initial seed panorama, we collect street view images by running a bread-first search, downloading each image and its associated metadata along the way. Thus far, our dataset contains one million GSV panoramas from the San Francisco Bay Area. GSV panoramas can be downloaded at several different resolutions (marked as "zoom levels"). Finding the higher zoom levels unnecessary for our purposes, we elected to download at a zoom level of 1, where each panorama has a size of 832×416 pixels.

Labelling - I

- Match panoramas to roads
  - panorama center location, orientation is known
  - (essentially) project to plane
  - thresholded nearest neighbor to road center polyline
    - thresholding removes panoramas inside buildings, etc.
  - some noise
    - under bridges, etc.
- Annotations
  - Intersections
  - Drivable heading
  - Heading angle
  - Bike lane
  - Speed limit, wrong way, etc.



Pred = 0.1 m
True = 1.9 m

Pred = 18.5 m
True = 19.2 m

Pred = 22.9 m
True = 22.4 m

Fig. 4. Distance to intersection estimation. For images within 30 m of true intersections, our model is trained to estimate the distance from the host car to the center of the intersection across a variety of road types.

Seff+Xiao

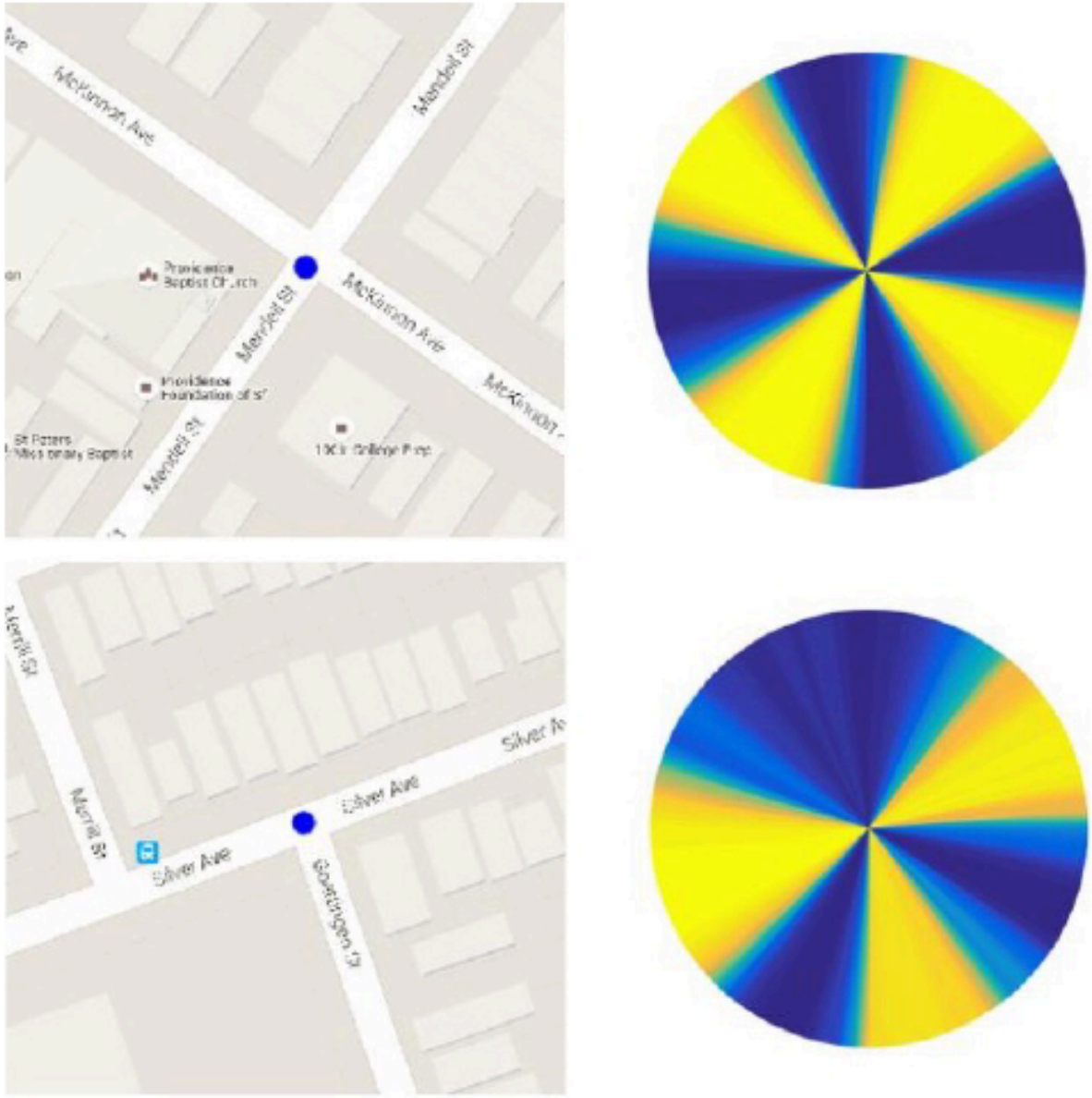
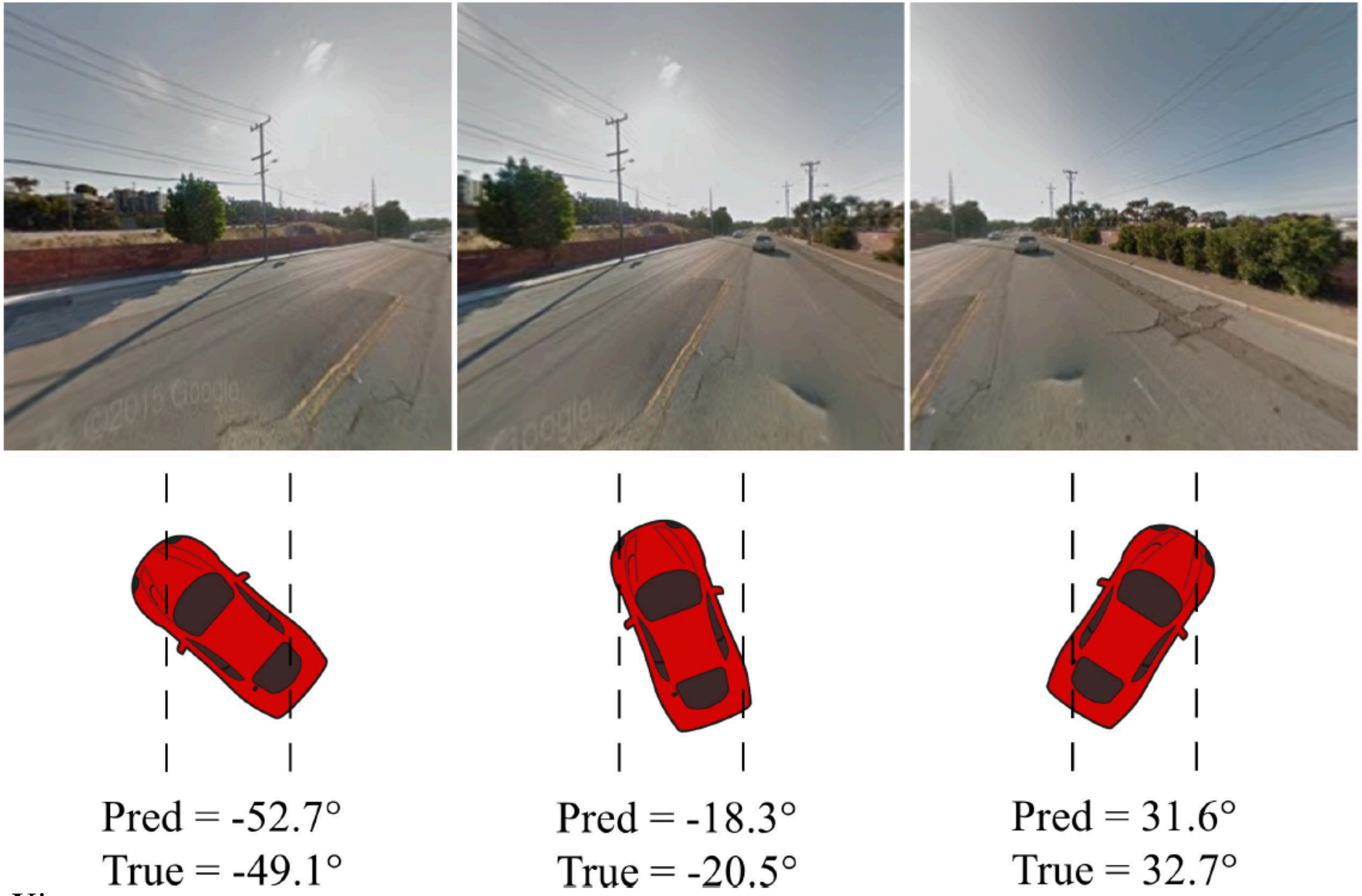


Fig. 5. Intersection topology is one of several attributes our model learns to infer from an input GSV panorama. The blue circles on the Google Maps extracts to the left show the locations of the input panoramas. The pie charts display the probabilities output by the trained ConvNet of each heading angle being on a driveable path (see Figure 3 for colormap legend).

Seff+Xiao

Fig. 6. Driveable headings. A ConvNet is trained to distinguish between non-drivable headings (left) and drivable headings aligned with the road (right). The ConvNet weakly classifies the middle example as drivable because the host car's heading is facing the alleyway between the buildings.

Seff+Xiao



Seff+Xiao

Fig. 7. Heading angle regression. The network learns to predict the relative angle between the street and host vehicle heading given a single image cropped from a GSV panorama. Below each GSV image, the graphic visualizes the ground truth heading angle.

p(bike lane) = 0.043 p(bike lane) = 0.604 p(bike lane) = 0.988

Fig. 8. The ConvNet learns to detect bike lanes adjacent to the vehicle. The GSV images are arranged from left to right in increasing order of probability output by the ConvNet of a bike lane being present (ground truth labels from left to right are negative, negative, positive). The middle example contains a taxi lane, resulting in a weak false positive.

Seff+Xiao

Pred = 26.1 mph
True = 30 mph

Pred = 30.0 mph
True = 50 mph

Pred = 54.3 mph
True = 50 mph

Fig. 9. Speed limit regression. The network learns to predict speed limits given a GSV image of road scene. The model significantly underestimates the speed limit in the middle example as this type of two-way road with a single lane in each direction would generally not have a speed limit as high as 50 mph.

Fig. 10. One-way vs. two-way road classification. The probability output by the ConvNet of each GSV scene being on a one-way road is shown. From left to right the ground truth labels are two-way, two-way, and one-way. The image on the left is correctly classified as two-way despite the absence of the signature double yellow lines.

Seff+Xiao

Fig. 11. Wrong way detection. The probability output by the ConvNet of each GSV image facing the wrong way on the road is displayed. From left to right the ground truth labels are wrong way, right way, and right way. For two-way roads with no lane markings (left), this is an especially difficult problem as it amounts to estimating the horizontal position of the host car. The problem can also be quite ill-defined if there are no context clues as is the case with the rightmost image.

Fig. 12. Number of lanes estimation. The predicted and true number of lanes for three roads are displayed along with the corresponding GSV images. For streets without clearly visible lane markings (left), this is especially challenging. Although the ground truth for the rightmost image is two lanes, there is a third lane that merges just ahead.

Seff+Xiao

Cues

- Incidental data
  - streetview+openmaps
- layout is stylized
- persistent categories have coherent appearance
- scene flow/photometric consistency