Data-Driven 3D Voxel Patterns for Object Category Recognition

Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese

Adam Allevato
April 6, 2016
CS381V: Visual Recognition, UT Austin

Outline

- Problem Statement
- Related Work
- Approach
  - Data Representation
  - Training
  - Testing
- Results
  - Examples
  - Quantitative
- Discussion

Problem Statement

How can we build a system that can **classify**, **locate**, and **orient** **occluded 3D** objects using **2D image** inputs?

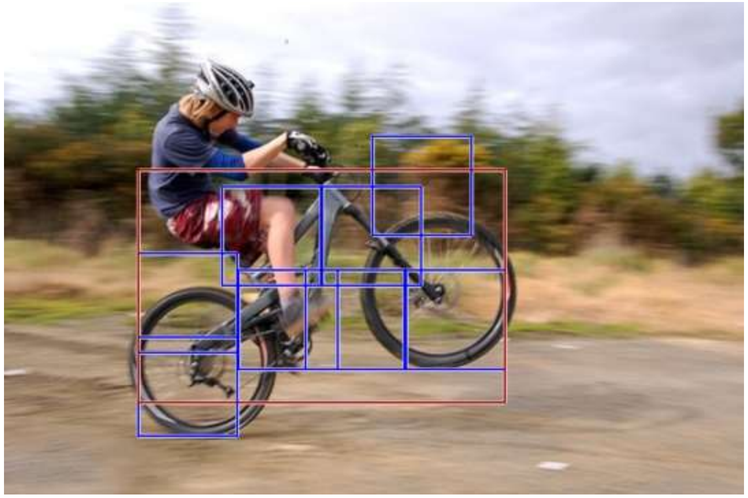
Related Work

How can we build a system that can **classify**, **locate**, and **orient** **occluded 3D** objects using **2D image** inputs?

Method	Locate in 3D	Orient in 3D	Occlusion	2D Input
2D Detection	✗	✗	✓	✓
3D Pose Estimation	✗	✓	✓	✓
Point-Cloud Based Methods	✓	✓	✓	✗
This Paper: 3D Voxel Patterns	✓	✓	✓	✓

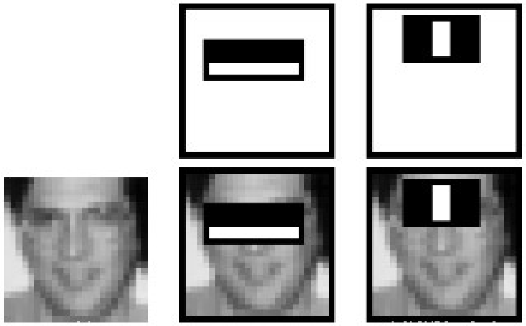
Related Work: 2D Detection

Method	Locate	Orient	Occlusion	2D Input
2D Detection (DPM)	✗	✗	✓	✓



Deformable Part Models

Felzenswalb et al. 2010

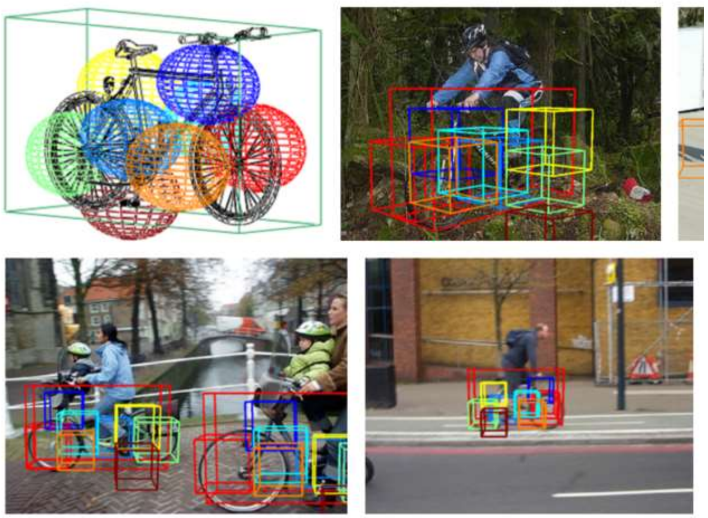


Face Detector

Viola and Jones 2004

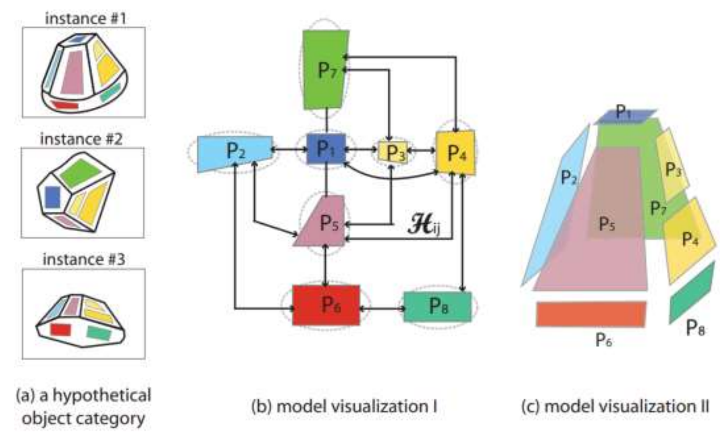
Related Work: 3D Pose Estimation

Method	Locate	Orient	Occlusion	2D Input
3D Pose Estimation	✗	✓	✓	✓



3D²PM

Pepik et al. 2012

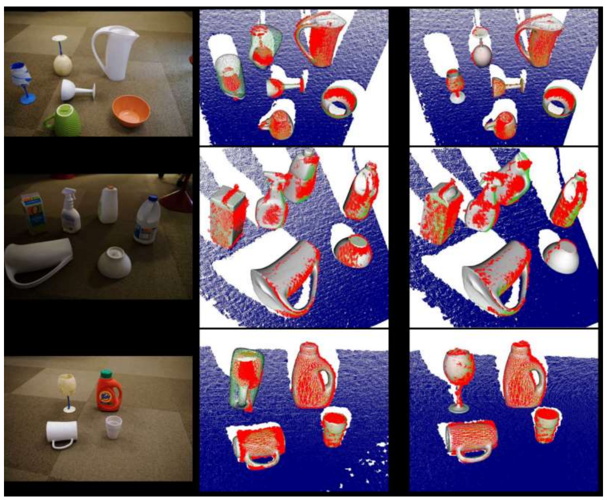


3D Category Classification

Savarese and Fei-Fei 2007

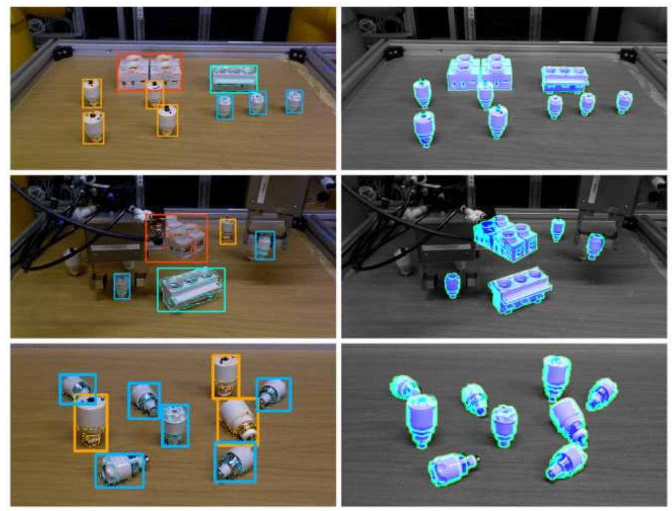
Related Work: Point-Cloud Based Methods

Method	Locate	Orient	Occlusion	2D Input
Point-Cloud Based Methods	✓	✓	✓	✗



Clustered Viewpoint Feature Histogram

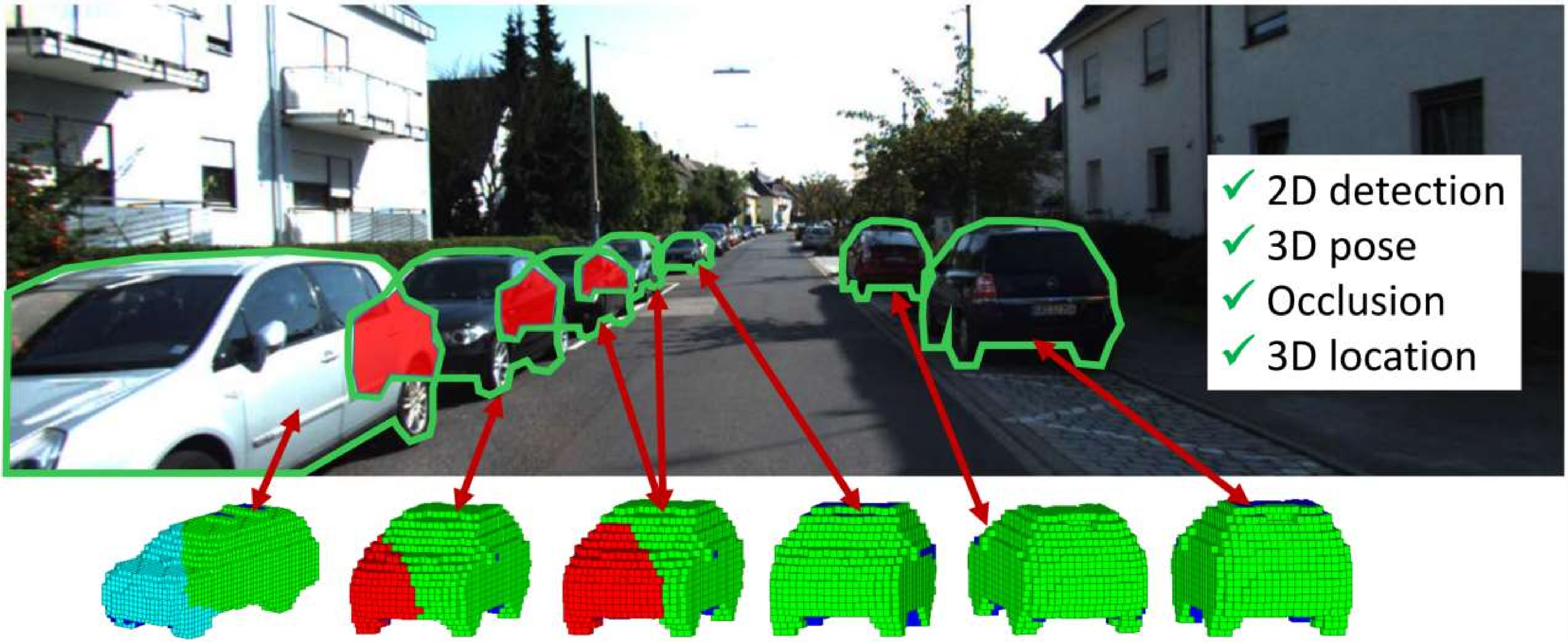
Aldoma 2011



Hashed 3D Voting

Hodaň et al. 2015

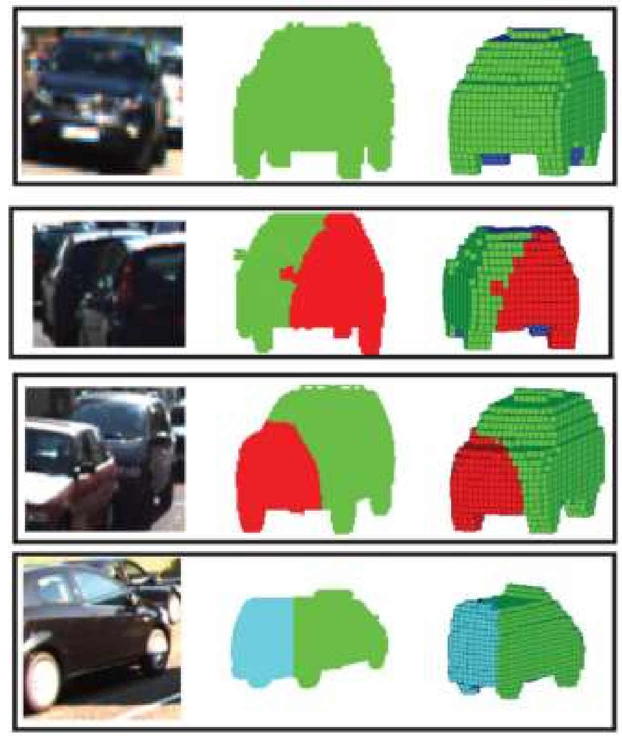
Data-Driven 3D Voxel Patterns



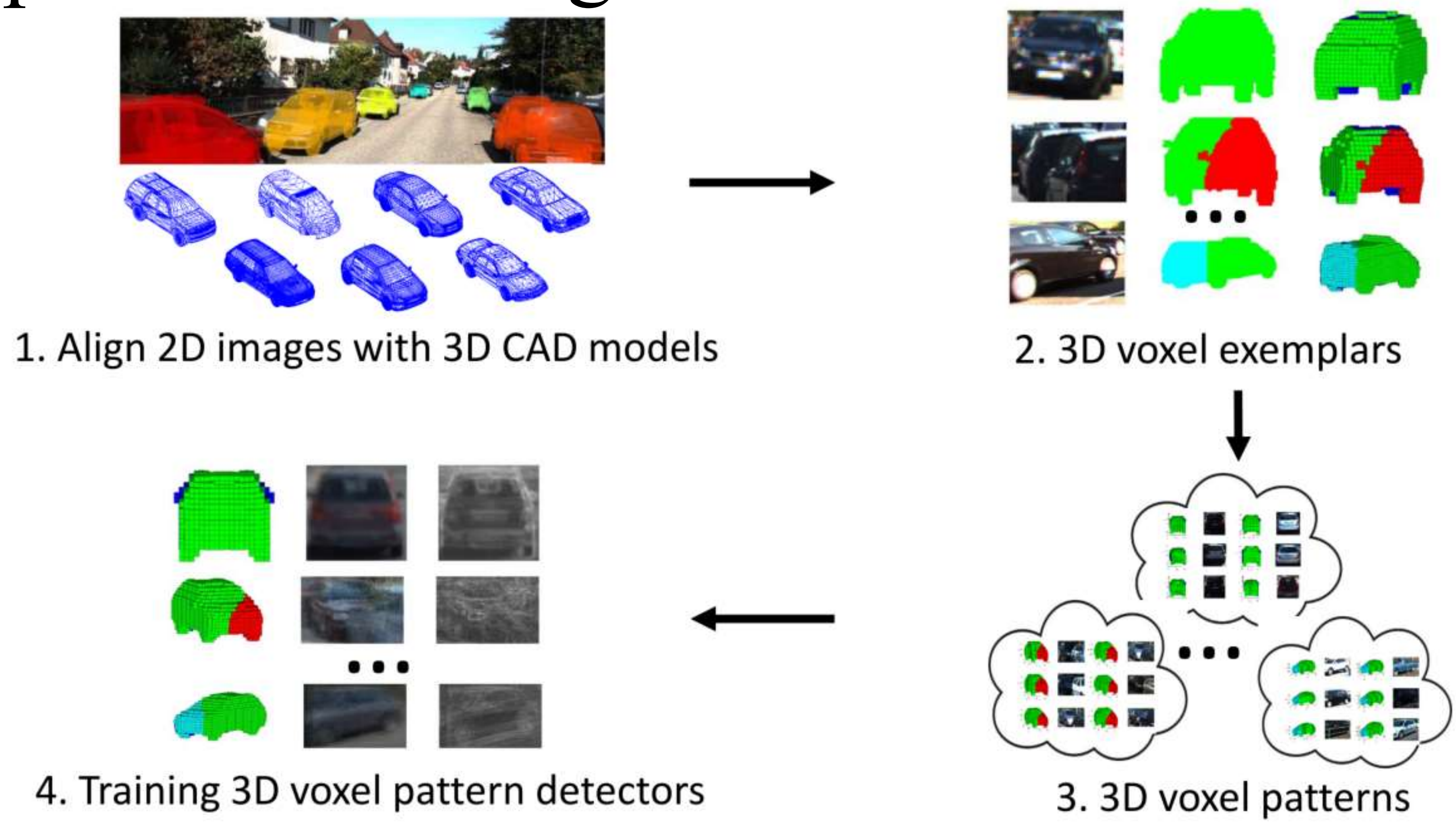
Slide Credit: Yu Xiang

Approach: Data Representation

- **3D Voxel Patterns** (3DVPs)
  - Capture "patterns of visibility"
  - Composed of four parts:
    - 2D object image
    - 2D segmentation mask
    - 3D voxel model
    - Metadata: pose, 3D model
- Voxels and pixels can be
  - Visible (green)
  - Occluded (red)
  - Truncated (cyan)
  - Self-occluded (blue)



Approach: Training

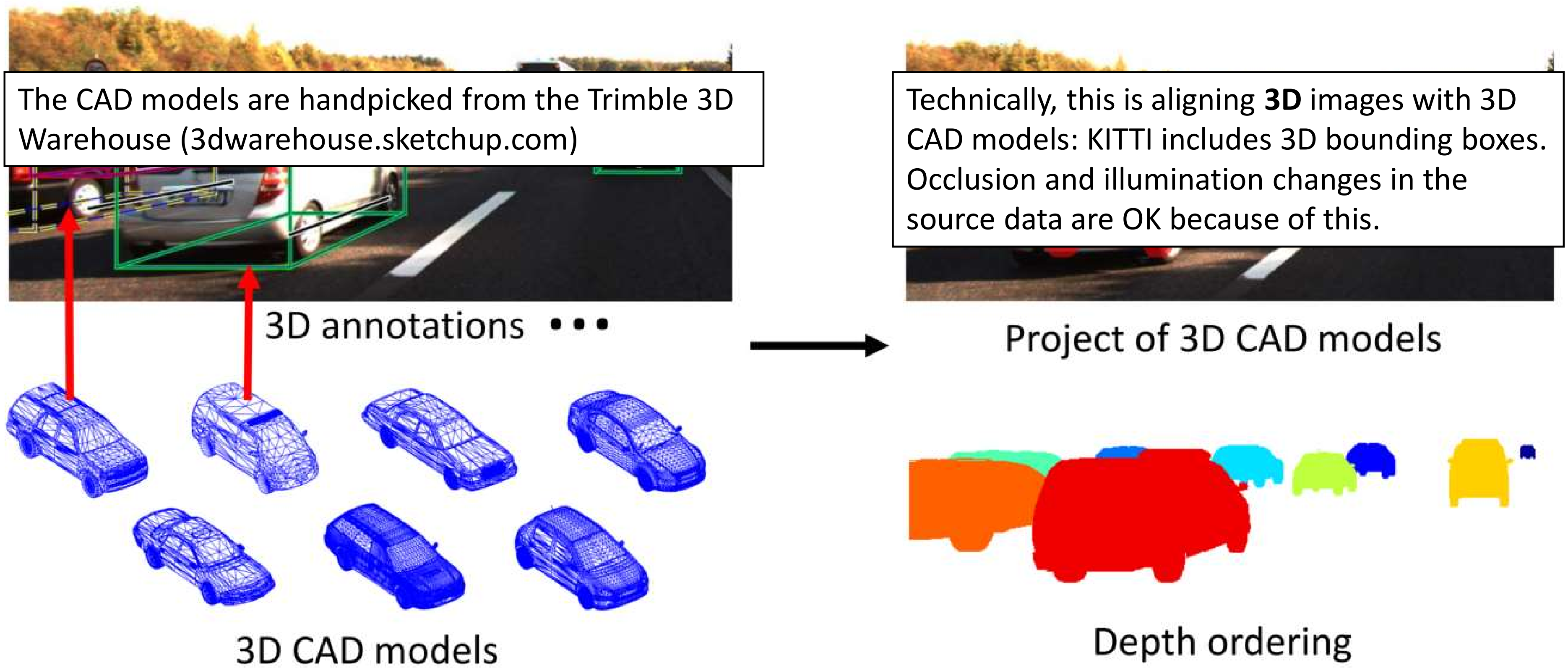


Slide Credit: Yu Xiang

1. Align 2D Images with 3D CAD Models

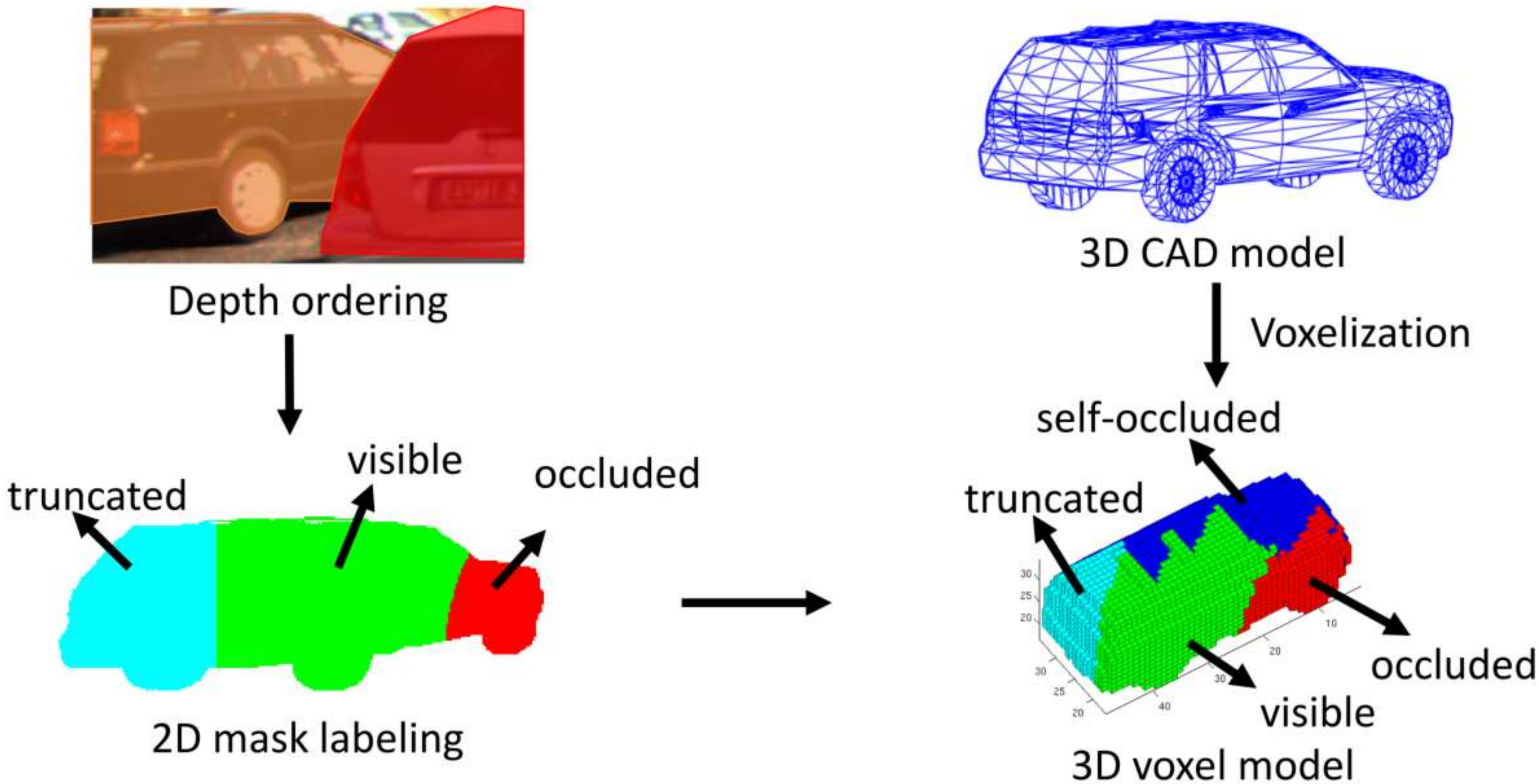
The CAD models are handpicked from the Trimble 3D Warehouse (3dwarehouse.sketchup.com)

Technically, this is aligning **3D** images with 3D CAD models: KITTI includes 3D bounding boxes. Occlusion and illumination changes in the source data are OK because of this.



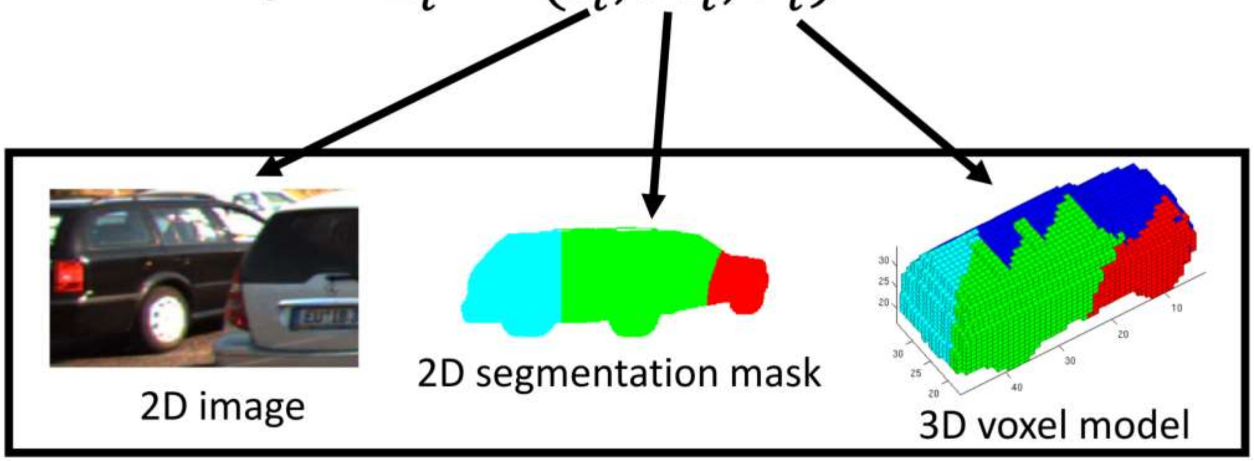
A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012

2. Building Voxel Exemplars: Baby 3DVPs



2. Building Voxel Exemplars: Baby 3DVPs

A 3D voxel exemplar $E_i = (I_i, M_i, V_i)$ *+ metadata*



Metadata includes 3D CAD model (classification) and 3D pose

57,224 voxel exemplars found in KITTI

Slide Credit: Yu Xiang

3. Discovering 3D Voxel Patterns

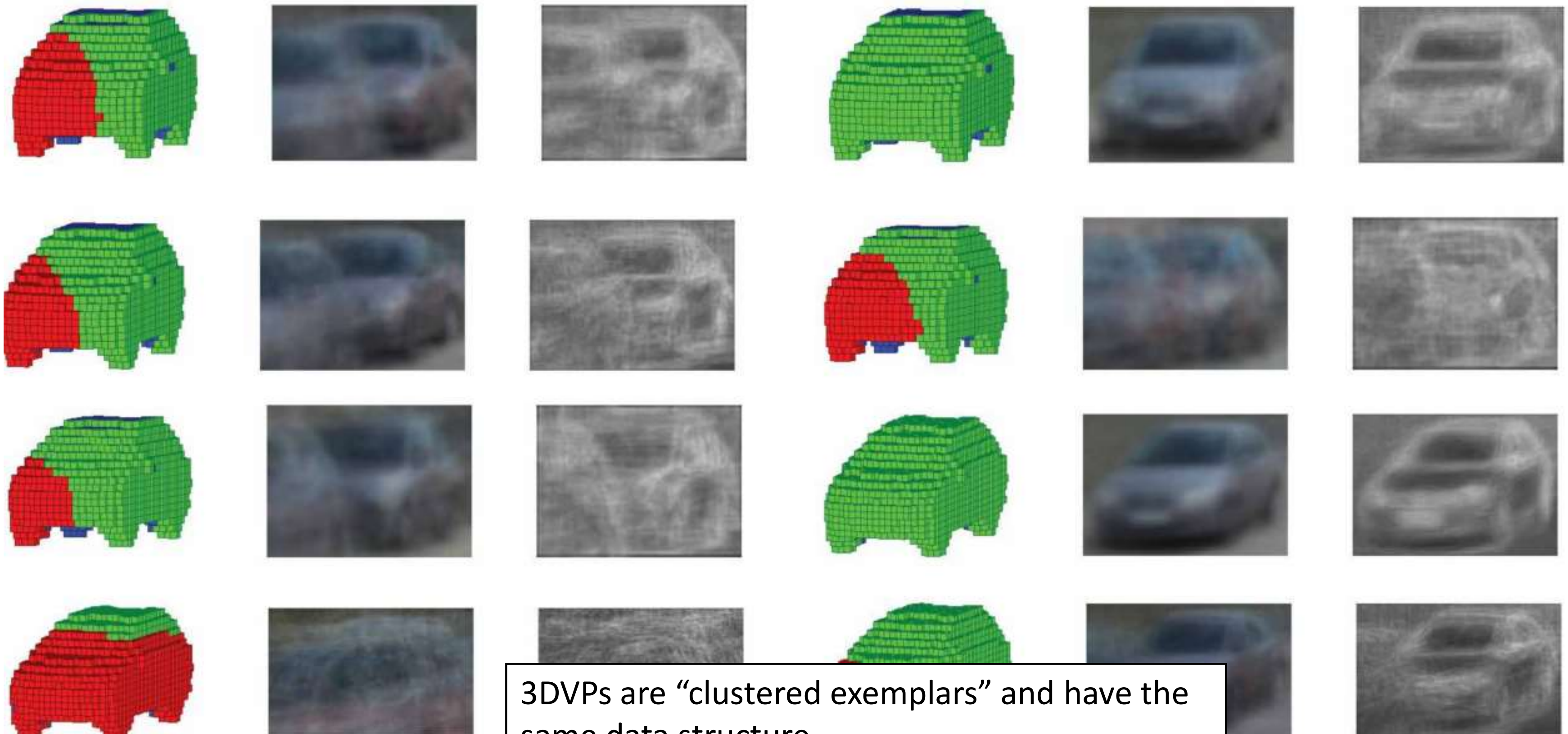
- First, generate mirror-image voxel patterns to increase training set
- Similarity metric for clustering:

$$s(\mathbf{x_1}, \mathbf{x_2}) = \frac{|\mathcal{S}|}{N^3} \sum_{i=1}^{N^3} \mathbb{1}(x_1^i = x_2^i) \cdot w(x_1^i),$$

$$\text{s.t.,} \sum_{i=0}^{|\mathcal{S}|-1} w(i) = 1,$$

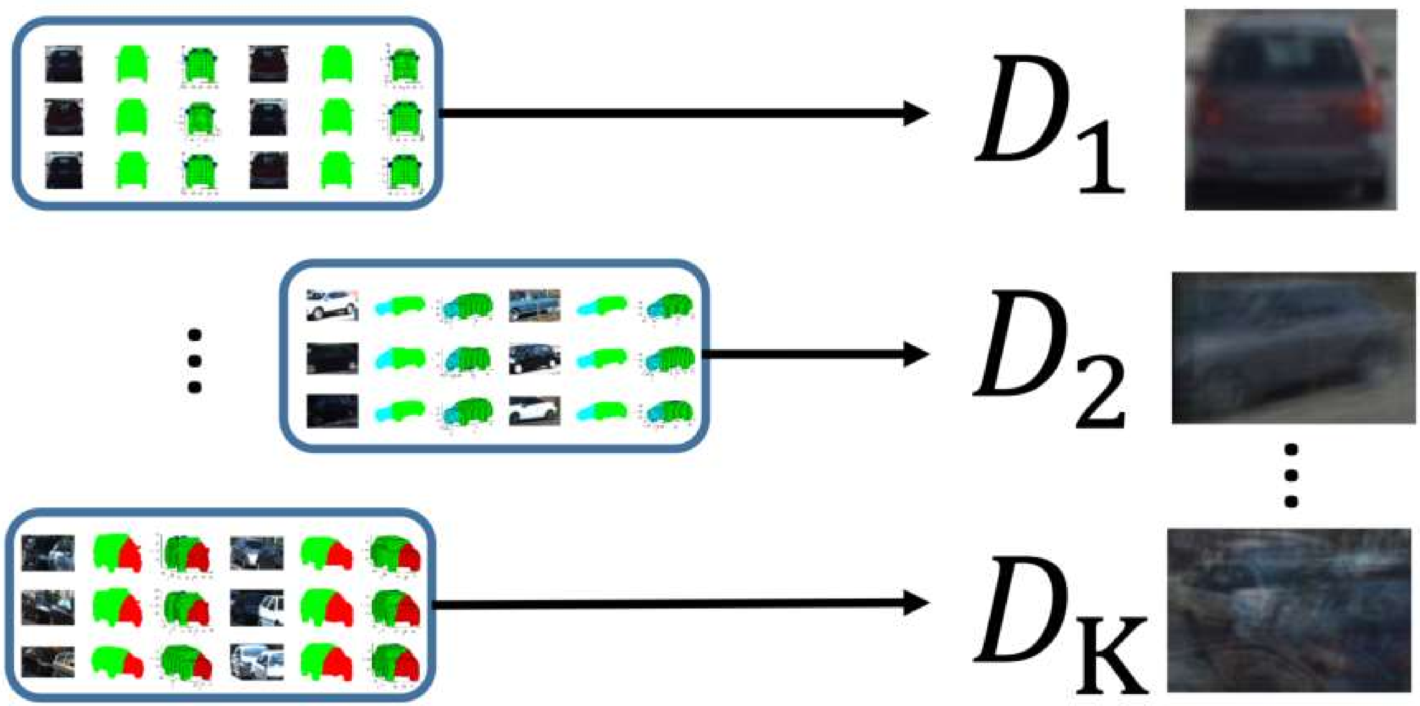
- Flexibility provided by $w(i)$, but authors use $w(i) = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$.
- Gives a “flat” similarity: evaluates agreement between voxel labels

3. Discovering 3D Voxel Patterns



3DVPs are “clustered exemplars” and have the same data structure

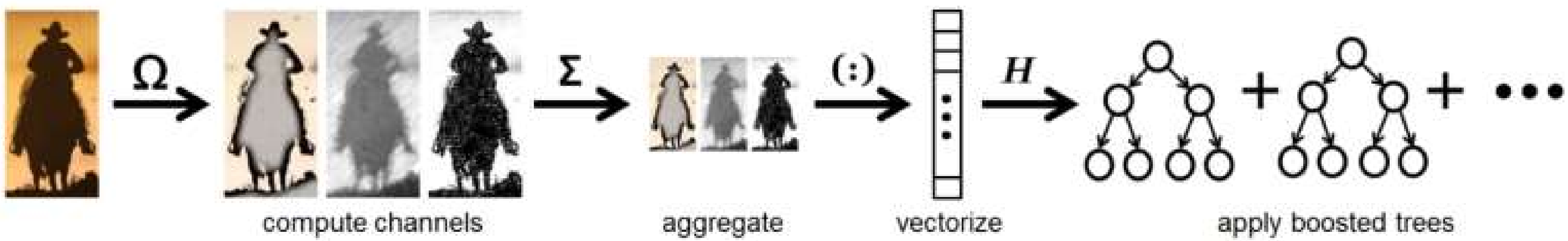
4. Training 3D Voxel Pattern Detectors



Slide Credit: Yu Xiang

4. Training 3D Voxel Pattern Detectors

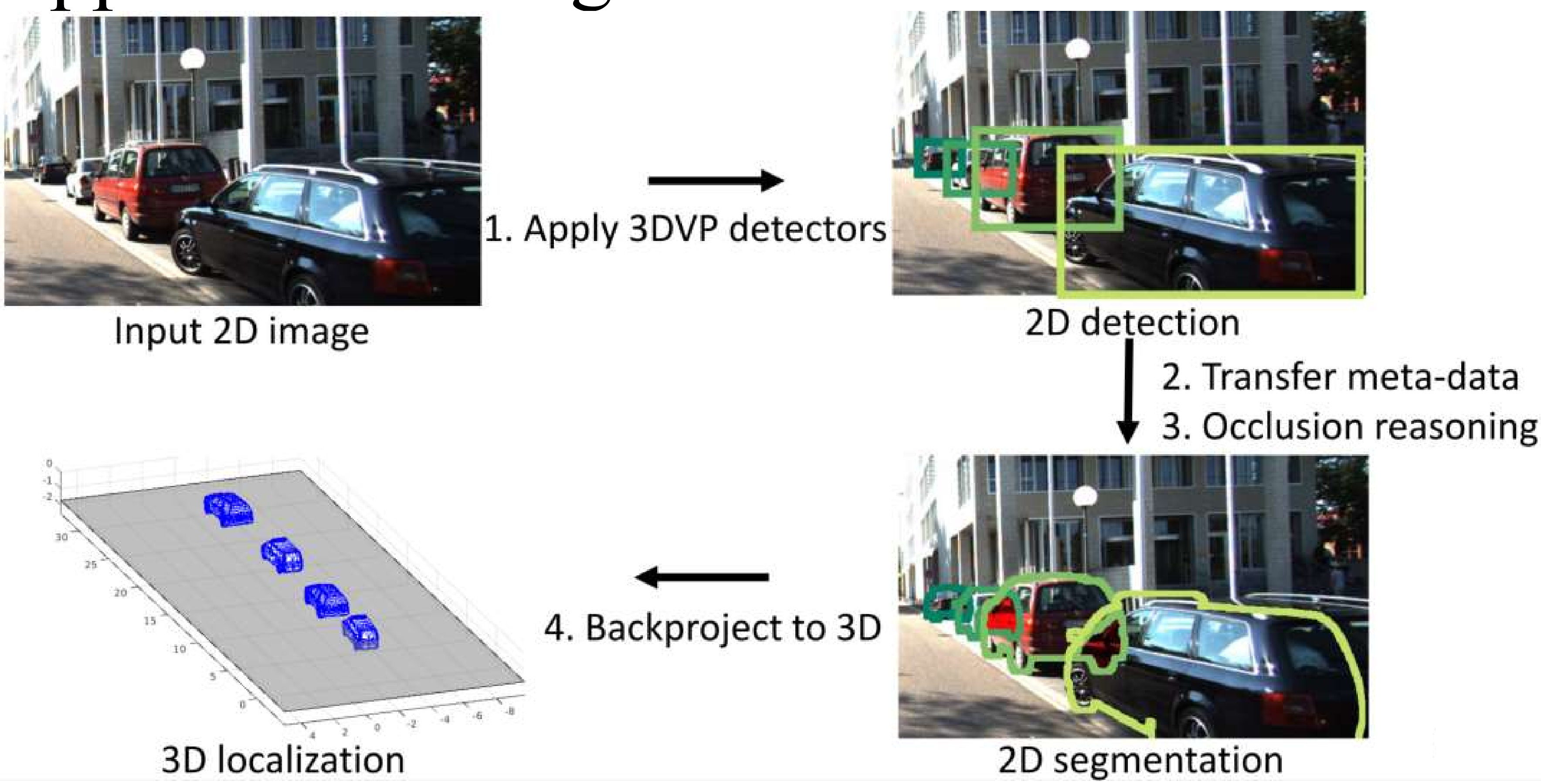
- Train a detector for each 3DVP using Aggregated Channel Features (ACF): Blur the image, then split into channels and downsample 4x



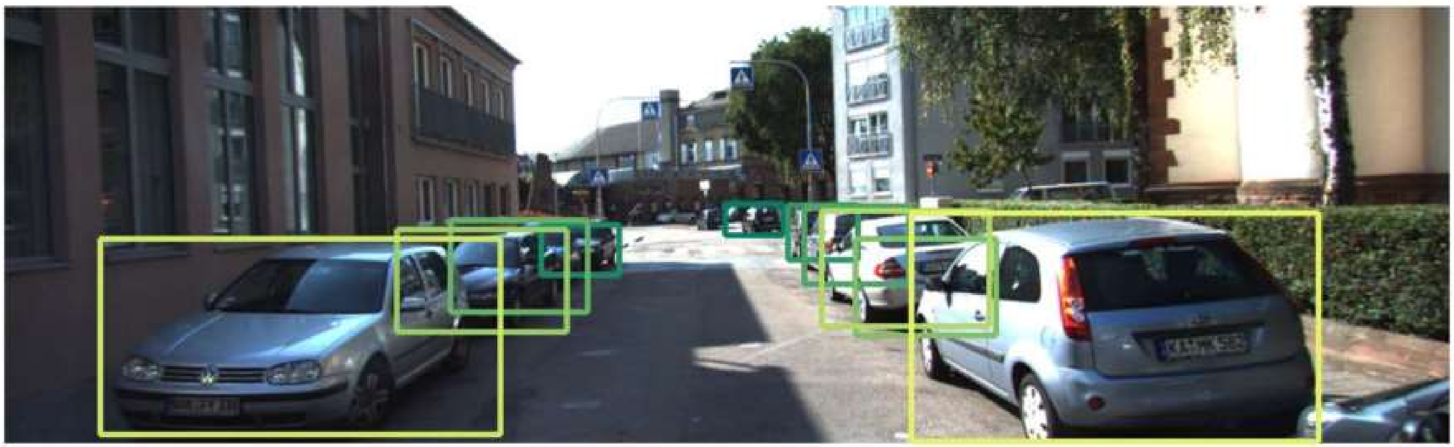
- Channels in Ω: gradients, HoG features, and LUV channels.
- Use boosting on pixels selected from all channels at once to build discriminative trees
- Realtime algorithm developed in 2014, over 200 citations

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

Approach: Testing

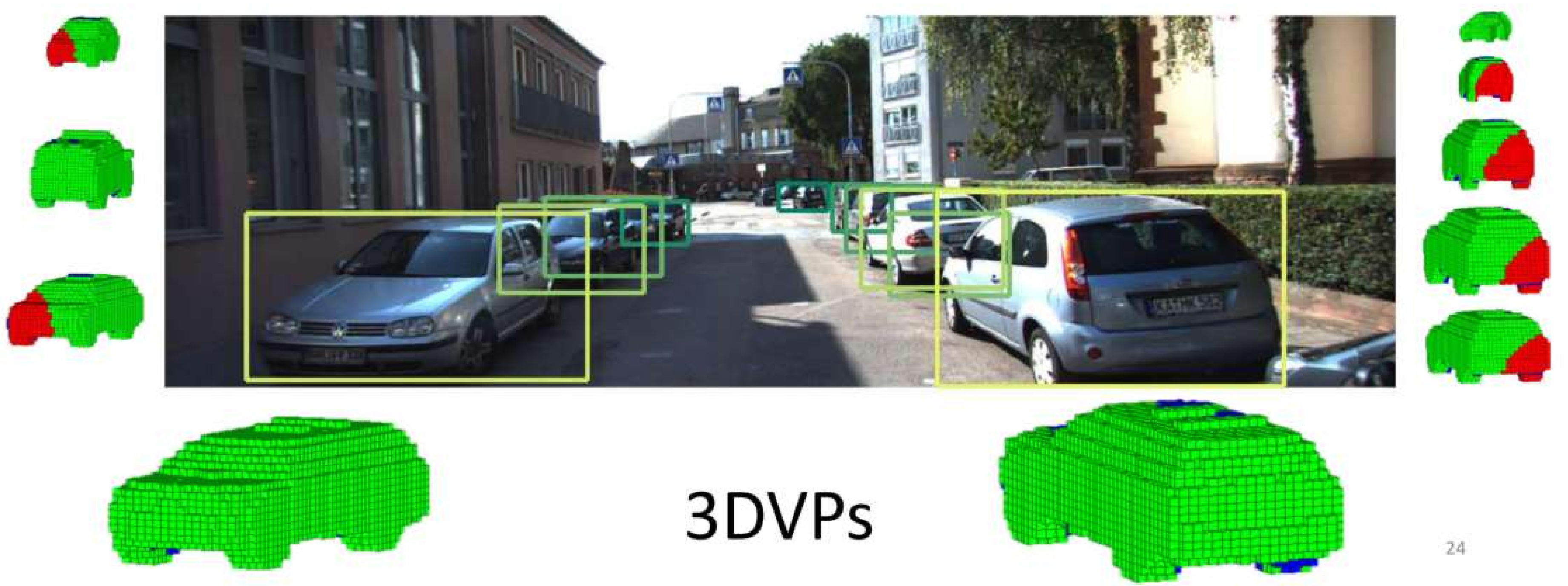


1. Apply 3DVP Detectors



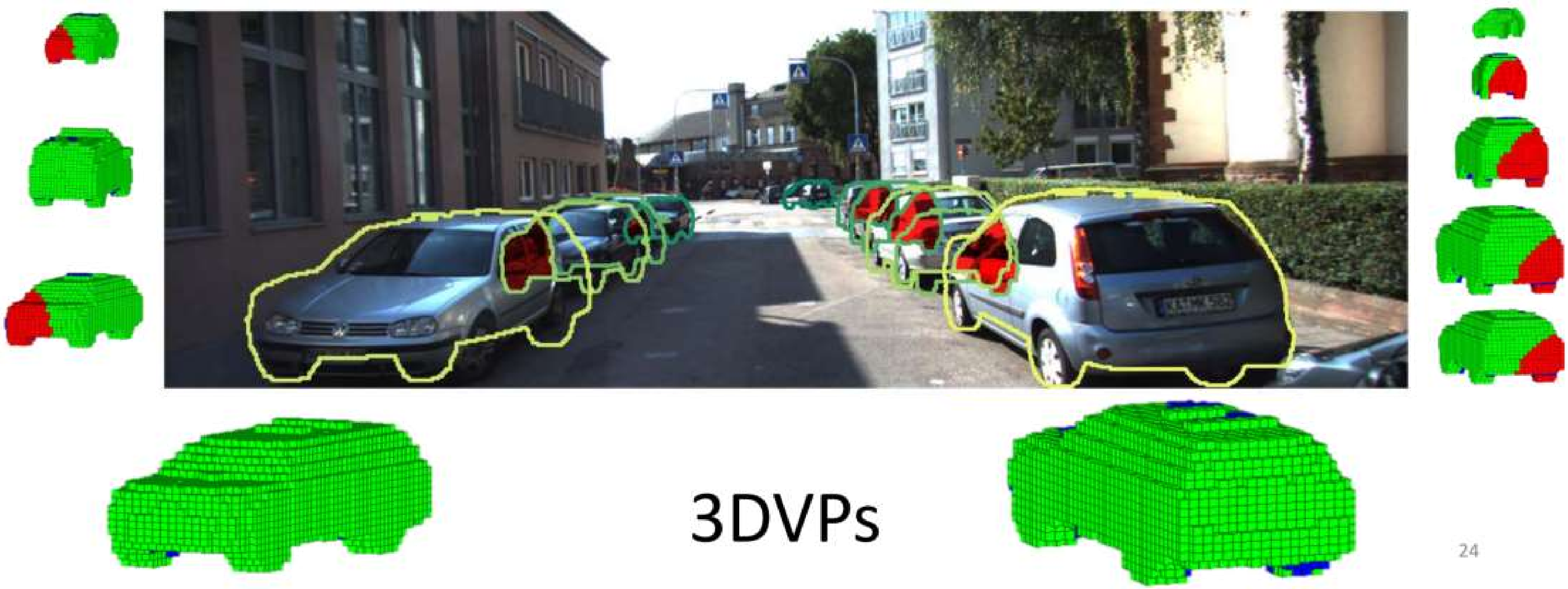
Testing images are
2D only

2. Collect and Apply Metadata



Slide Credit: Yu Xiang

2. Collect and Apply Metadata



3. Occlusion Reasoning

$$E = \sum_i (\psi_{\text{detection\_score}} + \psi_{\text{truncation}}) + \sum_{ij} \psi_{\text{occlusion}}$$

Slide Credit: Yu Xiang

3. Occlusion Reasoning

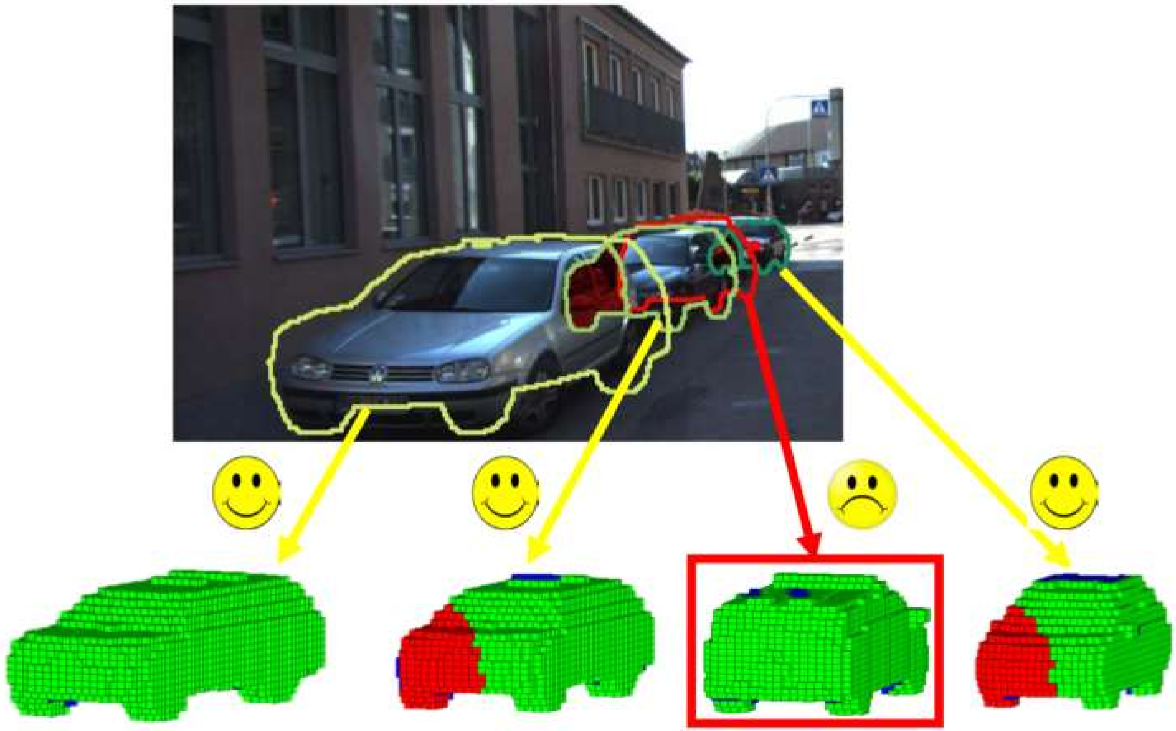
What we want:

$$E(\hat{\mathbb{D}}) = \sum_{i \in \hat{\mathbb{D}}} \Bigg(\underbrace{w_d(s_i - b)}_{\text{detection score}} - \underbrace{w_o \frac{|m_i^o| + |m_i^t|}{|m_i|}}_{\text{invisibility penalty}} + \underbrace{w_o \frac{|m_i^t \not\subset I|}{|m_i|}}_{\text{truncation explained}} \Bigg) +$$

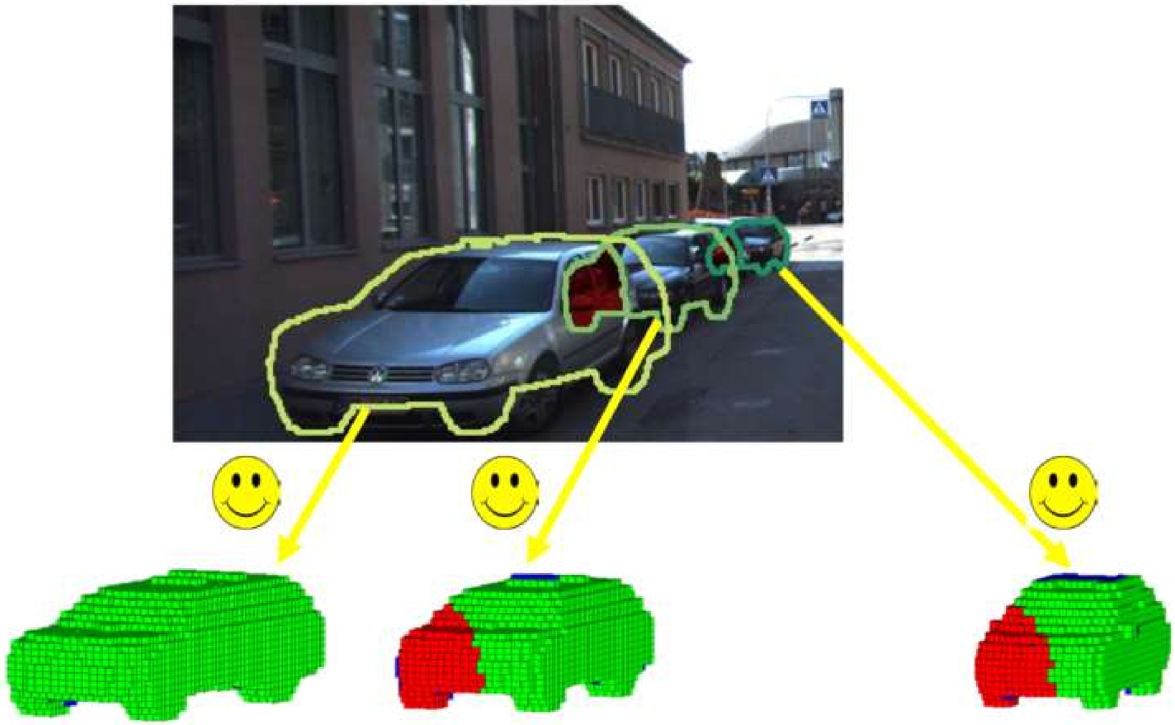
$$\sum_{i,j \in \hat{\mathbb{D}}, i \neq j} \Bigg(\underbrace{w_o \frac{|m^o_{\text{far}(i,j)} \cap m^v_{\text{near}(i,j)}|}{|m_{\text{far}(i,j)}|}}_{\text{occlusion explained}} - \underbrace{w_p \frac{\sum_{k=v,o,t} |m_i^k \cap m_j^k|}{\min(|m_i|, |m_j|)}}_{\text{overlap penalty}} \Bigg)$$

Greedy approach to maximize E – better methods may be possible

3. Occlusion Reasoning

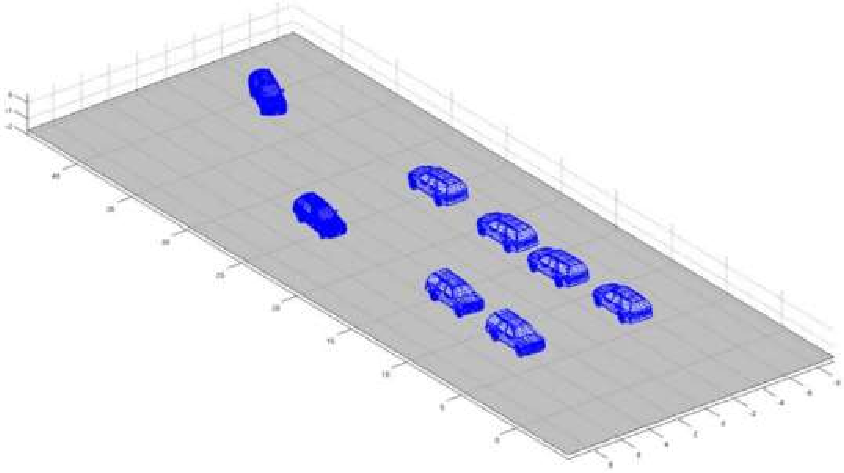


3. Occlusion Reasoning



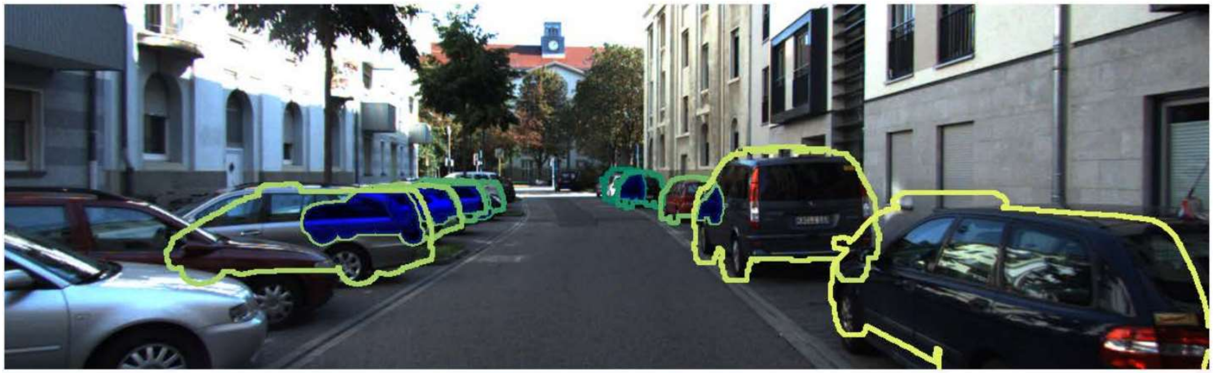
Slide Credit: Yu Xiang

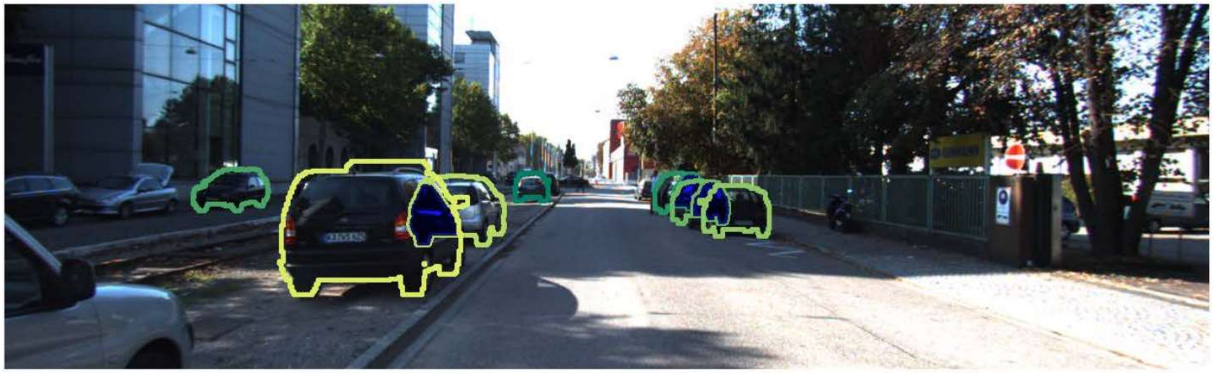
4. 3D Localization

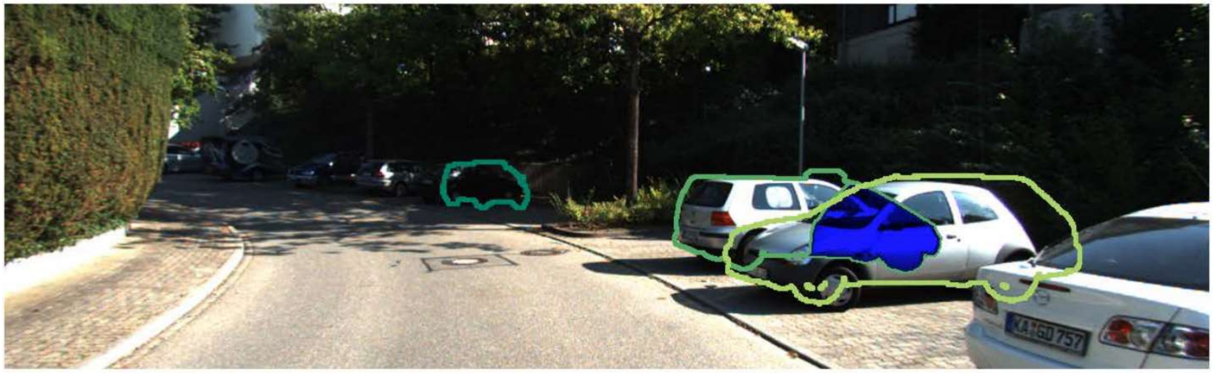


At this point we also **assign a single CAD model to each detection** based on the closest match from clustering**.**

Results

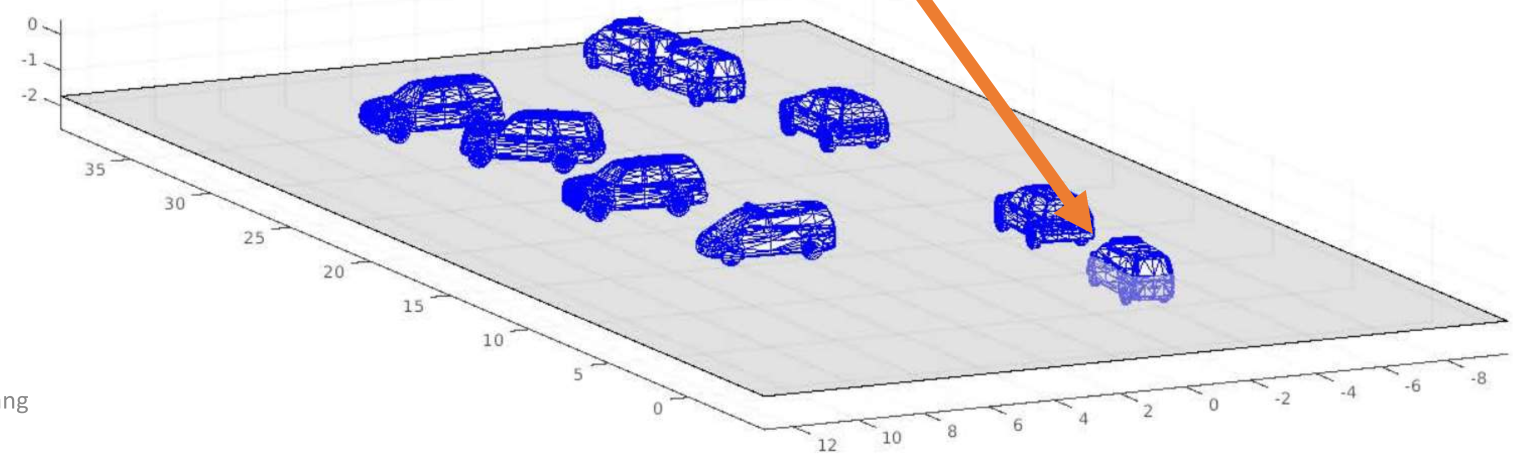






Slide Credit: Yu Xiang

This approach can generalize.



Car Detection and Orientation on KITTI

	Object Detection (AP)				Object Detection and Orientation estimation (AOS)		
Method	Easy	Moderate	Hard		Easy	Moderate	Hard
ACF [1]	55.89	54.77	42.98		N/A	N/A	N/A
DPM [2]	71.19	62.16	48.43		67.27	55.77	43.59
DPM-VOC+VP [3]	74.95	64.71	48.76		72.28	61.84	46.54
OC-DPM [4]	74.94	65.95	53.86		73.50	64.42	52.40
SubCat [5]	81.94	66.32	51.10		80.92	64.94	50.03
AOG [6]	84.36	71.88	59.27		43.81	38.21	31.53
SubCat [7]	84.14	75.46	59.71		83.41	74.42	58.83
Regionlets [8]	84.75	**76.45**	59.70		N/A	N/A	N/A
Ours NMS	84.81	73.02	63.22		84.31	71.99	62.11
Ours Occlusion	**87.46**	75.77	**65.38**		**86.92**	**74.59**	**64.11**

[1] P. Dolla´r, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. TPAMI, 2014.
[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.
[3] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Multi-view and 3d deformable part models. TPAMI, 2015.
[4] B. Pepikj, M. Stark, P. Gehler, and B. Schiele. Occlusion patterns for object class detection. In CVPR, 2013.
[5] E. Ohn-Bar and M. M. Trivedi. Fast and robust object detection using visual subcategories. In CVPRW, 2014.
[6] B. Li, T. Wu, and S.-C. Zhu. Integrating context and occlusion for car detection by hierarchical and-or model. In ECCV, 2014.
[7] E. Ohn-Bar and M. M. Trivedi. Learning to detect vehicles by clustering appearance patterns. T-ITS, 2015.
[8] X.Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In ICCV, 2013.

Slide Credit: Yu Xiang

Analysis

The Good

- State-of-the-art results
- Explicitly models occlusion
- Data driven approach
- Can perform segmentation

The Bad

- Complicated training pipeline
- Requires CAD models of objects
- Requires 3D data for training
- Little to no work on classification (sedans vs. vans)

?

Uses handcrafted features
How long does it take?

Discussion

- Will we ever not need 3D data for training?
- Does this work with deformable objects?
  - Does it need to?
- Possible extension: extend to more diverse classes
  - 227 different detectors just for “car”
  - What if we want to detect 20 different objects?
  - What kind of grouping would need to be done?

References

- Xiang, Yu, et al. "Data-driven 3d voxel patterns for object category recognition." *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.
- Dollár, Piotr, et al. "Fast feature pyramids for object detection." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.8 (2014): 1532-1545.
- Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.