3D Object Detection

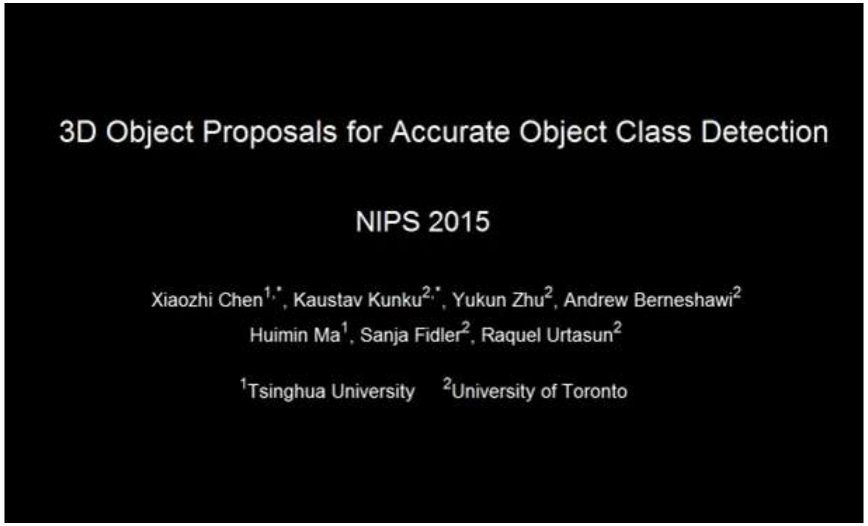
Zhen Li

CSC 2541 Presentation

Mar 8th, 2016

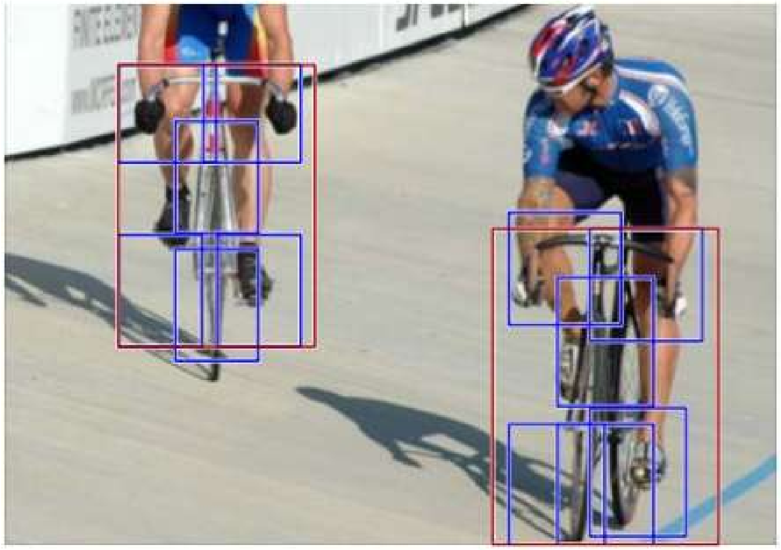
Object Detection: 2D vs 3D

[Video (Chen *et al*. 2015)]

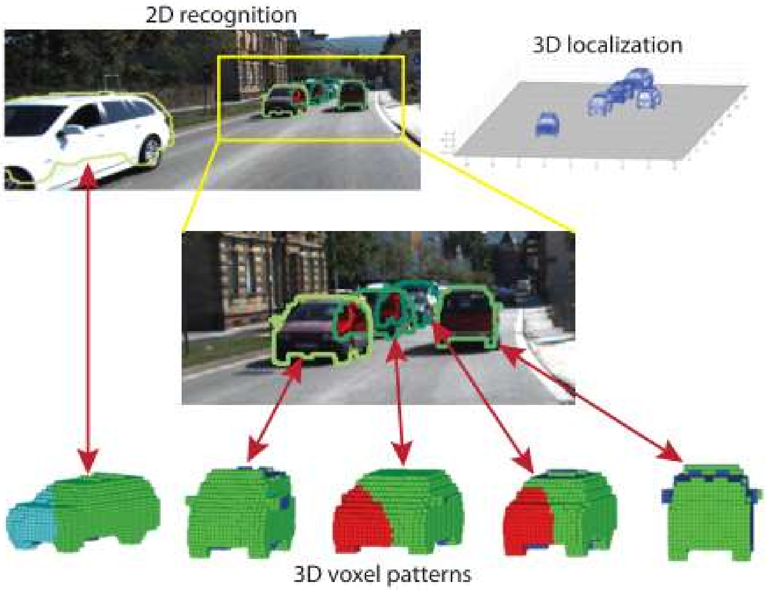


3D Object Detection: Motivation

- 2D bounding boxes are not sufficient
  - Lack of 3D pose, Occlusion information, and 3D location



(Figure from Felzenszwalb *et al*. 2010)



(Figure from Xiang *et al*. 2015)

3D Object Detection: Challenge

- Occlusion/Truncation: Only a small portion of the surface is visible
  - Leader board from KITTI website

		Easy	Moderate	Hard
1	SubCNN	90.49%	87.88%	77.10%
2	DJML	90.67%	87.51%	76.33%
3	3DOP	91.44%	86.10%	76.52%
4	Mono3D	88.31%	85.66%	75.89%
5	3DVP	86.92%	74.59%	64.11%

Easy: Max. occlusion 15%
Moderate: Max. occlusion 30%
Hard: Max. occlusion 50%

Outline

- Overview with contributions
- Main motivation
- Technical approach
- Experimental evaluation
- Discussion

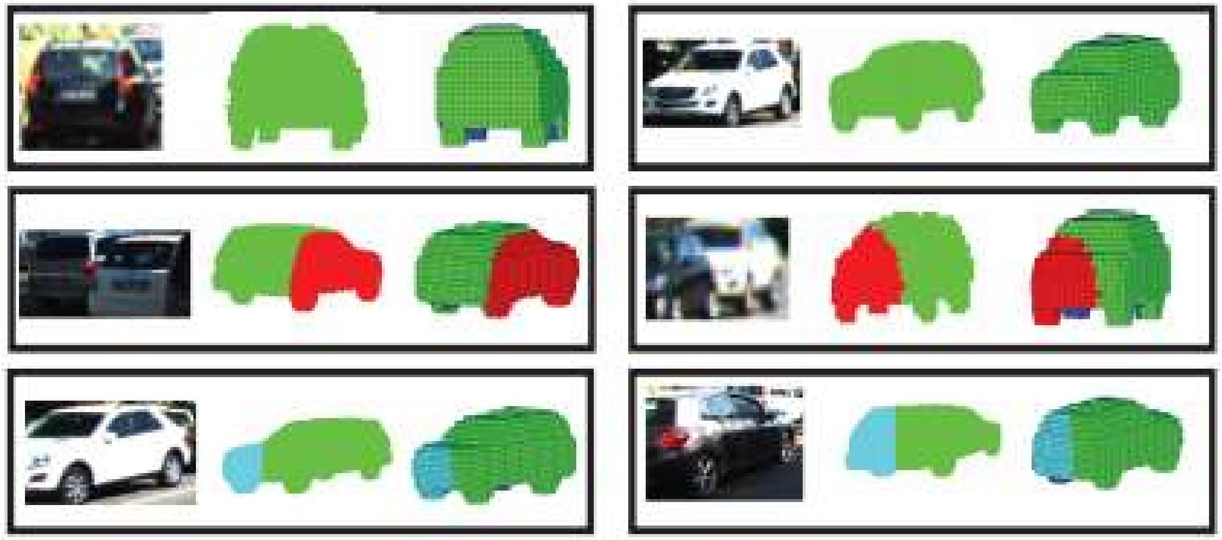
Paper #1

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.
yuxiang@umich.edu, {wongun, ylin}@nec-labs.com, ssilvio@stanford.edu

High-level Overview

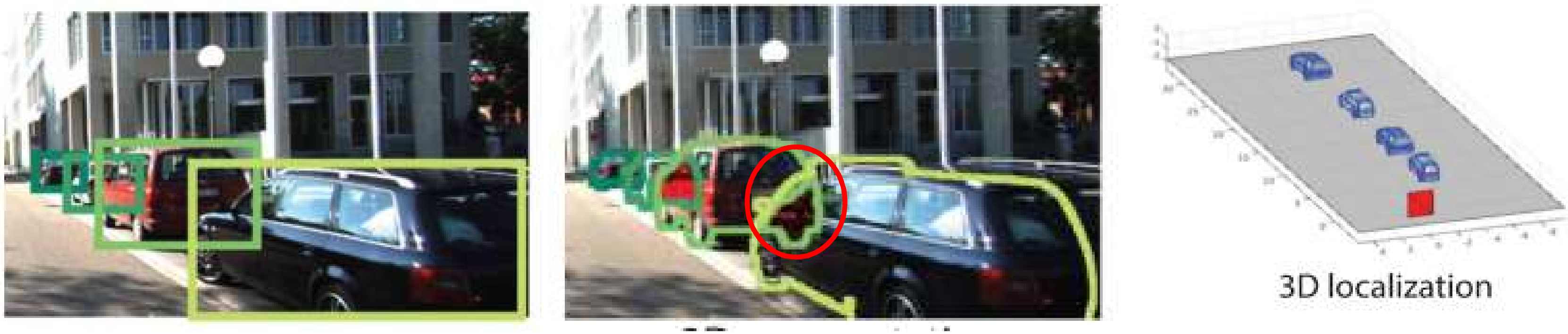
- Propose a novel object representation: 3D Voxel Pattern (3DVP)
  - Appearance, 3D shape, and occlusion masks



(Figure from Xiang *et al*. 2015)

High-level Overview

- Propose a novel object representation: 3D Voxel Pattern (3DVP)
  - Appearance, 3D shape, and occlusion masks
- Train specialized 3DVP detectors which are capable of:
  - 2D Object detection
  - Segmentation mask, occlusion or truncation boundaries
  - 3D localization, 3D pose



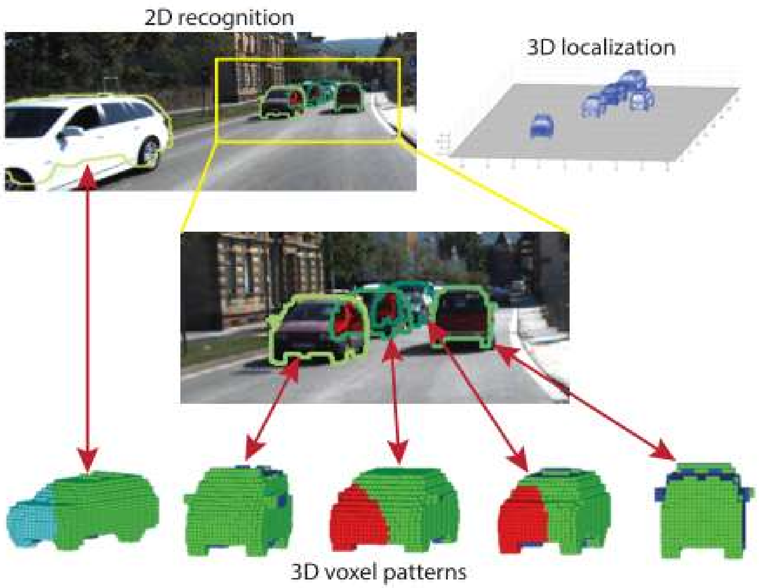
(Figure from Xiang *et al*. 2015)

High-level Overview

- Propose a novel object representation: 3D Voxel Pattern (3DVP)
  - Appearance, 3D shape, and occlusion masks
- Train specialized 3DVP detectors which are capable of:
  - 2D Object detection
  - Segmentation mask, occlusion or truncation boundaries
  - 3D localization, 3D pose
- Experiments on the KITTI benchmark and the OutdoorScene dataset
  - Improve the state-of-the-art results on detection and pose estimation with notable margins (6% in *difficult level* of KITTI)

Motivations

- What are the key challenges in this topic?
  - Occlusion/Truncation
    - Train partial object detectors for visible parts of objects (Wu and Nevatia 2005; Wojek *et al*. 2011; Xiang and Savarese 2013)



(Figure from Xiang *et al*. 2015)

Motivations

- What are the key challenges in this topic?
  - Occlusion/Truncation
  - Shape variation: Intra-class changes should be modeled
    - Discover and learn object sub-categories

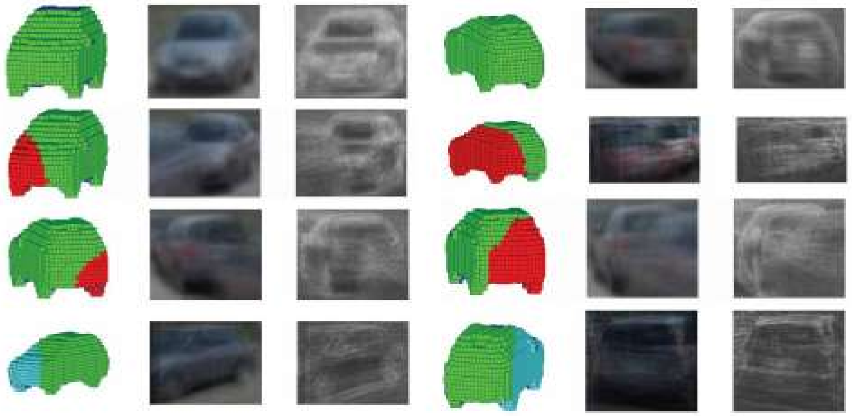
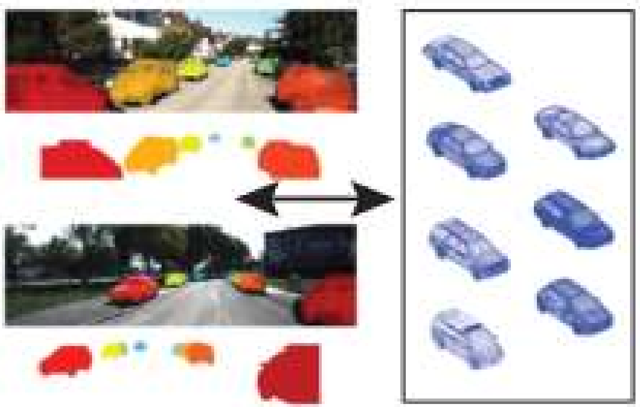


Figure 6. Visualization of selected 3DVPs. We show the 3D voxel model of the cluster center, the average RGB image, and the average gradient image of each 3DVP.

(Figure from Xiang *et al.* 2015)

Motivations

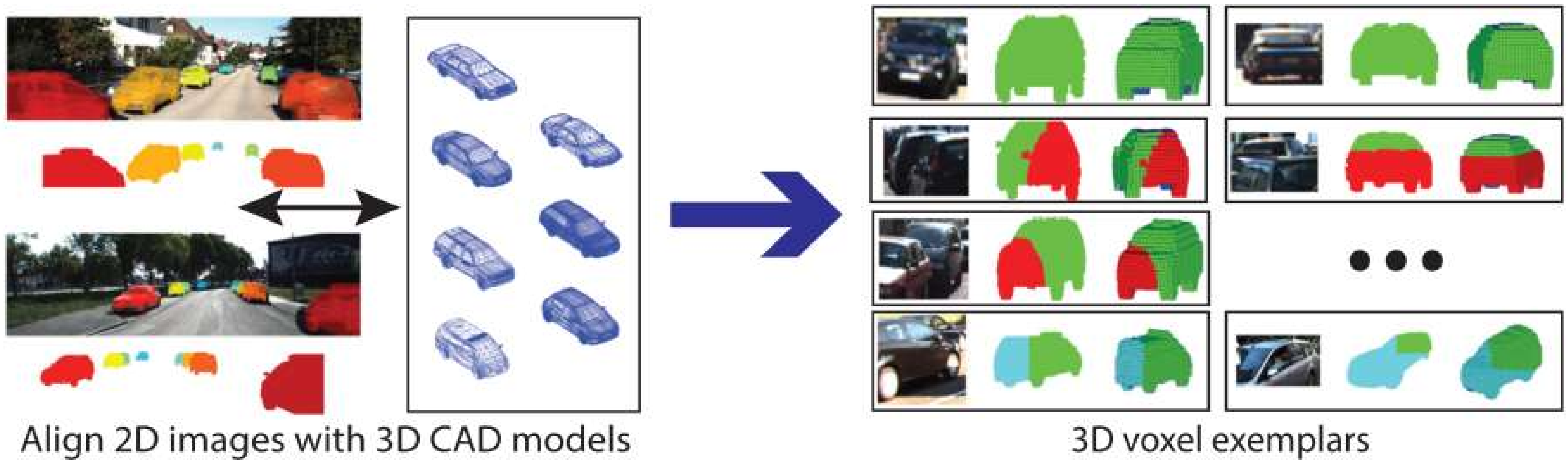
- What are the key challenges in this topic?
  - Occlusion/Truncation
  - Shape variation: Intra-class changes should be modeled
  - Viewpoint: Multiview object detection in 3D
    - Built from various 2D images (Yan *et al*. 2007; Glasner *et al.* 2011)
    - Constructed using CAD models (Liebelt *et al*. 2008)



(Figure from Xiang *et al.* 2015)

Technical approach

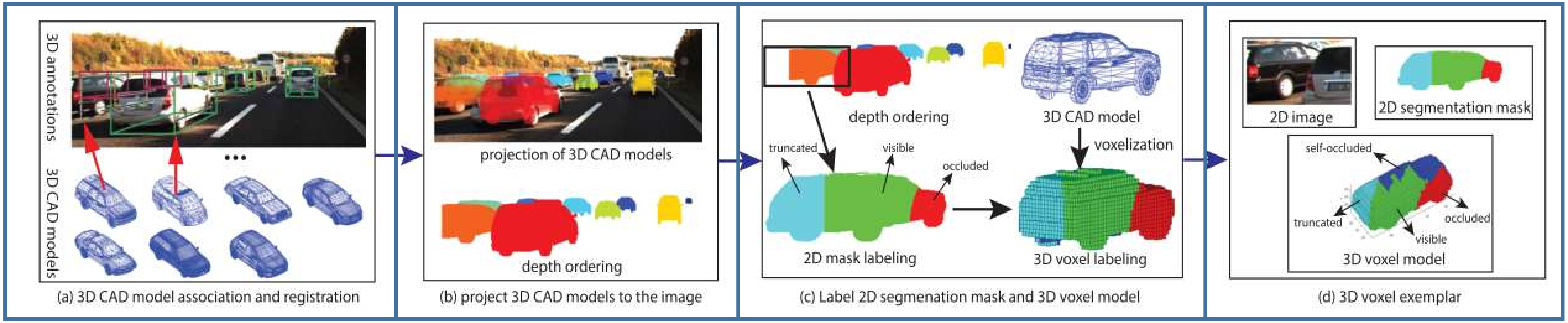
- Training: Generate *3D Voxel Exemplars*
  - A triplet of 2D image of the object, its 2D segmentation, and its 3D voxel model



(Figures from Xiang *et al.* 2015)

Technical approach

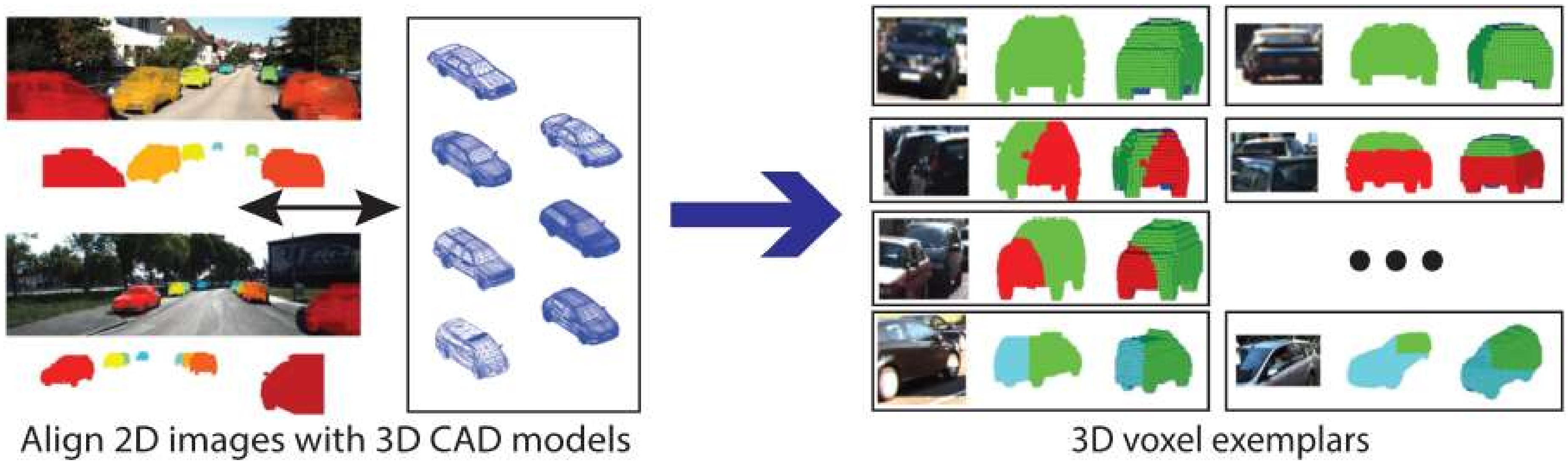
- Training: Generate *3D Voxel Exemplars*
  - 3D CAD model association and registration
  - Project 3D CAD models to the image
  - Label 2D segmentation mask and 3D voxel model
  - Generate a 3D voxel exemplar



(Figures from Xiang *et al.* 2015)

Technical approach

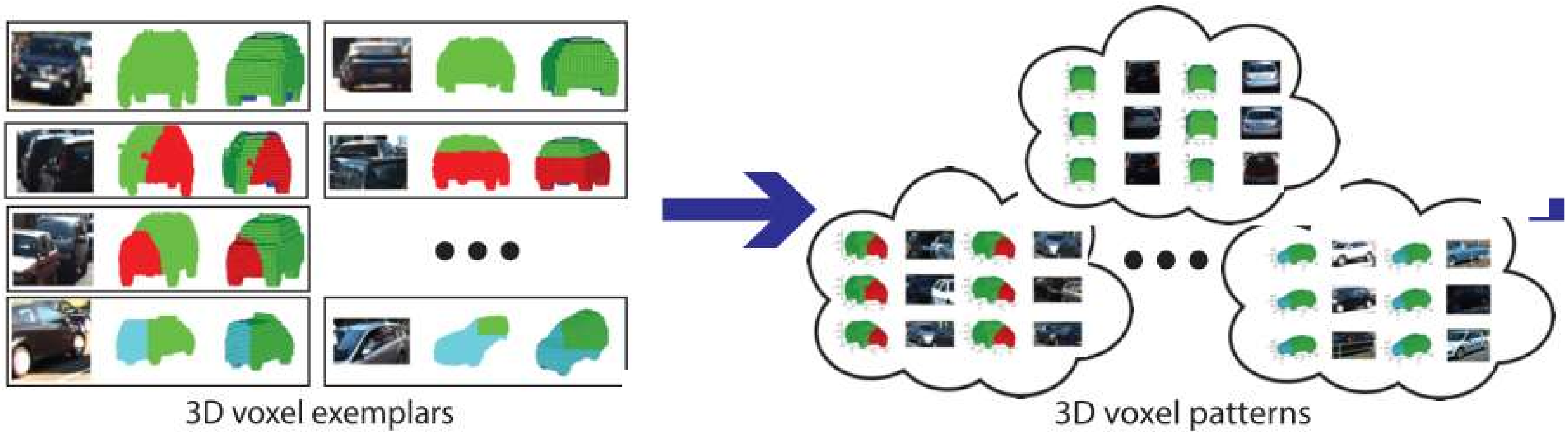
- Training: Generate *3D Voxel Exemplars*
  - A triplet of 2D image of the object, its 2D segmentation, and its 3D voxel model



(Figures from Xiang *et al.* 2015)

Technical approach

- Training: Build a representative set of 3DVPs



(Figures from Xiang *et al.* 2015)

Technical approach

- Training: Build a representative set of 3DVPs

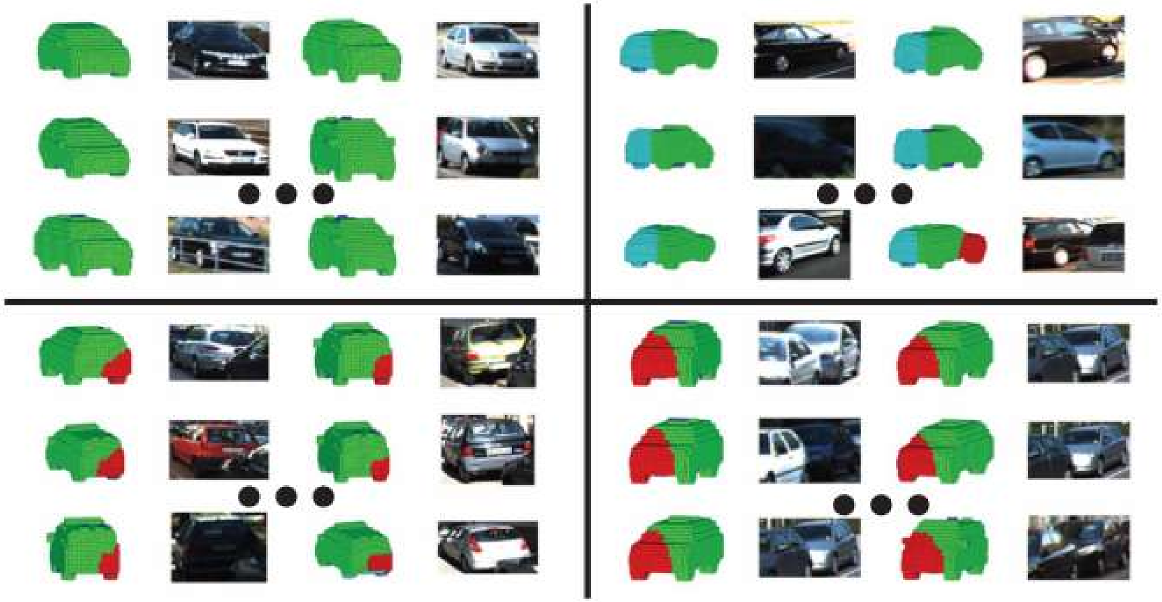


Figure 5. Examples of 3D clusters from the KITTI dataset.

(Figures from Xiang *et al.* 2015)

Technical approach

- Training: Build a representative set of 3DVPs
  - Define the 3D voxel exemplar feature vector $\mathbf{x}$ with dimension N^3
    - Encoding: 0 for empty voxels, 1 for visible voxels, 2 for self-occluded voxels, 3 for voxels occluded by other objects, and 4 for truncated voxels.
  - Define the similarity metric：

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

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

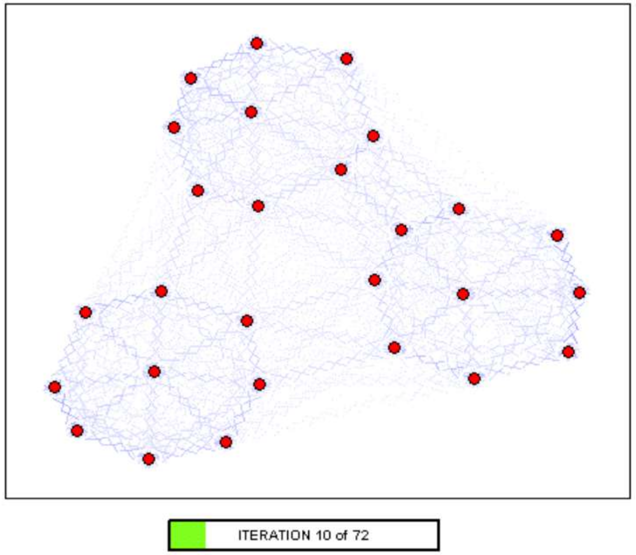
Technical approach

- Training: Build a representative set of 3DVPs
  - Define the 3D voxel exemplar feature vector $\mathbf{x}$ with dimension N^3
    - Encoding: 0 for empty voxels, 1 for visible voxels, 2 for self-occluded voxels, 3 for voxels occluded by other objects, and 4 for truncated voxels.
  - Define the similarity metric :

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

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

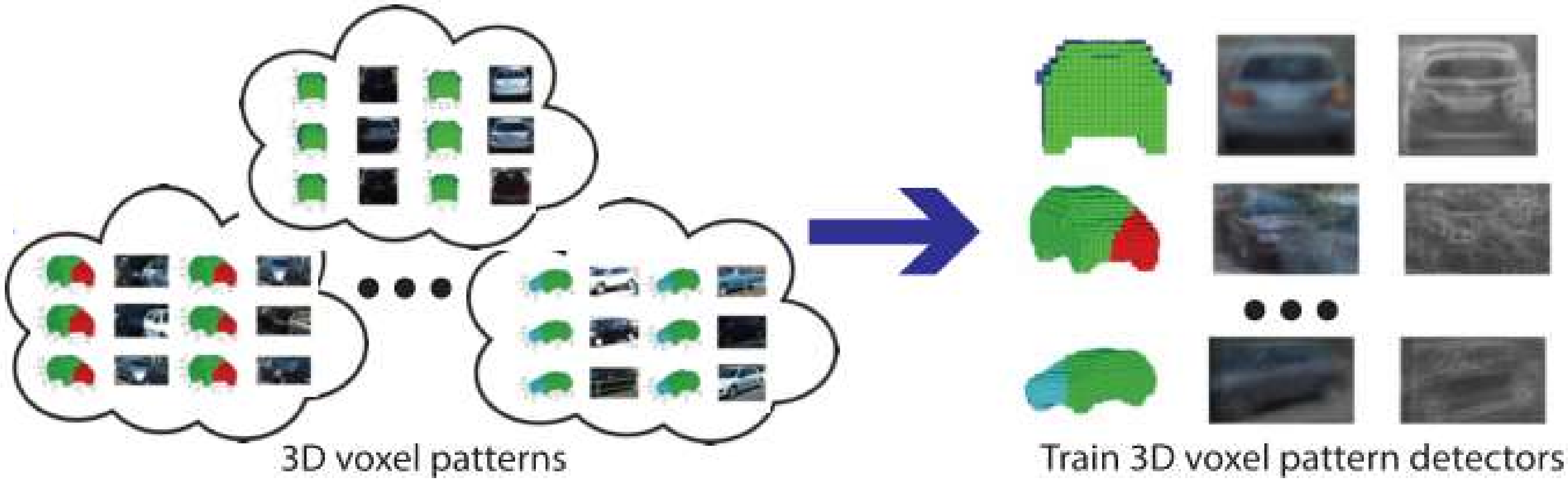
  - Employ clustering algorithms
    - K-means
    - [Affinity Propagation (AP) (Frey and Dueck 2007)]



(Video from http://www.psi.toronto.edu/affinitypropagation/)

Technical approach

- Training: Train 3DVP Detectors
  - SVM-based detectors for KITTI (Malisiewicz *et al.* 2011)
  - Boosting detector for KITTI
    - Aggregated Channel Features (ACF) (Dollár *et al.* 2014)



(Figures from Xiang *et al.* 2015)

Technical approach

- Training: Train 3DVP Detectors
  - SVM-based detectors for KITTI (Malisiewicz *et al.* 2011)
  - Boosting detector for KITTI
    - Aggregated Channel Features (ACF) (Dollár *et al.* 2014)

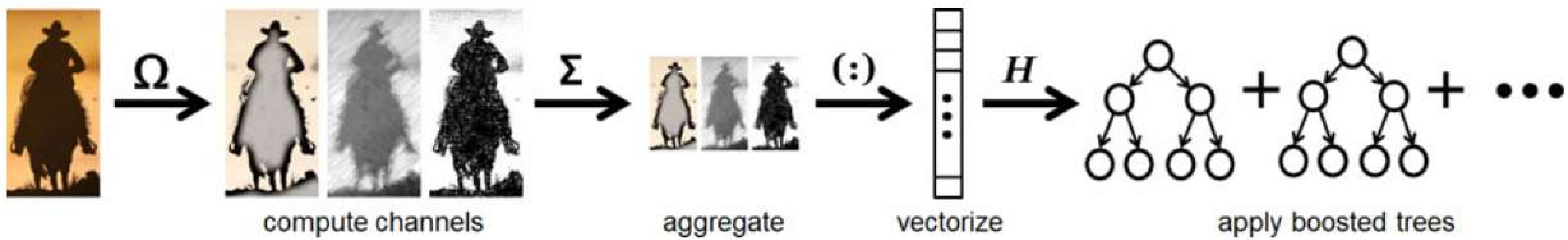
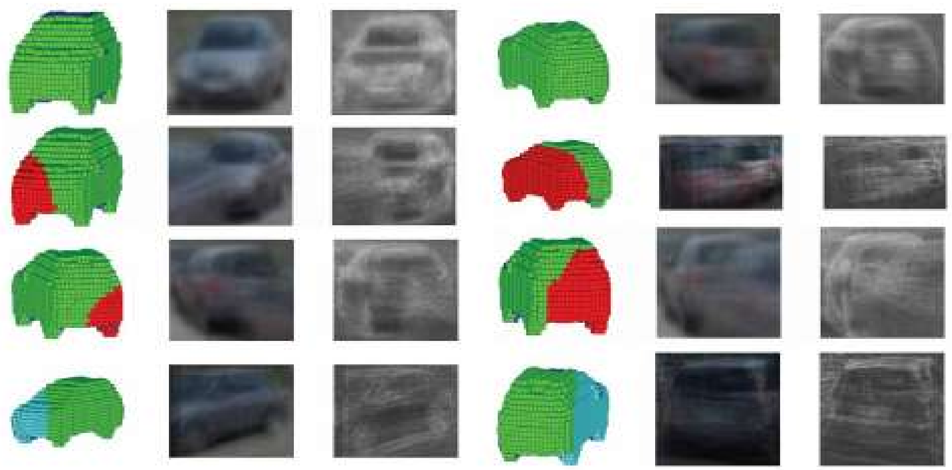


Fig. 8. Overview of the ACF detector. Given an input image I, we compute several channels $C = \Omega(I)$, sum every block of pixels in C, and smooth the resulting lower resolution channels. Features are single pixel lookups in the aggregated channels. Boosting is used to learn decision trees over these features (pixels) to distinguish object from background. With the appropriate choice of channels and careful attention to design, ACF achieves state-of-the-art performance in pedestrian detection.

(Images from Dollár *et al.* 2014)

Technical approach

- Training: Train 3DVP Detectors
  - SVM-based detectors for KITTI (Malisiewicz *et al.* 2011)
  - Boosting detector for KITTI
    - Aggregated Channel Features (ACF) (Dollár *et al.* 2014)
- A trick: Incorporate the appearance of the occluder



(Figures from Xiang *et al.* 2015)

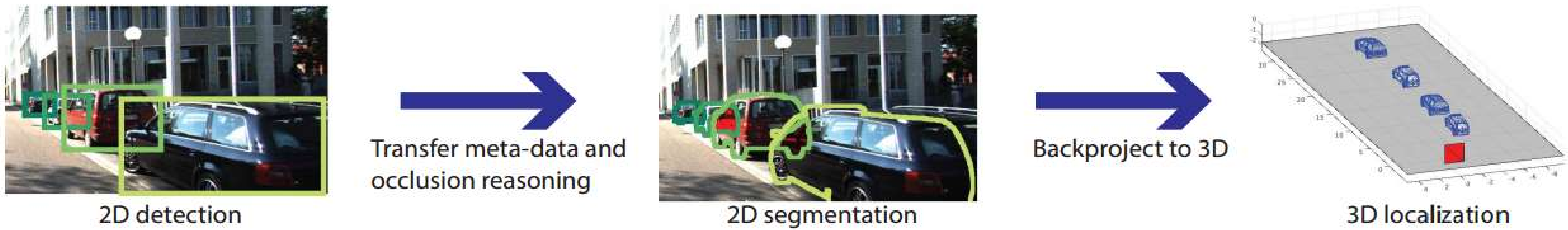
Technical approach

- Testing: Get 2D detection bounding boxes

(Figures from Xiang *et al.* 2015)

Technical approach

- Testing: Transfer the meta-data associated with the 3DVPs



(Figures from Xiang *et al.* 2015)

Technical approach

- Testing: Transfer the meta-data associated with the 3DVPs
  - Energy-based conditional random field model
    - $m_i = m_i^v + m_i^o + m_i^t$ (visible, occluded, and truncated)

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

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

Technical approach

- Testing: Transfer the meta-data associated with the 3DVPs
  - Energy-based conditional random field model
    - $m_i = m_i^v + m_i^o + m_i^t$ (visible, occluded, and truncated)

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

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

  - Implementation: Greedy algorithm

Technical approach

- Testing: Transfer the meta-data associated with the 3DVPs
  - Non –Maximum Suppression (NMS) (Felzenszwalb *et al.* 2010)
    - Sort the results, and pick the one with largest score
    - Computes the overlap between two bounding boxes by $\frac{|o_i \cap o_j|}{|o_i|}$
    - Greedily suppress detections that have larger than 0.5 overlap with selected ones
    - Noted by “NMS.5” in this paper
  - Intersection over Union (IoU) with 0.6 threshold
    - NMS-based, but keep more occluded detection hypotheses
    - Noted by “INMS.6” in this paper

Experimental evaluation

- Datasets
  - KITTI:
    - 7481 images (28,612 cars)
    - Split the training set into training set and validation set
  - OutdoorScene:
    - 200 images (focus on the presence of severe occlusions)
    - Only for testing

Experimental evaluation

- Evaluation metrics (threshold based metrics)
  - Object detection: Average Precision (AP) (Everingham *et al.* 2011)
  - Object orientation: Average Orientation Similarity (AOS) (Geiger *et al.* 2012)

$$AOS = \frac{1}{11} \sum_{r \in \{0, 0.1, .., 1\}} \max_{\tilde{r}:\tilde{r} \geq r} s(\tilde{r})$$

where $r = \frac{TP}{TP+FN}$ $\quad s(r) = \frac{1}{|\mathcal{D}(r)|} \sum_{i \in \mathcal{D}(r)} \frac{1 + \cos \Delta_\theta^{(i)}}{2} \delta_i \quad \in [0,1]$

Experimental evaluation

- Evaluation metrics (threshold based metrics)
  - Object detection: Average Precision (AP) (Everingham *et al.* 2011)
  - Object orientation: Average Orientation Similarity (AOS) (Geiger *et al.* 2012)

$$AOS = \frac{1}{11} \sum_{r \in \{0,0.1,..,1\}} \max_{\tilde{r}:\tilde{r} \geq r} s(\tilde{r})$$

where $r = \frac{TP}{TP+FN}$ $\quad s(r) = \boxed{\frac{1}{|\mathcal{D}(r)|}} \sum_{i \in \mathcal{D}(r)} \frac{1 + \boxed{\cos \Delta_\theta^{(i)}}}{2} \boxed{\delta_i}$ $\quad \in [0,1]$

Experimental evaluation

- Evaluation metrics (threshold based metrics)
  - Object detection: Average Precision (AP) (Everingham *et al.* 2011)
  - Object orientation: Average Orientation Similarity (AOS) (Geiger *et al.* 2012)

$$AOS = \frac{1}{11} \sum_{r \in \{0, 0.1, .., 1\}} \max_{\tilde{r}: \tilde{r} \geq r} s(\tilde{r})$$

where $r = \frac{TP}{TP+FN}$ $\quad s(r) = \frac{1}{|\mathcal{D}(r)|} \sum_{i \in \mathcal{D}(r)} \frac{1 + \cos \Delta_\theta^{(i)}}{2} \delta_i \quad \in [0,1]$

  - 2D segmentation: Average Segmentation Accuracy (ASA)
  - 3D localization: Average Localization Precision (ALP)

Experimental evaluation

- Result: 2D clustering vs 3D clustering

2D K-means				3D K-means				2D Affinity Propagation				3D Affinity Propagation			
K	Easy	Moderate	Hard	K	Easy	Moderate	Hard	K	Easy	Moderate	Hard	K	Easy	Moderate	Hard
5	44.21	31.23	25.42	5	41.78	31.63	28.06	137	**46.76**	**35.66**	**32.30**	87	74.28	62.54	52.87
10	47.78	38.13	32.26	10	52.55	39.44	32.76	156	46.12	34.44	30.35	125	78.28	**65.62**	**54.90**
20	61.24	48.04	40.27	20	61.52	49.33	42.07	189	44.97	34.88	31.53	135	78.13	65.44	54.79
30	**67.83**	51.68	43.63	30	63.29	49.46	41.55	227	39.66	31.67	29.62	152	77.96	64.45	53.93
40	66.49	**53.18**	**45.96**	40	69.46	56.13	47.26	273	36.52	28.51	27.08	180	79.02	65.55	54.72
50	66.65	51.90	43.28	50	70.76	58.77	50.30	335	27.96	22.74	22.22	229	79.94	64.87	53.53
100	58.45	46.15	39.34	100	75.73	61.06	51.29					284	79.91	64.04	53.10
150	56.74	43.84	37.75	150	77.15	63.25	53.13					333	**79.98**	63.95	52.99
200	53.57	41.26	33.61	200	78.00	**64.81**	**54.30**								
250	53.86	39.81	33.58	250	76.85	63.48	53.93								
300	48.81	35.53	29.10	300	**78.10**	62.11	51.99								
350	42.68	33.55	27.35	350	74.78	62.00	51.81								

Table 1. AP Comparison between 2D and 3D clustering with k-means and affinity propagation on our validation split. The table shows the average precision obtained by training ACF detectors in different settings.

(Table from Xiang *et al.* 2015)

Experimental evaluation

- Result: Occlusion(Energy-based) vs NMS.5 vs INMS.6
  - DPM: baselines (Felzenszwalb *et al.* 2010)

	Object Detection (AP)			Orientation (AOS)		
Methods	Easy	Moderate	Hard	Easy	Moderate	Hard
DPM [10] NMS.5	54.91	42.49	32.73	33.71	26.30	20.37
DPM [10] INMS.6	44.35	36.49	28.87	27.45	22.71	18.07
Ours NMS.5	79.06	64.72	50.38	77.65	62.75	48.57
Ours INMS.6	78.28	65.62	54.90	76.87	63.49	52.57
Ours Occlusion	**80.48**	**68.05**	**57.20**	**78.99**	**65.73**	**54.67**

Table 2. AP/AOS comparison between different detection/decoding methods on the validation set. We show the results of 3D AP with 125 clusters for **Ours**.

(Table from Xiang *et al.* 2015)

Experimental evaluation

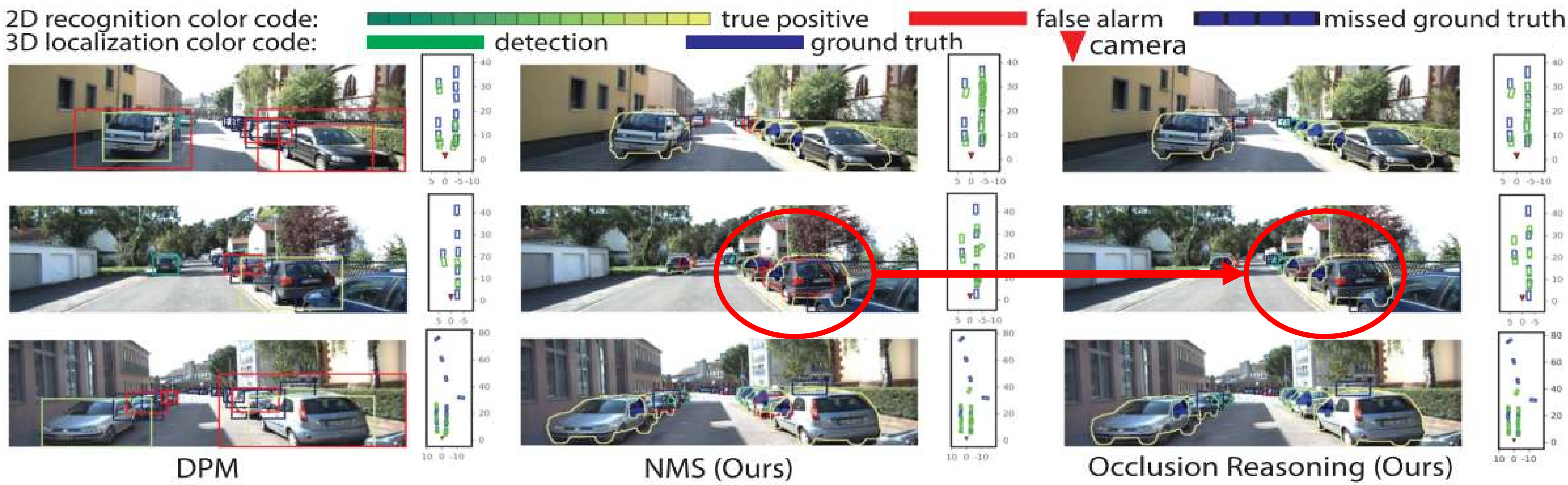
- Result: 2D segmentation
  - Lack of ground truth: projecting registered 3D CAD models

Method	Easy	Moderate	Hard
Joint 2D Detection and Segmentation (ASA)			
DPM [10]+box	38.09	29.42	22.65
Ours INMS.6+box	57.52	47.84	40.01
Ours Occlusion+box	59.21	49.74	41.71
Ours INMS.6+3DVP	63.88	52.57	43.82
Ours Occlusion+3DVP	**65.73**	**54.60**	**45.62**

(Table from Xiang *et al.* 2015)

Experimental evaluation

- Result: 2D segmentation
  - Qualitative result:



(Images from Xiang *et al.* 2015)

Experimental evaluation

- Result: 3D localization

Method	Easy	Moderate	Hard
Joint 2D Detection and 3D Localization (ALP)			
DPM [10] < 2m	40.21	29.02	22.36
Ours INMS.6 < 2m	64.85	49.97	41.14
Ours Occlusion < 2m	**66.56**	**51.52**	**42.39**
DPM [10] < 1m	24.44	18.04	14.13
Ours INMS.6 < 1m	44.47	33.25	26.93
Ours Occlusion < 1m	**45.61**	**34.28**	**27.72**

(Table from Xiang *et al.* 2015)

Experimental evaluation

- Result: 3D localization
  - Qualitative result:

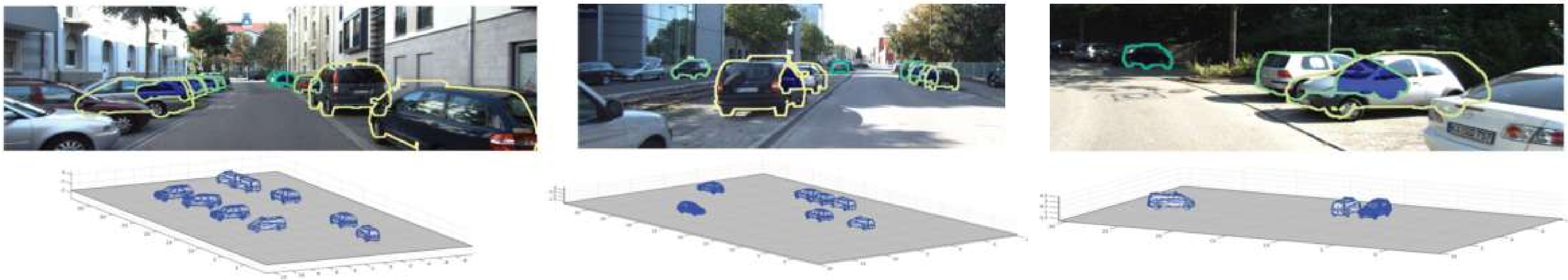
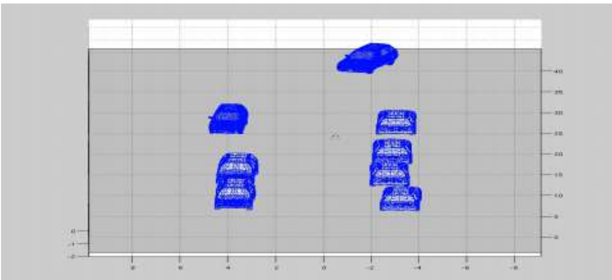


Figure 8. 2D recognition and 3D localization results on the KITTI test set. Blue regions in the images are the estimated occluded areas.



(Images and videos from Xiang *et al.* 2015)

Experimental evaluation

- Result: KITTI test set evaluation
  - Use the whole training set to generate the 3DVPs

	Object Detection (AP)			Orientation (AOS)		
Methods	Easy	Moderate	Hard	Easy	Moderate	Hard
ACF [8]	55.89	54.74	42.98	N/A	N/A	N/A
DPM [10]	71.19	62.16	48.43	67.27	55.77	43.59
DPM-VOC+VP [29]	74.95	64.71	48.76	72.28	61.84	46.54
OC-DPM [30]	74.94	65.95	53.86	73.50	64.42	52.40
SubCat [27]	81.94	66.32	51.10	80.92	64.94	50.03
AOG [24]	84.36	71.88	59.27	43.81	38.21	31.53
SubCat [28]	84.14	75.46	59.71	83.41	74.42	58.83
Regionlets [36]	84.75	**76.45**	59.70	N/A	N/A	N/A
Ours INMS.6	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**

Table 4. AP/AOS Comparison between different methods on the KITTI test set. We show the results of 3D AP with 227 clusters for **Ours**. More comparisons are available at [16].

(Table from Xiang *et al.* 2015)

Experimental evaluation

- Result: OutdoorScene dataset evaluation
  - 3DVP detectors are generalizable to other scenarios

% occlusion	< 0.3	$0.3 - 0.6$	> 0.6
# images	66	68	66
ALM [40]	72.3	42.9	35.5
DPM [10]	75.9	58.6	44.6
SLM [41]	80.2	63.3	52.9
Ours NMS.5	89.7	76.3	55.9
Ours Occlusion	**90.0**	**76.5**	**62.1**

Table 5. AP of the car detection on the OutdoorScene dataset [41].

(Table from Xiang *et al.* 2015)

Discussion

- Strength of the approach
  - Estimate detailed properties of objects beyond 2D bounding boxes

- Weakness of the approach
  - Running time: not mentioned in this paper
  - KITTI website

Discussion

	Method	Setting	Code	Moderate	Easy	Hard	Runtime	Environment	Compare
1	SubCNN			**87.88 %**	90.49 %	**77.10 %**	2 s	GPU @ 3.5 Ghz (Python + C/C++)	
Anonymous submission									
2	DJML			87.51 %	90.67 %	76.33 %	x s	GPU @ 1.5 Ghz (Matlab + C/C++)	
3	3DOP		code	86.10 %	**91.44 %**	76.52 %	3s	GPU @ 2.5 Ghz (Matlab + C/C++)	
X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler and R. Urtasun: 3D Object Proposals for Accurate Object Class Detection. NIPS 2015.									
4	Mono3D			85.66 %	88.31 %	75.89 %	x s	GPU @ 2.5 Ghz (Matlab + C/C++)	
X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler and R. Urtasun: Monocular 3D Object Detection for Autonomous Driving. CVPR 2016.									
5	3DVP			74.59 %	86.92 %	64.11 %	40 s	8 cores @ 3.5 Ghz (Matlab + C/C++)	
Y. Xiang, W. Choi, Y. Lin and S. Savarese: Data-Driven 3D Voxel Patterns for Object Category Recognition. IEEE Conference on Computer Vision and Pattern Recognition 2015.									
6	SubCat		code	74.42 %	83.41 %	58.83 %	0.7 s	6 cores @ 3.5 Ghz (Matlab + C/C++)	
E. Ohn-Bar and M. Trivedi: Learning to Detect Vehicles by Clustering Appearance Patterns. T-ITS 2015.									
7	SubCat+HSC			73.95 %	83.07 %	58.29 %	5.5 s	2 cores @ 2.5 Ghz (Matlab + C++)	
Anonymous submission									
8	SS			73.06 %	83.87 %	58.38 %	0.3 s	4 cores @ 2.5 Ghz (Matlab + C/C++)	
Anonymous submission									
9	SubCat			64.94 %	80.92 %	50.03 %	0.3 s	6 cores @ 2.5 Ghz (Matlab + C/C++)	
E. Ohn-Bar and M. Trivedi: Learning to Detect Vehicles by Clustering Appearance Patterns. T-ITS 2015. E. Ohn-Bar and M. Trivedi: Fast and Robust Object Detection Using Visual Subcategories. Computer Vision and Pattern Recognition Workshops Mobile Vision 2014.									
10	OC-DPM			64.42 %	73.50 %	52.40 %	10 s	8 cores @ 2.5 Ghz (Matlab)	
B. Pepik, M. Stark, P. Gehler and B. Schiele: Occlusion Patterns for Object Class Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013.									

(Screenshot from KITTI website: Geiger *et al.* 2012)

Discussion

- Strength of the approach
  - Estimate detailed properties of objects beyond 2D bounding boxes
- Weakness of the approach
  - Running time: not mentioned in this paper
  - KITTI website
- Future direction
  - Be able to adapt to different problems using different CAD models (*e.g.*, Cyclists, Pedestrians)

Paper #2

3D Object Proposals for Accurate Object Class Detection

Xiaozhi Chen[*,1] **Kaustav Kundu** [*,2] **Yukun Zhu**[2] **Andrew Berneshawi**[2]
Huimin Ma[1] **Sanja Fidler**[2] **Raquel Urtasun**[2]

[1]Department of Electronic Engineering
Tsinghua University

[2]Department of Computer Science
University of Toronto

High-level Overview

- Propose a new object proposal approach: 3D object proposals (3DOP)
  - In the context of autonomous driving
  - Exploits stereo imagery to place 3D bounding boxes
- Complete the full pipeline combing 3DOP and CNN

(Images from Chen *et al.* 2015)

High-level Overview

- Propose a new object proposal approach: 3D object proposals (3DOP)
  - In the context of autonomous driving
  - Exploits stereo imagery to place 3D bounding boxes
- Complete the full pipeline combing 3DOP and CNN
- Experiments on KITTI benchmark
  - Outperforms all existing approaches on all three categories (cars, cyclists, and pedestrians)

Motivation

- Why generating the proposal before object detection?
  - Proposals: at least a few accurately cover the ground-truth objects
  - Split the system into two phases:
    - i) generate the image proposals and ii) classify each proposal
  - Combine with other algorithm like R-CNN
    - Challenging conditions in autonomous driving

Motivation

- Why generating the proposal before object detection?
  - Proposals: at least a few accurately cover the ground-truth objects
  - Split the system into two phases:
    - i) generate the image proposals and ii) classify each proposal
  - Combine with other algorithm like R-CNN
    - Challenging conditions in autonomous driving
- Inspired by previous work
  - Selective Search (Van de Sande *et al.* 2011)
  - Contours-based method (Zitnick and Dollár 2014)

Motivation

- Challenges
  - High computational complexity of sliding windows
  - Produce perfect recall with fewer proposals
    - Trade-off between recall rate and precision rate
  - Exploit the stereo imagery to improve the performance

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x}, \mathbf{y}) = \mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x}, \mathbf{y}) + \mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x}, \mathbf{y}) + \mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x}, \mathbf{y}) + \mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x}, \mathbf{y})$$

  - $\mathbf{x}$**:** point cloud
  - $\mathbf{y}$**:** tuple (x, y, z, θ, c, t)
  - $\mathbf{w}_c^{\mathrm{T}}$**:** class-specific weights

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x},\mathbf{y}) = \boxed{\mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x},\mathbf{y})} + \mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x},\mathbf{y})$$

  - Point cloud density

$$\phi_{pcd}(\mathbf{x},\mathbf{y}) = \frac{\sum_{p\in\Omega(\mathbf{y})} S(p)}{|\Omega(\mathbf{y})|}$$

(Image from Chen *et al.* 2015)

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x},\mathbf{y}) = \mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x},\mathbf{y}) + \boxed{\mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x},\mathbf{y})} + \mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x},\mathbf{y})$$

  - Free space

$$\phi_{fs}(\mathbf{x},\mathbf{y}) = \frac{\sum_{p\in\Omega(\mathbf{y})}(1-F(p))}{|\Omega(\mathbf{y})|}$$

(Image from Chen *et al.* 2015)

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x},\mathbf{y}) = \mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x},\mathbf{y}) + \boxed{\mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x},\mathbf{y})} + \mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x},\mathbf{y})$$

  - Height prior

$$\phi_{ht}(\mathbf{x},\mathbf{y}) = \frac{1}{|\Omega(\mathbf{y})|}\sum_{p\in\Omega(\mathbf{y})} H_c(p)$$

with

$$H_c(p) = \begin{cases} \exp\left[-\frac{1}{2}\left(\frac{d_p-\mu_{c,ht}}{\sigma_{c,ht}}\right)^2\right], & \text{if } S(p)=1 \\ 0, & \text{o.w.} \end{cases}$$

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x},\mathbf{y}) = \mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x},\mathbf{y}) + \boxed{\mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x},\mathbf{y})}$$

  - Height contrast

$$\phi_{ht-contr}(\mathbf{x},\mathbf{y}) = \frac{\phi_{ht}(\mathbf{x},\mathbf{y})}{\phi_{ht}(\mathbf{x},\mathbf{y}^{+}) - \phi_{ht}(\mathbf{x},\mathbf{y})}$$

Technical approach

- Proposal Generation as Energy Minimization

$$\boxed{E(\mathbf{x},\mathbf{y}) = \mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x},\mathbf{y})}$$

  - Inference

$$\mathbf{y}^* = \mathrm{argmin}_{\mathbf{y}} E(\mathbf{x},\mathbf{y})$$

  - Get N diverse proposals
    - Sort the values of E(x,y) for all y
    - Greedy inference: pick top scoring proposal, perform NMS (Felzenszwalb *et al.* 2010), and iterate

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x},\mathbf{y}) = \mathbf{w}_{c,pcd}^{\top}\phi_{pcd}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,fs}^{\top}\phi_{fs}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht}^{\top}\phi_{ht}(\mathbf{x},\mathbf{y}) + \mathbf{w}_{c,ht-contr}^{\top}\phi_{ht-contr}(\mathbf{x},\mathbf{y})$$

  - Speed up tricks
    - Integral image (summed area table)
    - Skipping configurations which do not overlap with the point cloud
    - Place all our bounding boxes on the road plane
      - Sample additional proposal boxes at large locations: $y = y_{road} \pm \sigma_{road}$

Technical approach

- Proposal Generation as Energy Minimization

$$E(\mathbf{x},\mathbf{y}) = \underline{\mathbf{w}_{c,pcd}^{\top}}\phi_{pcd}(\mathbf{x},\mathbf{y}) + \underline{\mathbf{w}_{c,fs}^{\top}}\phi_{fs}(\mathbf{x},\mathbf{y}) + \underline{\mathbf{w}_{c,ht}^{\top}}\phi_{ht}(\mathbf{x},\mathbf{y}) + \underline{\mathbf{w}_{c,ht-contr}^{\top}}\phi_{ht-contr}(\mathbf{x},\mathbf{y})$$

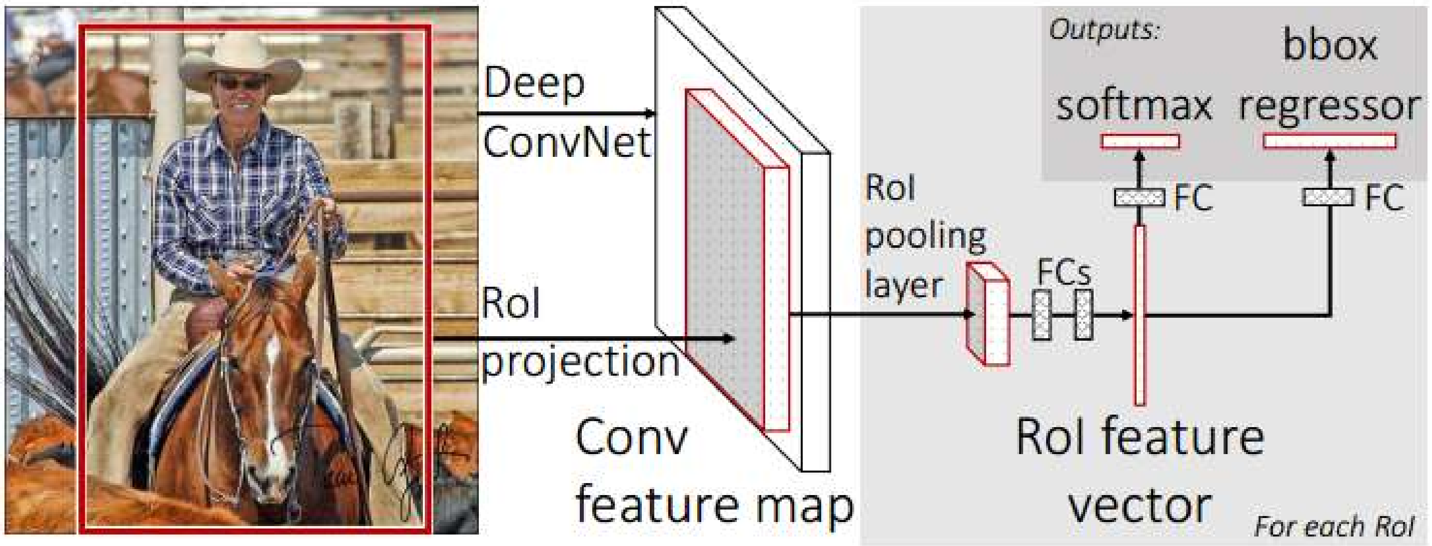
  - Learn the weights $\mathbf{w}_c^{\mathrm{T}}$ using structured SVM (Tsochantaridis *et al.* 2004)
    - Given N ground truth input-output pairs $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1,\ldots,N}$, solve the optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^D} \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.:}\quad \mathbf{w}^T(\phi(\mathbf{x}^{(i)},\mathbf{y}) - \phi(\mathbf{x}^{(i)},\mathbf{y}^{(i)})) \geq \boxed{\Delta(\mathbf{y}^{(i)},\mathbf{y})} - \xi_i, \quad \forall \mathbf{y} \setminus \mathbf{y}^{(i)}$$

Technical approach

- Object Detection and Orientation Estimation Network
  - 3DOP is combined with Fast R-CNN (Girshick 2015)



(Figure from Girshick 2015)

Technical approach

- Object Detection and Orientation Estimation Network
  - 3DOP is combined with Fast R-CNN (Girshick 2015)
  - A context branch after the last convolutional layer
    - Enlarging the candidate regions by a factor of 1.5 (Zhu *et al.* 2015)

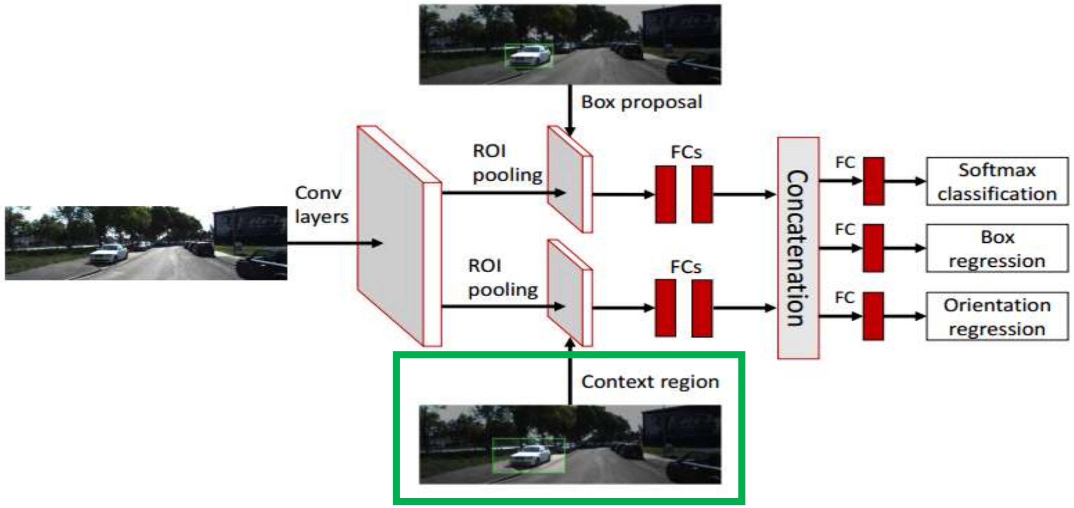


Figure 1: CNN architecture used to score our proposals for object detection.

(Figures from Chen *et al.* 2015)

Technical approach

- Object Detection and Orientation Estimation Network
  - 3DOP is combined with Fast R-CNN (Girshick 2015)
  - A context branch after the last convolutional layer
    - Enlarging the candidate regions by a factor of 1.5 (Zhu *et al.* 2015)
  - Orientation regression loss
    - Jointly learn object location and orientation
    - Smooth L_1 loss: Less sensitive to outliers than L2 loss used in R-CNN (Girshick *et al.* 2014) and SPPnet (He *et al.* 2015)

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{\text{x,y,w,h}\}} \text{smooth}_{L_1}(t_i^u - v_i),$$

in which

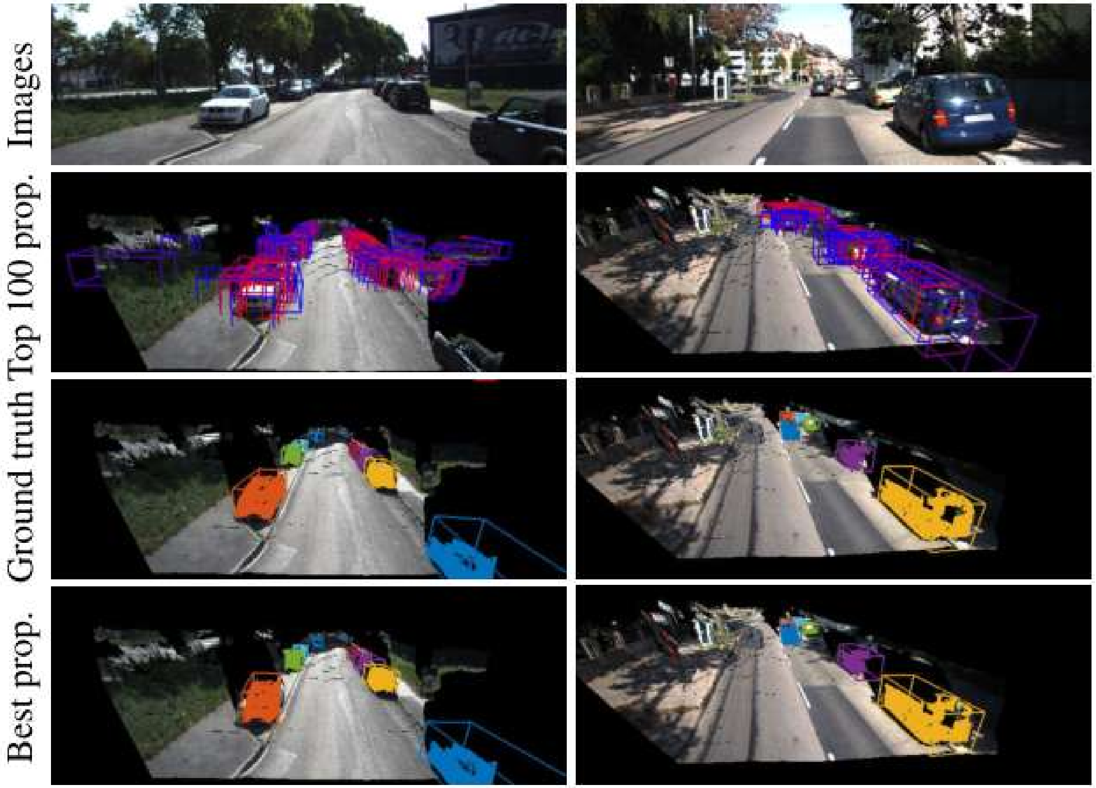
$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases}$$

Technical approach

- Object Detection and Orientation Estimation Network
  - 3DOP is combined with Fast R-CNN (Girshick 2015)
  - A context branch after the last convolutional layer
    - Enlarging the candidate regions by a factor of 1.5 (Zhu *et al.* 2015)
  - Orientation regression loss
    - Jointly learn object location and orientation
    - Smooth L_1 loss: Less sensitive to outliers than L2 loss used in R-CNN (Girshick *et al.* 2014) and SPPnet (He *et al.* 2015)
  - Initialization of weights on CNN
    - Use OxfordNet (Simonyan and Zisserman 2014) trained on ImageNet

Technical approach

- Object Detection and Orientation Estimation Network



(Figures from Chen *et al.* 2015)

Experimental evaluation

- Dataset: KITTI
  - 7481 training images, which contains three classes: Car, Pedestrian, and Cyclist
  - Three regimes based on the occlusion levels: Easy, Moderate, and Hard
  - Split the training set into training set and validation set
- Evaluation metric: Oracle recall (Van de Sande *et al.* 2011; Hosang *et al.* 2015)
  - For each ground truth (GT) object we found the proposal that overlaps the most in Intersection over Union (IoU)
  - Then we say it is recalled if IoU exceeds 70% for cars and 50% for pedestrians and cyclists

Experimental evaluation

- Results: Recall as a function of the number of candidates

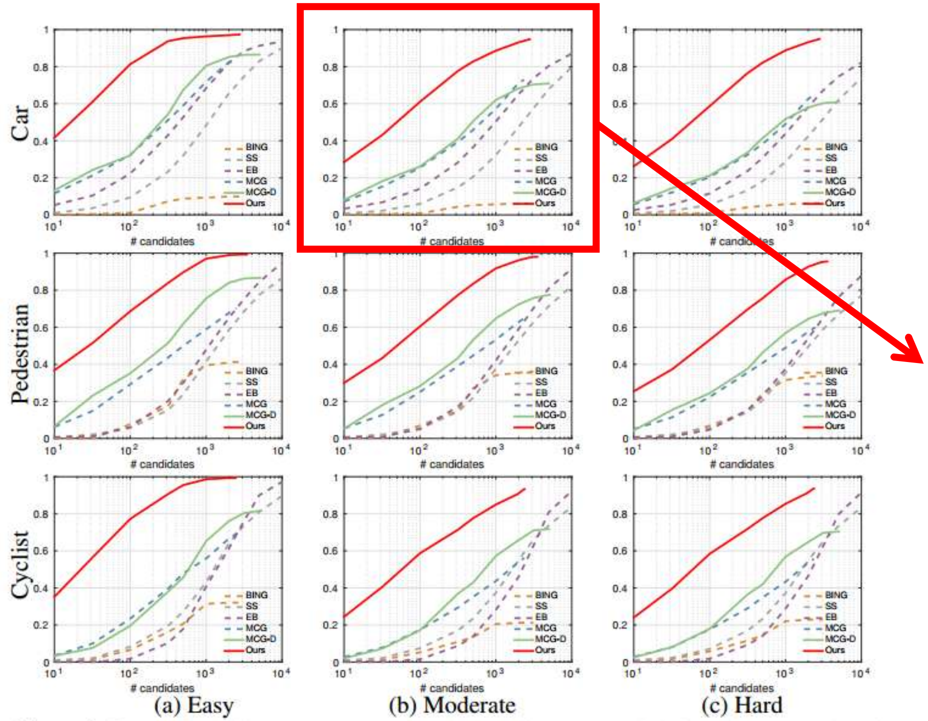
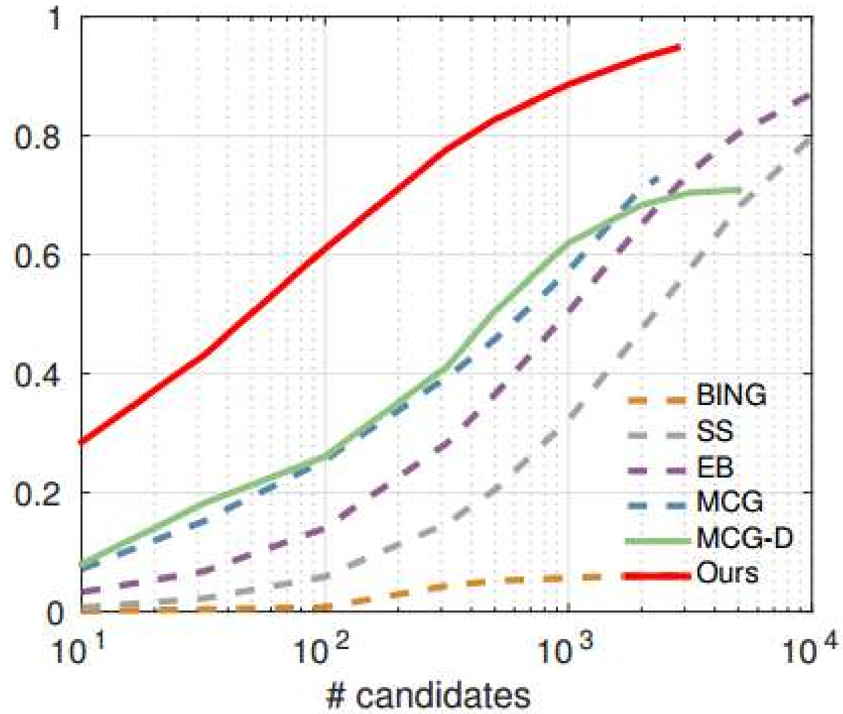


Figure 2: **Proposal recall:** We use 0.7 overlap threshold for *Car* , and 0.5 for *Pedestrian* and *Cyclist*.



(Figures from Chen *et al.* 2015)

Experimental evaluation

- Results: Recall as a function of the number of candidates

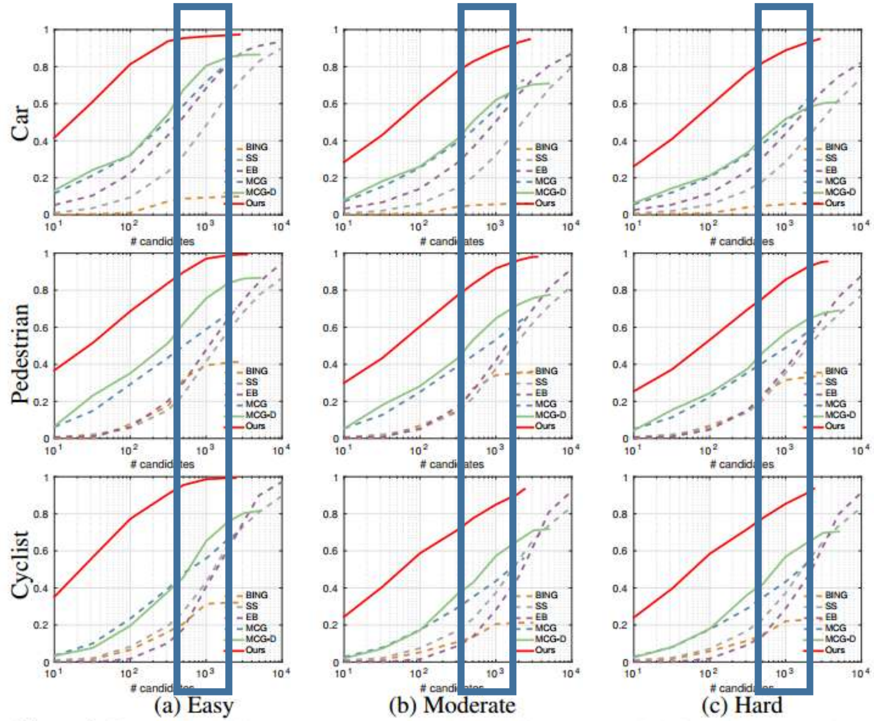
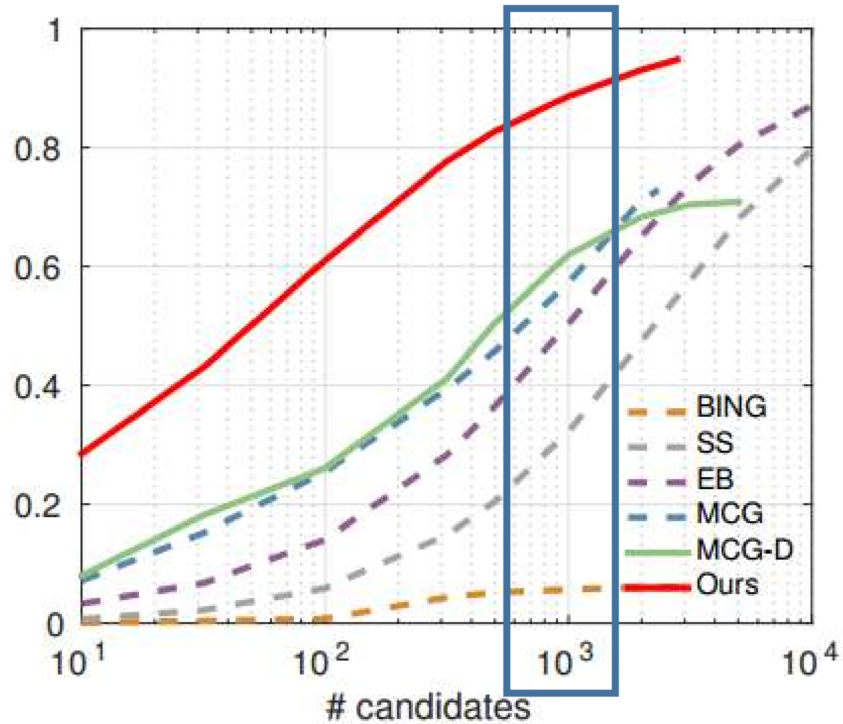


Figure 2: **Proposal recall:** We use 0.7 overlap threshold for *Car* , and 0.5 for *Pedestrian* and *Cyclist*.



(Figures from Chen *et al.* 2015)

Experimental evaluation

- Results: Recall vs IoU for 500 proposals

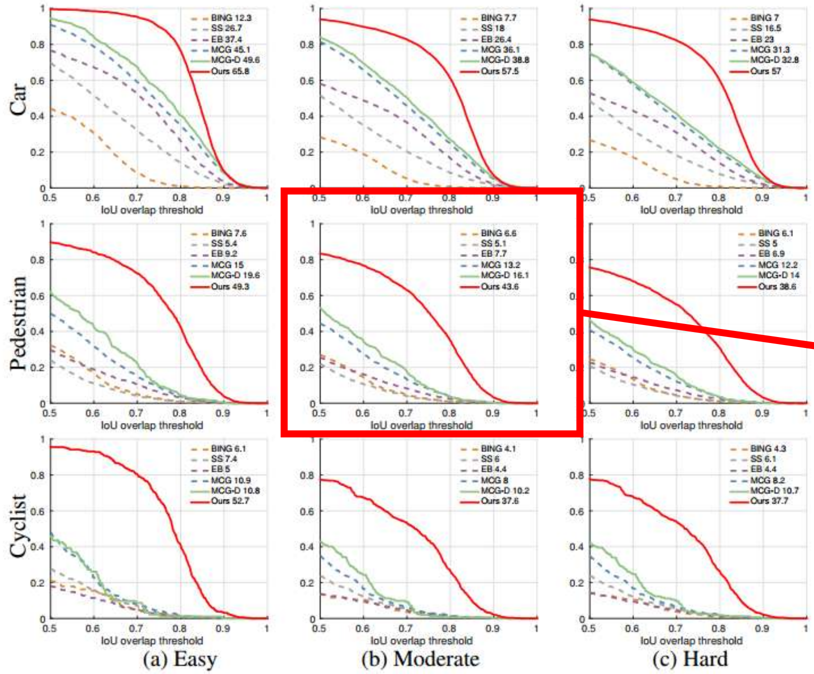
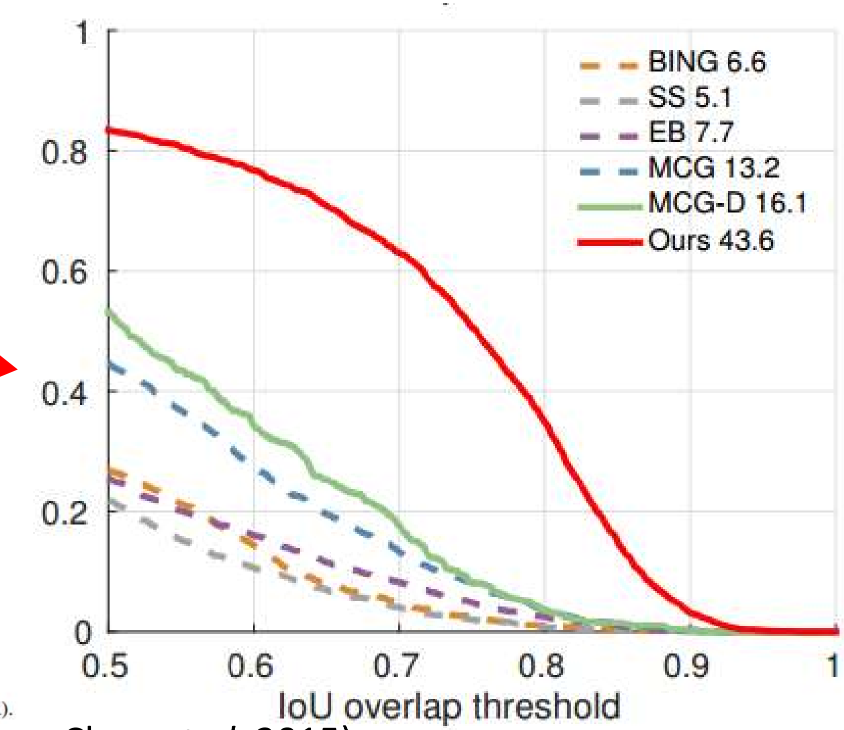


Figure 3: **Recall vs IoU for 500 proposals**. The number next to the labels indicates the average recall (AR).



(Figures from Chen *et al.* 2015)

Experimental evaluation

- Results: Running time

Method	BING	Selective Search	Edge Boxes (EB)	MCG	MCG-D	**Ours**
Time (seconds)	0.01	15	1.5	100	160	1.2

Table 3: Running time of different proposal methods.

(Table from Chen *et al.* 2015)

Experimental evaluation

- Results: Full object detection pipeline

	Cars			Pedestrians			Cyclists		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
LSVM-MDPM-sv [35, 1]	68.02	56.48	44.18	47.74	39.36	35.95	35.04	27.50	26.21
SquaresICF [36]	-	-	-	57.33	44.42	40.08	-	-	-
DPM-C8B1 [37]	74.33	60.99	47.16	38.96	29.03	25.61	43.49	29.04	26.20
MDPM-un-BB [1]	71.19	62.16	48.43	-	-	-	-	-	-
DPM-VOC+VP [27]	74.95	64.71	48.76	59.48	44.86	40.37	42.43	31.08	28.23
OC-DPM [38]	74.94	65.95	53.86	-	-	-	-	-	-
AOG [39]	84.36	71.88	59.27	-	-	-	-	-	-
SubCat [28]	84.14	75.46	59.71	54.67	42.34	37.95	-	-	-
DA-DPM [40]	-	-	-	56.36	45.51	41.08	-	-	-
Fusion-DPM [41]	-	-	-	59.51	46.67	42.05	-	-	-
R-CNN [42]	-	-	-	61.61	50.13	44.79	-	-	-
FilteredICF [43]	-	-	-	61.14	53.98	49.29	-	-	-
pAUCEnsT [44]	-	-	-	65.26	54.49	48.60	51.62	38.03	33.38
MV-RGBD-RF [45]	-	-	-	70.21	54.56	51.25	54.02	39.72	34.82
3DVP [12]	87.46	75.77	65.38	-	-	-	-	-	-
Regionlets [13]	84.75	76.45	59.70	73.14	61.15	55.21	70.41	58.72	51.83
Ours	**93.04**	**88.64**	**79.10**	**81.78**	**67.47**	**64.70**	**78.39**	**68.94**	**61.37**

Table 1: Average Precision (AP) (in %) on the test set of the KITTI Object Detection Benchmark.

(Table from Chen *et al.* 2015)

Experimental evaluation

- Results: Full object orientation estimation pipeline

	Cars			Pedestrians			Cyclists		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
AOG [39]	43.81	38.21	31.53	-	-	-	-	-	-
DPM-C8B1 [37]	59.51	50.32	39.22	31.08	23.37	20.72	27.25	19.25	17.95
LSVM-MDPM-sv [35, 1]	67.27	55.77	43.59	43.58	35.49	32.42	27.54	22.07	21.45
DPM-VOC+VP [27]	72.28	61.84	46.54	53.55	39.83	35.73	30.52	23.17	21.58
OC-DPM [38]	73.50	64.42	52.40	-	-	-	-	-	-
SubCat [28]	83.41	74.42	58.83	44.32	34.18	30.76	-	-	-
3DVP [12]	86.92	74.59	64.11	-	-	-	-	-	-
Ours	**91.44**	**86.10**	**76.52**	**72.94**	**59.80**	**57.03**	**70.13**	**58.68**	**52.35**

Table 2: AOS scores (in %) on the test set of KITTI's Object Detection and Orientation Estimation Benchmark.

(Table from Chen *et al.* 2015)

Experimental evaluation

• Results: Full object orientation estimation pipeline

3rd : 3DOP (this paper), Dec 2015, NIPS.

5th: 3DVP (previous paper), June 2015, CVPR.

	Method	Setting	Code	Moderate	Easy	Hard	Runtime	Environment	Compare
1	SubCNN			**87.88 %**	90.49 %	**77.10 %**	2 s	GPU @ 3.5 Ghz (Python + C/C++)	
Anonymous submission									
2	DJML			87.51 %	90.67 %	76.33 %	x s	GPU @ 1.5 Ghz (Matlab + C/C++)	
3	[illegible]	[illegible]	code	[illegible]	**91.44 %**	76.52 %	3s	GPU @ 2.5 Ghz (Matlab + C/C++)	
X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler and R. Urtasun: 3D Object Proposals for Accurate Object Class Detection. NIPS 2015.									
4	Mono3D			85.66 %	88.31 %	75.89 %	x s	GPU @ 2.5 Ghz (Matlab + C/C++)	
X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler and R. Urtasun: Monocular 3D Object Detection for Autonomous Driving. CVPR 2016.									
5	[illegible]			[illegible]	[illegible]	64.11 %	40 s	8 cores @ 3.5 Ghz (Matlab + C/C++)	
Y. Xiang, W. Choi, Y. Lin and S. Savarese: Data-Driven 3D Voxel Patterns for Object Category Recognition. IEEE Conference on Computer Vision and Pattern Recognition 2015.									
6	SubCat		code	74.42 %	83.41 %	58.83 %	0.7 s	6 cores @ 3.5 Ghz (Matlab + C/C++)	
E. Ohn-Bar and M. Trivedi: Learning to Detect Vehicles by Clustering Appearance Patterns. T-ITS 2015.									
7	SubCat+HSC			73.95 %	83.07 %	58.29 %	5.5 s	2 cores @ 2.5 Ghz (Matlab + C++)	
Anonymous submission									
8	SS			73.06 %	83.87 %	58.38 %	0.3 s	4 cores @ 2.5 Ghz (Matlab + C/C++)	
Anonymous submission									
9	SubCat			64.94 %	80.92 %	50.03 %	0.3 s	6 cores @ 2.5 Ghz (Matlab + C/C++)	
E. Ohn-Bar and M. Trivedi: Learning to Detect Vehicles by Clustering Appearance Patterns. T-ITS 2015. E. Ohn-Bar and M. Trivedi: Fast and Robust Object Detection Using Visual Subcategories. Computer Vision and Pattern Recognition Workshops Mobile Vision 2014.									
10	OC-DPM			64.42 %	73.50 %	52.40 %	10 s	8 cores @ 2.5 Ghz (Matlab)	
B. Pepik, M. Stark, P. Gehler and B. Schiele: Occlusion Patterns for Object Class Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013.									

(Screenshot from KITTI website: Geiger *et al.* 2012)

Discussion

- Strength of the approach
  - Generating proposals
    - 3DOP achieves higher recall rate on challenging KITTI benchmark
  - Full object detection/orientation estimation pipeline
    - 3DOP + Fast R-CNN outperforms state-of-the-art methods on KITTI testing set

- Weakness of the approach
  - Rely on stereo images
  - Still not a real-time algorithm (1.2 seconds for proposals, 3 seconds for full pipeline)

- Future work
  - Implement monocular 3D Object Detection
  - Improve efficiency by reducing spurious false positives

Reference

Chen, X., Kundu, K., Zhu, Y., Berneshawi, A. G., Ma, H., Fidler, S., & Urtasun, R. (2015). 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems* (pp. 424-432).

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *36*(8), 1532-1545.

Everingham, M., Van Gool, L., Williams, C., Winn, J., & Zisserman, A. (2011). The pascal visual object classes challenge 2012.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(9), 1627-1645.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, *315*(5814), 972-976.

Geiger, A., Lenz, P., & Urtasun, R. (2012, June). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3354-3361). IEEE.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In*Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

Glasner, D., Galun, M., Alpert, S., Basri, R., & Shakhnarovich, G. (2011, November). Viewpoint-aware object detection and pose estimation. In*Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 1275-1282). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *37*(9), 1904-1916.

Hosang, J., Benenson, R., Dollár, P., & Schiele, B. (2015). What makes for effective detection proposals?.

Liebelt, J., Schmid, C., & Schertler, K. (2008, June). Viewpoint-independent object class detection using 3d feature maps. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-8). IEEE.

Reference

Malisiewicz, T., Gupta, A., & Efros, A. A. (2011, November). Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 89-96). IEEE.

Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. (2004, July). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning* (p. 104). ACM.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Van de Sande, K. E., Uijlings, J. R., Gevers, T., & Smeulders, A. W. (2011, November). Segmentation as selective search for object recognition. In*Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 1879-1886). IEEE.

Wojek, C., Walk, S., Roth, S., & Schiele, B. (2011, June). Monocular 3D scene understanding with explicit occlusion reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1993-2000). IEEE.

Wu, B., & Nevatia, R. (2005, October). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (Vol. 1, pp. 90-97). IEEE.

Xiang, Y., Choi, W., Lin, Y., & Savarese, S. (2015, June). Data-driven 3d voxel patterns for object category recognition. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 1903-1911). IEEE.

Xiang, Y., & Savarese, S. (2013). Object detection by 3d aspectlets and occlusion reasoning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 530-537).

Yan, P., Khan, S. M., & Shah, M. (2007, October). 3d model based object class detection in an arbitrary view. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (pp. 1-6). IEEE.

Zhu, Y., Urtasun, R., Salakhutdinov, R., & Fidler, S. (2015). segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4703-4711).

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014* (pp. 391-405). Springer International Publishing.

Thank you!