3D Object Representations for Recognition

Yu Xiang
Computational Vision and Geometry Lab
Stanford University
2D Object Recognition

Image classification/tagging/annotation

- Ordonez et al. ICCV13
- horse, pasture, field, cow, fence

Object detection

- Ren et al. NIPS15

Object segmentation

- Long et al. CVPR15

Image description generation

- Karpathy et al. CVPR15
- "man in black shirt is playing guitar."
Applications of 2D Object Recognition

- Image search/indexing
- Photo editing
- Visual surveillance
- Biometrics authentication
These are all great, but...

2D recognition is NOT enough!
Applications that need 3D Object Recognition

- Autonomous Driving
- Robotics
- Augmented Reality
- Gaming

Any application that requires interaction with the 3D world!
Goal: Infer the 3D World

- Interaction
- Control
- Decision making
- Navigation
- Etc.

A 2D image

The 3D world
Goal: Infer the 3D World

1965
Blocks World
Larry Roberts, 1965

2010
Blocks World Revisited
Gupta et al., ECCV’10

Surface Layout/Normal Estimation
Hoiem et al., ICCV’05
Fouhey et al. ICCV’13, ECCV’14

Room Layout Estimation
Lee et al. CVPR’09
Hedau, el al., ICCV’09
Mallya & Lazebnik, ICCV’15

3D Object Recognition
Kar et al., ICCV’15
Tulsiani & Malik, CVPR’15

Marr’s Theory
David Marr, 1978

2-D Primal sketch
edges contours blobs

2.5-D Sketch
depth & orientation

3-D model representation
real shape

3-D Object Recognition
Kar et al., ICCV’15
Tulsiani & Malik, CVPR’15
My Work: 2D Object Detection

The image is from the KITTI detection benchmark (Geiger et al. CVPR’12)
My Work: 2D Object Detection
My Work: 2D Segmentation and 3D Pose Estimation
My Work: Occlusion Reasoning
My Work: 3D Localization
Contribution: 3D Object Representations

3D Object Representation

A 2D image

The 3D world
Related Work: 2D Object Representations

- Deformable part model
  Felzenszwalb et al., TPAMI’10

- Viola & Jones, IJCV’01
- Fergus et al., CVPR’03
- Leibe et al., ECCVW’04
- Hoiem et al., CVPR’06
- Vedaldi et al., ICCV’09
- Maji & Malik, CVPR’09
- Felzenszwalb et al., TPAMI’10
- Malisiewicz et al., ICCV’11
- Divvala et al., ECCVW’12
- Dolla´r et al., TPAMI’14
  Etc.

- ✓ 2D detection
- ✗ 3D pose
- ✗ Occlusion
- ✗ 3D location
Related Work: 2.5D Object Representations

- Thomas et al., CVPR’06
- Savarese & Fei-Fei ICCV’07
- Kushal et al., CVPR’07

- Su et al., ICCV’09
- Sun et al., CVPR’10
- Etc.

- 2D detection
- 3D pose
- Occlusion
- 3D location
Related Work: 3D Object Representations

- 2D detection
- 3D pose
- Occlusion
- 3D location

3DDPM
Pepik et al., CVPR’12

- Yan et al., ICCV’07
- Hoiem et al., CVPR’07
- Liebelt et al., CVPR’08, 10
- Glasner et al. ICCV’11

- Pepik et al., CVPR’12
- Xiang & Savarese, CVPR’12
- Hejrati & Ramanan, NIPS’12
- Fidler et al., NIPS’12
- Etc
Contribution: 3D Object Representations

- ✓ 2D detection
- ✓ 3D pose
- ✓ Occlusion
- ✓ 3D location
Outline

• 3D Aspect Part Representation

• 3D Voxel Pattern Representation

• A Benchmark for 3D Object Recognition in the Wild

• Conclusion and Future Work
Outline

• 3D Aspect Part Representation

• 3D Voxel Pattern Representation

• A Benchmark for 3D Object Recognition in the Wild

• Conclusion and Future Work
3D Aspect Part Representation

Viewpoint Variation
3D Aspect Part Representation

Viewpoint: Azimuth 315°, Elevation 30°, Distance 2

3D Aspect Parts from 3D CAD Models

Mean Shape
3D Aspect Part Representation

Bicycle | Car | Cellphone | Iron | Mouse | Shoe
Stapler | Toaster | Bed | Chair | Sofa | Table
Aspect Layout Model

An input image

3D aspect part representation

Viewpoint: Azimuth 315°, Elevation 30°, Distance 2

Output
Aspect Layout Model

• Posterior distribution

\[ P(Y, L, O, V | I) \propto \exp(E(Y, L, O, V, I)) \]

\[ L = (l_1, \ldots, l_n), \quad l_i = (x_i, y_i) \]
Aspect Layout Model

• Energy function

\[ E(Y, L, O, V, I) = \begin{cases} 
\sum_i V_1 (l_i, O, V, I) + \sum_{(i,j)} V_2 (l_i, l_j, O, V), & \text{if } Y = +1 \\
0, & \text{if } Y = -1 
\end{cases} \]

unary potential

pairwise potential
Aspect Layout Model

• Unary potential

\[ V_1(l_i, O, V, I) = \begin{cases} w_i^T \phi(l_i, O, V, I), & \text{if unoccluded} \\ \alpha_i, & \text{if self-occluded} \end{cases} \]
Aspect Layout Model

\[ V_1(l_i, O, V, I) = \begin{cases} w_i^T \phi(l_i, O, V, I), & \text{if unoccluded} \\ \alpha_i, & \text{if occluded} \end{cases} \]
Aspect Layout Model

- Pairwise potential

\[ V_2(\mathbf{l}_i, \mathbf{l}_j, O, V) = -w_x (x_i - x_j + d_{ij,o,v} \cos(\theta_{ij,o,v}))^2 - w_y (y_i - y_j + d_{ij,o,v} \sin(\theta_{ij,o,v}))^2 \]
Aspect Layout Model

• Training with Structural SVM [1]

$$\min_{\theta} \frac{1}{2} \|\theta\|^2 + \lambda \sum_{t=1}^{N} \left[ \max_{Y_{t},L_{t},O_{t},V_{t}} \left[ \theta^T \Psi_{t,Y_{t},L_{t},O_{t},V_{t}} + \Delta_{t,Y_{t},L_{t},O_{t},V_{t}} \right] - \theta^T \Psi_{t,Y^{*}_{t},L^{*}_{t},O^{*}_{t},V^{*}_{t}} \right]$$

• Inference  $$(Y^{*}, L^{*}, O^{*}, V^{*}) = \arg \max_{Y,L,O,V} E(Y,L,O,V,I|\theta)$$
  • Loop over discretized viewpoints
  • Run Belief Propagation [2] under each viewpoint to predict part locations


### Aspect Layout Model

- **Best results upon publication in pose estimation and 3D part estimation**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Viewpoint (cars)</strong></td>
<td>Ours</td>
<td><strong>93.4%</strong></td>
<td>85.4</td>
<td>85.3</td>
<td>81</td>
<td>70</td>
<td>67</td>
<td>48.5</td>
</tr>
<tr>
<td><strong>Viewpoint (cars)</strong></td>
<td>Ours - baseline</td>
<td>58.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Viewpoint</strong></td>
<td>Ours</td>
<td><strong>63.4%</strong></td>
<td>34.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Method</th>
<th>Ours</th>
<th>Ours - baseline</th>
<th>DPM [7]</th>
<th>[8]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Viewpoint (cars)</strong></td>
<td>Ours</td>
<td>56.6</td>
<td></td>
<td>41.6</td>
<td></td>
</tr>
</tbody>
</table>

---

Aspect Layout Model

Prediction: a=225, e=30, d=7

Prediction: a=330, e=15, d=7

Prediction: a=150, e=15, d=7

Prediction: a=300, e=45, d=23

Prediction: a=45, e=90, d=5

Prediction: a=240, e=45, d=11
Aspect Layout Model

Prediction: $a=30$, $e=15$, $d=2.5$

Prediction: $a=345$, $e=15$, $d=3.5$

Prediction: $a=60$, $e=-30$, $d=2.5$

Prediction: $a=0$, $e=15$, $d=1.5$

Prediction: $a=315$, $e=30$, $d=2$

Prediction: $a=60$, $e=15$, $d=2$

Prediction: $a=0$, $e=30$, $d=7$

ImageNet dataset [Deng et al. 2010]
Wrong examples

Prediction: $a=45$, $e=15$, $d=1.5$

Prediction: $a=0$, $e=30$, $d=7$

Prediction: $a=225$, $e=30$, $d=7$

Prediction: $a=345$, $e=15$, $d=2.5$
Application I: Object Co-detection with 3D Aspect Parts

Application II: Multiview Object Tracking with 3D Aspect Parts

Frame 9: Azimuth=315.48, Elevation=4.56, Distance=4.98

Frame 29: Azimuth=1.34, Elevation=2.78, Distance=6.58

Frame 69: Azimuth=89.12, Elevation=3.73, Distance=2.34

Frame 48: Azimuth=25.17, Elevation=3.60, Distance=3.56

Application II: Multiview Object Tracking with 3D Aspect Parts

How to handle occlusion?

Occlusion changes the appearances of objects.
3D Aspectlet Representation

3D Aspectlets
3D Aspectlet Representation
# Object Detection Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Outdoor-scene</th>
<th>Indoor-scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>% occlusion</td>
<td>&lt; 0.3</td>
<td>0.3 – 0.6</td>
</tr>
<tr>
<td># images</td>
<td>66</td>
<td>68</td>
</tr>
<tr>
<td>ALM [1]</td>
<td>72.3</td>
<td>42.9</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>75.9</td>
<td>58.6</td>
</tr>
<tr>
<td><strong>Ours 3D Aspectlets</strong></td>
<td><strong>80.2</strong></td>
<td><strong>63.3</strong></td>
</tr>
</tbody>
</table>

Object Detection Experiments

Outdoor Scenes

Indoor Scenes
Outline

• 3D Aspect Part Representation

• 3D Voxel Pattern Representation

• A Benchmark for 3D Object Recognition in the Wild

• Conclusion and Future Work
What are the 3D aspect parts for aeroplane and bottle?
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
1. Align 2D Images with 3D CAD Models

2. Building 3D Voxel Exemplars

Depth ordering

2D mask labeling

visible

occluded

truncated

Voxelization

3D CAD model

self-occluded

truncated

visible

occluded

3D voxel model
2. Building 3D Voxel Exemplars

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

2D image
2D segmentation mask
3D voxel model
3D Voxel Exemplars

3D Discovering 3D Voxel Patterns

Clustering in 3D voxel space

3D Voxel Patterns (3DVPs)
4. Training 3D Voxel Pattern detectors

• Train an ACF detector for each 3DVP.

Testing Pipeline Overview

1. Apply 3DVP detectors
2. Transfer meta-data
3. Occlusion reasoning
4. Backproject to 3D

Input 2D image

2D detection
2D segmentation

3D localization
1. Apply 3DVP Detectors
1. Apply 3DVP Detectors
2. Transfer Meta-Data

3D Voxel Patterns
2. Transfer Meta-Data
3. Occlusion Reasoning

Occlusion reasoning: find a set of visibility-compatible detections

\[ E = \sum_i (\psi_{\text{detection\_score}} + \psi_{\text{truncation}}) + \sum_{ij} \psi_{\text{occlusion}} \]
3. Occlusion Reasoning
3. Occlusion Reasoning
4. 3D Localization

Backprojection
## Car Detection and Orientation Estimation on KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF [1]</td>
<td>55.89</td>
<td>54.77</td>
<td>42.98</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>68.02</td>
<td>56.48</td>
<td>44.18</td>
<td>67.27</td>
<td>55.77</td>
<td>43.59</td>
</tr>
<tr>
<td>DPM-VOC+VP [3]</td>
<td>74.95</td>
<td>64.71</td>
<td>48.76</td>
<td>72.28</td>
<td>61.84</td>
<td>46.54</td>
</tr>
<tr>
<td>OC-DPM [4]</td>
<td>74.94</td>
<td>65.95</td>
<td>53.86</td>
<td>73.50</td>
<td>64.42</td>
<td>52.40</td>
</tr>
<tr>
<td>SubCat [5]</td>
<td>84.14</td>
<td>75.46</td>
<td>59.71</td>
<td>83.41</td>
<td>74.42</td>
<td>58.83</td>
</tr>
<tr>
<td>Regionlets [6]</td>
<td>84.75</td>
<td><strong>76.45</strong></td>
<td>59.70</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>AOG [7]</td>
<td>84.80</td>
<td>75.94</td>
<td>60.70</td>
<td>33.79</td>
<td>30.77</td>
<td>24.75</td>
</tr>
<tr>
<td><strong>Ours 3DVP</strong></td>
<td><strong>84.81</strong></td>
<td>73.02</td>
<td><strong>63.22</strong></td>
<td><strong>84.31</strong></td>
<td>71.99</td>
<td><strong>62.11</strong></td>
</tr>
</tbody>
</table>

## Car Detection and Orientation Estimation on KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF [1]</td>
<td>55.89</td>
<td>54.77</td>
<td>42.98</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>68.02</td>
<td>56.48</td>
<td>44.18</td>
<td>67.27</td>
<td>55.77</td>
<td>43.59</td>
</tr>
<tr>
<td>DPM-VOC+VP [3]</td>
<td>74.95</td>
<td>64.71</td>
<td>48.76</td>
<td>72.28</td>
<td>61.84</td>
<td>46.54</td>
</tr>
<tr>
<td>OC-DPM [4]</td>
<td>74.94</td>
<td>65.95</td>
<td>53.86</td>
<td>73.50</td>
<td>64.42</td>
<td>52.40</td>
</tr>
<tr>
<td>SubCat [5]</td>
<td>84.14</td>
<td>75.46</td>
<td>59.71</td>
<td>83.41</td>
<td>74.42</td>
<td>58.83</td>
</tr>
<tr>
<td>Regionlets [6]</td>
<td>84.75</td>
<td><strong>76.45</strong></td>
<td>59.70</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>AOG [7]</td>
<td>84.80</td>
<td>75.94</td>
<td>60.70</td>
<td>33.79</td>
<td>30.77</td>
<td>24.75</td>
</tr>
<tr>
<td><strong>Ours</strong> 3DVP</td>
<td>84.81</td>
<td>73.02</td>
<td>63.22</td>
<td>84.31</td>
<td>71.99</td>
<td>62.11</td>
</tr>
<tr>
<td><strong>Ours</strong> Occlusion</td>
<td><strong>87.46</strong></td>
<td>75.77</td>
<td><strong>65.38</strong></td>
<td><strong>86.92</strong></td>
<td><strong>74.59</strong></td>
<td><strong>64.11</strong></td>
</tr>
</tbody>
</table>

Apply 3DVPs in CNN-based Object Detection

• 3DVPs as subcategories
• Two-stage detection framework

An input image → Region proposals → CNN → Car, Cat, Dog, Person, ...

- R. Girshick et al., CVPR’14
- R. Girshick, ICCV’15
- S. Ren et al., NIPS’15
- S. Gidaris and N. Komodakis, CoRR’15
Subcategory-ware Region Proposal Network

Input image (image pyramid) ➔ Conv layers ➔ Feature extraction ➔ Feature map ➔ Subcategory Conv filters ➔ Heatmaps ➔ Region proposals

**Feature Extrapolating Layer**
- Generate features in nearby scales by extrapolating

**RoI Generating Layer**
- Training: hard positives and hard negatives
- Testing: high score boxes

arXiv:1604.04693
Subcategory-ware Detection Network

Input image → 5 Conv layers

Region proposals

5 Conv layers Feature extraction → RoI pooling Layer [1] → FC6 (4096) → FC7 (4096) → FC8 (K+1)

Class loss → Bounding box Regression loss → Subcategory classification loss

## Car Detection and Orientation Estimation on KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF [1]</td>
<td>55.89</td>
<td>54.77</td>
<td>42.98</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>68.02</td>
<td>56.48</td>
<td>44.18</td>
<td>67.27</td>
<td>55.77</td>
<td>43.59</td>
</tr>
<tr>
<td>DPM-VOC+VP [3]</td>
<td>74.95</td>
<td>64.71</td>
<td>48.76</td>
<td>72.28</td>
<td>61.84</td>
<td>46.54</td>
</tr>
<tr>
<td>OC-DPM [4]</td>
<td>74.94</td>
<td>65.95</td>
<td>53.86</td>
<td>73.50</td>
<td>64.42</td>
<td>52.40</td>
</tr>
<tr>
<td>SubCat [5]</td>
<td>84.14</td>
<td>75.46</td>
<td>59.71</td>
<td>83.41</td>
<td>74.42</td>
<td>58.83</td>
</tr>
<tr>
<td>Regionlets [6]</td>
<td>84.75</td>
<td>76.45</td>
<td>59.70</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>AOG [7]</td>
<td>84.80</td>
<td>75.94</td>
<td>60.70</td>
<td>33.79</td>
<td>30.77</td>
<td>24.75</td>
</tr>
<tr>
<td>Mono3D [8]</td>
<td>92.33</td>
<td>88.66</td>
<td>78.96</td>
<td>91.01</td>
<td>86.62</td>
<td>76.84</td>
</tr>
<tr>
<td><strong>Ours</strong> 3DVP</td>
<td>84.81</td>
<td>73.02</td>
<td>63.22</td>
<td>84.31</td>
<td>71.99</td>
<td>62.11</td>
</tr>
<tr>
<td><strong>Ours</strong> Occlusion</td>
<td>87.46</td>
<td>75.77</td>
<td>65.38</td>
<td>86.92</td>
<td>74.59</td>
<td>64.11</td>
</tr>
<tr>
<td><strong>Ours</strong> CNN</td>
<td>90.86</td>
<td><strong>89.32</strong></td>
<td><strong>79.33</strong></td>
<td>90.72</td>
<td><strong>88.97</strong></td>
<td><strong>78.83</strong></td>
</tr>
</tbody>
</table>


Detection: Rank 3
Pose: Rank 1
3D Voxel Patterns from PASCAL3D+ [1]

12 Rigid Categories

## Detection and Pose Estimation on PASCAL3D+

<table>
<thead>
<tr>
<th>Method</th>
<th>Detection (AP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM [1]</td>
<td>29.6</td>
</tr>
<tr>
<td>R-CNN [2]</td>
<td>56.9</td>
</tr>
<tr>
<td><strong>Ours</strong> CNN</td>
<td><strong>60.7</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>4 Views (AVP)</th>
<th>8 Views (AVP)</th>
<th>16 Views (AVP)</th>
<th>24 Views (AVP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VDPM [3]</td>
<td>19.5</td>
<td>18.7</td>
<td>15.6</td>
<td>12.1</td>
</tr>
<tr>
<td>DPM-VOC+VP [4]</td>
<td>24.5</td>
<td>22.2</td>
<td>17.9</td>
<td>14.4</td>
</tr>
<tr>
<td><strong>Ours</strong> CNN</td>
<td><strong>47.5</strong></td>
<td><strong>31.9</strong></td>
<td><strong>24.5</strong></td>
<td><strong>19.3</strong></td>
</tr>
</tbody>
</table>

Application: Online Multi-Object Tracking

Outline

• 3D Aspect Part Representation

• 3D Voxel Pattern Representation

• A Benchmark for 3D Object Recognition in the Wild

• Conclusion and Future Work
Goal

• Build a large scale dataset for 3D object recognition in the wild
# 3D Object Dataset

<table>
<thead>
<tr>
<th>#category</th>
<th>#instance</th>
<th>Non-centered objects</th>
<th>Dense viewpoint</th>
<th>3D Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Object [1]</td>
<td>10</td>
<td>100</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

## EPFL Car Dataset

<table>
<thead>
<tr>
<th></th>
<th>#category</th>
<th>#instance</th>
<th>Non-centered objects</th>
<th>Dense viewpoint</th>
<th>3D Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Object [1]</td>
<td>10</td>
<td>100</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>EPFL Car [2]</td>
<td>1</td>
<td>20</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

### PASCAL VOC Dataset

<table>
<thead>
<tr>
<th></th>
<th>#category</th>
<th>#instance</th>
<th>Non-centered objects</th>
<th>Dense viewpoint</th>
<th>3D Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Object [1]</td>
<td>10</td>
<td>100</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>EPFL Car [2]</td>
<td>1</td>
<td>20</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>PASCAL VOC [3]</td>
<td>20</td>
<td>27,450</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

# KITTI Dataset

<table>
<thead>
<tr>
<th></th>
<th>#category</th>
<th>#instance</th>
<th>Non-centered objects</th>
<th>Dense viewpoint</th>
<th>3D Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Object [1]</td>
<td>10</td>
<td>100</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>EPFL Car [2]</td>
<td>1</td>
<td>20</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>PASCAL VOC [3]</td>
<td>20</td>
<td>27,450</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>KITTI [4]</td>
<td>3</td>
<td>80,256</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

Our Contribution: PASCAL3D+

<table>
<thead>
<tr>
<th></th>
<th>#category</th>
<th>#instance</th>
<th>Non-centered objects</th>
<th>Dense viewpoint</th>
<th>3D Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Object [1]</td>
<td>10</td>
<td>100</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>EPFL Car [2]</td>
<td>1</td>
<td>20</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>PASCAL VOC [3]</td>
<td>20</td>
<td>27,450</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>KITTI [4]</td>
<td>3</td>
<td>80,256</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td><strong>PASCAL3D+ (Ours)</strong></td>
<td><strong>12</strong></td>
<td><strong>35,672</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

3D Annotation: 2D-3D Alignment
PASCAL3D+: A Benchmark for 3D Object Recognition
## Our Contribution: ObjectNet3D

<table>
<thead>
<tr>
<th></th>
<th>#category</th>
<th>#instance</th>
<th>Non-centered objects</th>
<th>Dense viewpoint</th>
<th>3D Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Object [1]</td>
<td>10</td>
<td>100</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>EPFL Car [2]</td>
<td>1</td>
<td>20</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>PASCAL VOC [3]</td>
<td>20</td>
<td>27,450</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>KITTI [4]</td>
<td>3</td>
<td>80,256</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>PASCAL3D+ (Ours)</td>
<td>12</td>
<td>35,672</td>
<td>✓</td>
<td>✓</td>
<td>✓ 79</td>
</tr>
<tr>
<td><strong>ObjectNet3D (Ours)</strong></td>
<td><strong>100</strong></td>
<td><strong>178,633</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓ 44,147</td>
</tr>
</tbody>
</table>

Under Review
ObjectNet3D: A Large Scale Database for 3D Object Recognition

2D-3D alignment

Images from ImageNet

3D Shapes from 3D Warehouse and ShapeNet

100 rigid object categories
Annotation Demo
Viewpoint Distribution

bed  car  cup  hammer

laptop  mouse  shoe  teapot
Image-based 3D Shape Retrieval

Conclusion

• 3D Aspect Part Representation

• 3D Voxel Pattern Representation

• A Benchmark for 3D Object Recognition in the Wild
  • PASCAL3D+ and ObjectNet3D
Future Work: Generalization of Object Recognition

Training Data (with annotations)

Testing Data (with few annotations or even no annotation)

- How to achieve generalization across domains?
- Can 3D object representations improve generalization?
- Can we find better 3D object representations for recognition?
Future Work: Joint Recognition and Reconstruction

Utilize large scale 3D shape data

Karsch et al., CVPR’13
Future Work: End-to-End Multi-Object Tracking

Video frames → Object Detector → Object Tracker → Object Trajectories

Tracking by Detection

Video frames → Object Detection and Tracking → Object Trajectories

Learning to Detect and Track
Future Work: Putting Objects in the Scene

- 3D object recognition and scene geometry understanding
  - Holistic 3D scene understanding
Conclusion

• 3D aspect part representation

• 3D voxel pattern representation

• A Benchmark for 3D Object Recognition in the Wild

Thank you!