Synthesizing Training Data for Object Detection in Indoor Scenes

Georgios Georgakis
Arsalan Mousavian
Alexander C. Berg
Jana Kosecka
Object Detection in Indoor Scenes

- Given an RGB-D image $I$, localize all bounding boxes $\{x,y,w,h\}$, of object instances of interest $c$ - instance recognition.
Challenges

- Viewpoint variation
- Clutter and scale variations
- Occlusions

- State-of-the-art object detectors SSD / Faster R-CNN – deep learning.
- These methods require large amounts of training data.
- Manual annotation is time consuming.

How can we train SSD and Faster R-CNN with the least amount of manual annotation?
System Overview

Cropped Objects (BigBird) -> Background Scenes (NYUv2) -> Synthetic Training Set

Set Generation

GEORGE MASON UNIVERSITY

THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
System Overview

Cropped Objects (BigBird)

Background Scenes (NYUv2)

Synthetic Training Set

Train

Test

CNN

Real Images

Detections
Synthetic Set Generation

• Important observations:
  • Household objects are found on support surfaces.
  • The distance from camera defines the scale.
  • Objects in natural images have smooth boundary discontinuities with the background.

• Three simple constraints for increased realism:
  • Selective positioning, Selective scaling, and Blending.
Automatic Generation

RGB

Depth
Automatic Generation

Semantic segmentation state-of-the-art approach:

Automatic Generation

Selective Positioning

Semantic segmentation

Support surface estimation

RGB

Depth

Selective Scaling
Automatic Generation

Off-the-shelf blending technique:

Composited Examples

Our approach

Randomized approach
Experiments

• GMU-Kitchens dataset
  • Georgakis et al. “Multiview RGB-D dataset for object instance detection”, 3DV 2016

• Synth to Real
  • Informatically generated synthetic sets (~20000 images) outperform random sets.

• Synth+Real to Real
  • Full real training set ~4000 images.
  • Augmenting existing few real annotations (400 real + 20000 synthetic) can yield comparable or superior results.

<table>
<thead>
<tr>
<th>Train Set</th>
<th>SSD / Faster R-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real to Real</td>
</tr>
<tr>
<td>1. Real Scenes</td>
<td>65.6 / 82.5</td>
</tr>
<tr>
<td></td>
<td>Synthetic to Real</td>
</tr>
<tr>
<td>2. RP-SI-RS</td>
<td>23.2 / 42.1</td>
</tr>
<tr>
<td>3. RP-BL-RS</td>
<td>22.2 / 44.2</td>
</tr>
<tr>
<td>4. SP-SI-SS</td>
<td>19.5 / 48.6</td>
</tr>
<tr>
<td>5. SP-BL-SS</td>
<td><strong>33.5 / 51.7</strong></td>
</tr>
<tr>
<td></td>
<td>Synthetic+Real to Real</td>
</tr>
<tr>
<td>6. 1% real</td>
<td>60.3 / 69.3</td>
</tr>
<tr>
<td>7. 10% real</td>
<td>71.6 / 79.2</td>
</tr>
<tr>
<td>8. 50% real</td>
<td><strong>74.1 / 83.8</strong></td>
</tr>
<tr>
<td>9. 100% real</td>
<td>73.7 / <strong>85.0</strong></td>
</tr>
</tbody>
</table>
Qualitative Results

Faster R-CNN

SSD

Real

Real + Synth
Summary

• Synthetic training data are more effective when they are generated with semantic and geometric information.

• Object detectors can be trained with significantly less annotated data using our proposed synthetic data augmentation.