Estimating Generic 3D Room Structures from 2D Annotations

Denys Rozumnyi  Stefan Popov  Kevis-Kokitsi Maninis  Matthias Nießner  Vittorio Ferrari

Introduction

Goal: Create 3D room layouts from RGB video (no depth) → commonly available

Method: create 3D room layouts only from 2D annotations → easy for humans!
- Few (real) prior datasets, all requiring special acquisition devices (RGB-D, panos)
- The dataset is released here: https://github.com/google-research/cad-estate

Pipeline

- Inputs are familiar 2D segmentation
- Each frame is annotated independently, without any correspondences

Method

Given an input video and manual 2D annotations of structural elements and their visible parts, we combine point tracks fitting, edge matching, and perpendicularity constraints to generate a 3D room layout.

Evaluation

- Low depth errors and very high IoU values → high quality reconstructions
- Automatic quality control: reject reconstructions with IoU < 0.8
  → IoU and depth worse when turned off
  → works well
- Run method many times and select automatically based on IoU
  → indirectly minimize depth error → good to have many runs
- We train and evaluate a baseline method [31] that performs at the state-of-the-art on the existing datasets, with a low error around 6% ~ 7%.
- Instead, it performs much worse on our dataset (20%), demonstrating it offers a harder challenge than the previous one

Spatial extent refinement, before (top) and after (bottom).

We cut hanging walls extending outside the room boundary, and fill in the holes between neighboring planes (blue).

Input annotations and final reconstructions


<table>
<thead>
<tr>
<th>Runs</th>
<th>Depth Error (%)</th>
<th>Pixel Error (%)</th>
<th>Pixel Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (full method)</td>
<td>10</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Ours (no quality control)</td>
<td>30</td>
<td>0.84</td>
<td>0.33</td>
</tr>
<tr>
<td>Ours (no quality control)</td>
<td>1</td>
<td>0.72</td>
<td>0.79</td>
</tr>
</tbody>
</table>

We train and evaluate the method [31] on each frame independently. In Table 3, we report the average depth error of all frames of all runs, and the per-element average depth error, for each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pixel Error (%)</th>
<th>Pixel Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedau et al. [19]</td>
<td>4.83</td>
<td>12.83</td>
</tr>
<tr>
<td>Malby et al. [35]</td>
<td>16.71</td>
<td>12.83</td>
</tr>
<tr>
<td>DeLay [42]</td>
<td>10.63</td>
<td>9.73</td>
</tr>
<tr>
<td>CPHLE [45]</td>
<td>7.57</td>
<td>8.67</td>
</tr>
<tr>
<td>Zhang et al. [68]</td>
<td>6.38</td>
<td>12.70</td>
</tr>
<tr>
<td>STE-PRO [71]</td>
<td>5.29</td>
<td>6.60</td>
</tr>
<tr>
<td>Lin et al. [51] (baseline)</td>
<td>6.23</td>
<td>7.41</td>
</tr>
</tbody>
</table>

We validate the quality of our reconstructions on the ScanNet dataset [11], which gathered 50 scenes from the ScanNet dataset.

• Limitations

- Estimating Generic 3D Room Structures from 2D Annotations

- 2D projection of a 3D room layout

- Human annotators

- Machine

- Result: 3D mesh