PQ-NET:
A Generative Part Seq2Seq Network for 3D Shapes

Rundi Wu$^{1,4}$ Yixin Zhuang$^1$ Kai Xu$^2$ Hao Zhang$^{3,4}$ Baoquan Chen$^{1,4}$

$^1$ Center on Frontiers of Computing Studies, Peking University
$^2$ National University of Defense Technology
$^3$ Simon Fraser University
$^4$ AICFVE, Beijing Film Academy
3D shape generation

Voxel grid

[3DGAN, NIPS 2016]

Point cloud

[Pointflow, ICCV 2019]

Mesh

[AtlasNet, CVPR 2018]

Implicit function

[DeepSDF, CVPR 2019]


Structural 3D shape generation


Shape structure presentations

① *hierarchical* part organization \(\approx\) phrases nested in phrases

② *linear* part order \(\approx\) linear string of words

"the men will find the books"
Generate as a sequence

- Our network, PQ-NET, learns 3D shape representation via *sequential part assembly*
Method

a. Apply IM-NET to encode each scaled part’s geometry
b. Model sequential part assembly using a Sequence-to-Sequence Auto-encoder (Seq2Seq AE)
Method - Part geometry encoding

Similar architecture as IM-NET\(^1\):
- a CNN encoder \( e \) maps \( 64^3 \) voxelized part \( P \) to 128D vector
- a MLP decoder \( d \) that predicts the occupancy of a given point \( p \)

\[
\mathcal{L}(P) = \mathbb{E}_{p \in T_P} |d(e(P), p) - F(p)|^2
\]

A set of sampled points from \( P \)
ground truth signed function

---

Method - Seq2Seq AE

Encoder:
- a bidirectional stacked RNN to encode part sequence

- Stacked GRU Cell
- Number of parts in one-hot representation
- Part Box Parameter: 6D, position + size
- Part Geometry Feature: latent vector encoded by IM-NET
Method - Seq2Seq AE

Decoder:
- a stacked RNN to predict geometry and structure feature separately

- GRU Cell
- $I_0$: Initial input: zero vector
- Stop sign: a confidence value between 0~1
- Part Box Parameter: 6D, position + size
- Part Geometry Feature: latent vector to be decoded by IM-NET
Method - Seq2Seq AE

Training losses

- **MSE** loss on the reconstruction of geometry feature $g_i$ and structure feature $b_i$
- **Binary Cross Entropy** loss on the stop sign $s_i$ predicted by decoder

$$
L_r(S) = \frac{1}{k} \sum_{i=1}^{k} [\beta ||g'_i - g_i||_2 + ||b'_i - b_i||_2] \\
L_{stop}(S) = \frac{1}{k} \sum_{i=1}^{k} [-s_i \log s'_i - (1 - s_i) \log(1 - s'_i)]
$$
Results: shape auto-encoding

a) Ground Truth

b) IM-NET-256

c) Ours-256

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Method</th>
<th>Chair</th>
<th>Table</th>
<th>Lamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoU</td>
<td>Ours-64</td>
<td>67.29</td>
<td>47.39</td>
<td>39.56</td>
</tr>
<tr>
<td></td>
<td>IM-NET-64</td>
<td>62.93</td>
<td><strong>56.14</strong></td>
<td><strong>41.29</strong></td>
</tr>
<tr>
<td></td>
<td>Ours-64</td>
<td>3.38</td>
<td>5.49</td>
<td>11.49</td>
</tr>
<tr>
<td></td>
<td>Ours-256</td>
<td>2.86</td>
<td>5.69</td>
<td>10.32</td>
</tr>
<tr>
<td></td>
<td>Ours-Cross-256</td>
<td>2.46</td>
<td><strong>4.50</strong></td>
<td>4.87</td>
</tr>
<tr>
<td></td>
<td>IM-NET-64</td>
<td>3.64</td>
<td>6.75</td>
<td>12.43</td>
</tr>
<tr>
<td></td>
<td>IM-NET-256</td>
<td>3.59</td>
<td>6.31</td>
<td>12.19</td>
</tr>
<tr>
<td></td>
<td>Ours-64</td>
<td>2734</td>
<td>2824</td>
<td>6254</td>
</tr>
<tr>
<td></td>
<td>Ours-256</td>
<td><strong>2441</strong></td>
<td>2609</td>
<td>5941</td>
</tr>
<tr>
<td></td>
<td>Ours-Cross-256</td>
<td>2501</td>
<td><strong>2415</strong></td>
<td><strong>4875</strong></td>
</tr>
<tr>
<td></td>
<td>IM-NET-64</td>
<td>2830</td>
<td>3446</td>
<td>6262</td>
</tr>
<tr>
<td></td>
<td>IM-NET-256</td>
<td>2794</td>
<td>3397</td>
<td>6622</td>
</tr>
</tbody>
</table>
Results: shape generation

a) Ours

b) IM-NET

c) StructureNet

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>COV</th>
<th>MMD</th>
<th>JSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair</td>
<td>Ours</td>
<td>54.91</td>
<td>8.34</td>
<td>0.0083</td>
</tr>
<tr>
<td></td>
<td>IM-NET</td>
<td>52.35</td>
<td>7.44</td>
<td>0.0084</td>
</tr>
<tr>
<td></td>
<td>StructureNet</td>
<td>29.51</td>
<td>9.67</td>
<td>0.0477</td>
</tr>
<tr>
<td>Table</td>
<td>Ours</td>
<td>56.51</td>
<td>7.56</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>IM-NET</td>
<td>56.67</td>
<td>6.90</td>
<td>0.0047</td>
</tr>
<tr>
<td></td>
<td>StructureNet</td>
<td>16.04</td>
<td>14.98</td>
<td>0.0725</td>
</tr>
<tr>
<td>Lamp</td>
<td>Ours</td>
<td>87.95</td>
<td>10.01</td>
<td>0.0215</td>
</tr>
<tr>
<td></td>
<td>IM-NET</td>
<td>81.25</td>
<td>10.45</td>
<td>0.0230</td>
</tr>
<tr>
<td></td>
<td>StructureNet</td>
<td>35.27</td>
<td>17.29</td>
<td>0.1719</td>
</tr>
</tbody>
</table>
Results: latent space interpolation
Results: single view reconstruction

a) Input image

b) IM-NET

c) Ours

d) Ground Truth
Results: comparison to 3D-PRNN

- Shape reconstruction from single depth image
- Compare on two orders: (A) PartNet default (B) enforced top-down

<table>
<thead>
<tr>
<th>Method</th>
<th>Order</th>
<th>Chair</th>
<th>Table</th>
<th>Lamp</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>A</td>
<td>61.47</td>
<td>53.67</td>
<td>52.94</td>
<td>56.03</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>58.68</td>
<td>48.58</td>
<td>52.17</td>
<td>53.14</td>
</tr>
<tr>
<td>3D-PRNN</td>
<td>A</td>
<td>37.26</td>
<td>51.30</td>
<td>47.26</td>
<td>45.27</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>36.46</td>
<td>51.93</td>
<td>43.83</td>
<td>44.07</td>
</tr>
</tbody>
</table>
Results: applications

- Order denosing and part correspondence
- Re-train the model to correct the input order
Results: applications

- Partial shape completion
  - Re-train the model to reconstruct from partial shape input
Limitation

- PQ-NET do not produce part relations
  - Comparing to prior works that seek to hierarchical representation
- The order of parts could affect the performance
  - A consistent part order over the dataset is required
Thanks!

Code and data: https://github.com/ChrisWu1997/PQ-NET