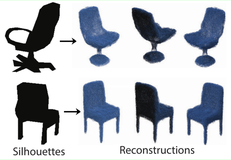
Synthesizing 3D Shapes via Modeling Multi-View Depth Maps and Silhouettes with Deep Generative Networks

Amir Arsalan Soltani[1], Haibin Huang[2], Jiajun Wu[1], Tejas D. Kulkarni[3], Joshua B. Tenenbaum[1]

[1] Massachusetts Institute of Technology [2] University of Massachusetts, Amherst [3] Google DeepMind London



1. Motivation

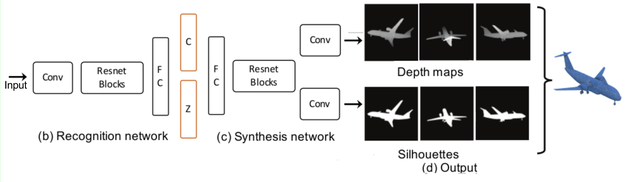
- **Multi-View** Representations:
  - Ease of Computation, Flexible, Easy to Obtain (E.g. RGB-D)
- Still **Explicitly** Model 3D Shapes
- Synthesize/Generate **Hi-Res, Detailed, Novel** 3D Shapes
- Recover the Underlying 3D Representation
- Have **Out-of-Sample** Generalizability
  - NYUD Data Set and Silhouettes From the Internet
  - Despite Training on Limited Data and Limited Views
- Using **2D** to obtain the underlying 3D representation
- **Share** the same representations for all objects

2. Pipeline & Architecture

- Used **ShapeNet Core** Data set
  - 54 Categories, ~37k and ~3k samples for train and test
- Render **20** depth maps from them
  - **Camera Angle Fixed, Shapes Normalized and Aligned**
- Train Models via the **VAE** Framework
  - 20 (**AllVPNet**), 2-5 (**DropoutNet**) or a single (**SingleVPNet**) view
  - 100 Hidden Notes for Unconditional Models, 40 for Conditional
- **L_1** loss function is used to train the models
- **Fusing depth maps** to give **3D** point clouds
- Number of Points as low as **30k** and as High as **400k**

Generated Depth Maps by our data generation pipeline

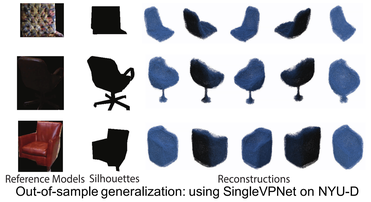
Input Types: AllVPNet takes in as input all 20 views, DropoutNet 2-5 views and SingleVPNet only a single view



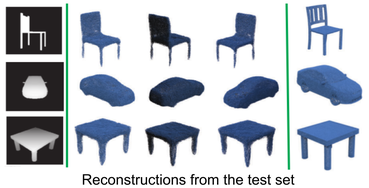
3. Results

Random Sampling

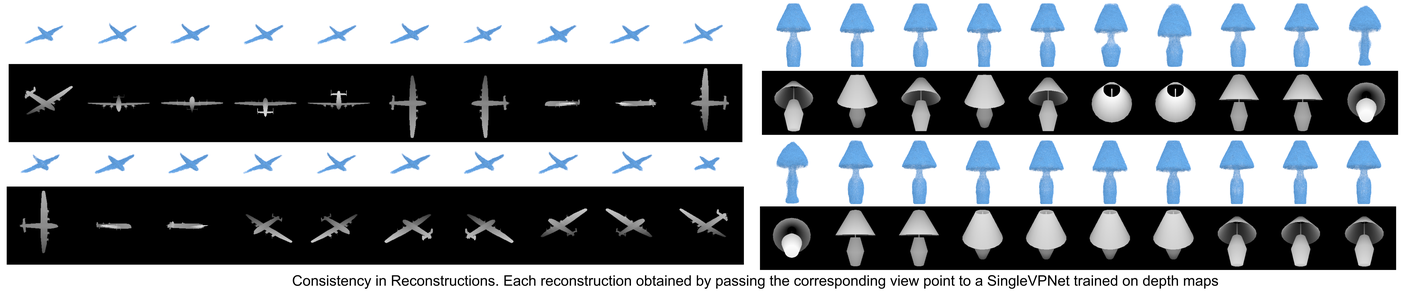
Conditional Samples: Note that there is only about **~250** and **~50** motor cycle/bike and headphones samples in the entire data set respectively



Out-of-sample generalization: using SingleVPNet on NYU-D



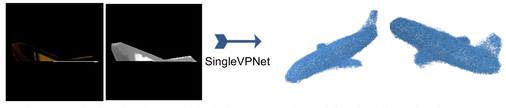
Reconstructions from the test set



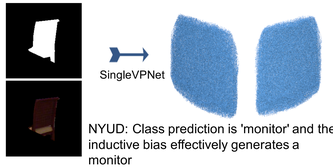
Consistency in Reconstructions. Each reconstruction obtained by passing the corresponding view point to a SingleVPNet trained on depth maps

Viewpoint	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Acc.	83.5	84.0	82.1	81.5	80.0	80.7	80.6	85.3	83.8	79.4	79.1	85.1	83.7	85.7	85.5	84.8	85.0	84.7	85.9	84.1
IoU	73.3	70.2	69.4	69.7	68.8	67.2	67.4	736	73.2	68.0	67.9	73.0	73.0	73.6	73.2	73.4	73.6	73.3	73.4	73.3

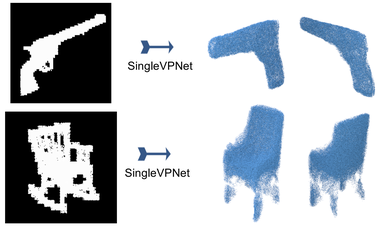
Consistency in Representations: **SingleVPNet** is able to obtain similar representations given different viewpoints



A very occluded depth map in NYUD. The conditional SingleVPNet classifies the image as an airplane and induces the effective bias to generate an airplane

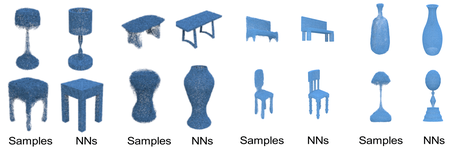


NYUD: Class prediction is 'monitor' and the inductive bias effectively generates a monitor



Feeding silhouettes from the Internet to SingleVPNet

4. Analysis



Random samples and their **nearest neighbor** training examples

Randomly generated samples: The models seems to **implicitly** learn about the different parts and their relations.

Network	Training Set		Test Set		Acc. (%)
	Depth	Sil.	Depth	Sil.	
AllVP	0.016	0.019	0.019	0.022	88.8
Dropout	0.022	0.027	0.024	0.029	86.1 ± 0.1
SingleVP	0.027	0.034	0.031	0.039	83.2 ± 0.2
AllVP	0.015	0.017	0.019	0.022	88.3
Dropout	0.023	0.027	0.024	0.029	85.0 ± 0.1
SingleVP	0.030	0.037	0.036	0.044	80.9 ± 0.2

Classification Accuracy and Reconstruction Error For ShapeNet Core. Top 3: trained on depth maps

Models	Representation	Accuracy (%)
DeepPano [21]	panorama	78%
3D ShapeNet [29]	voxel	77%
VoxNet [16]	voxel	83%
MVCNN [24]	multi-view	90%
AllVPNet	multi-view	$82.1\% \pm 0.1$
DropoutNet	multi-view	$74.2\% \pm 0.2$
SingleVPNet	single-view	$65.3\% \pm 0.3$

Classification Accuracy and Reconstruction Error For ModelNet40. Top 3: trained on depth maps

Acknowledgements

This work is supported by ONR MURI N00014-16-1-2007, the Center for Brain, Minds and Machines (NSF STC award CCF-1231216), and grants from the Toyota Research Institute and Adobe Research.