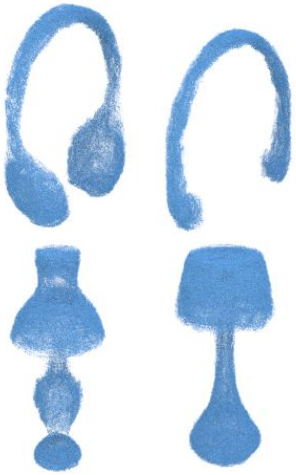
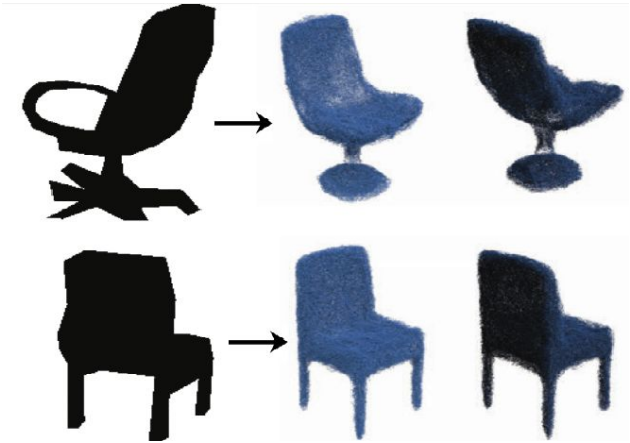
Synthesizing 3D Shapes via Modeling Multi-View Depth Maps and Silhouettes with Deep Generative Networks



Samples



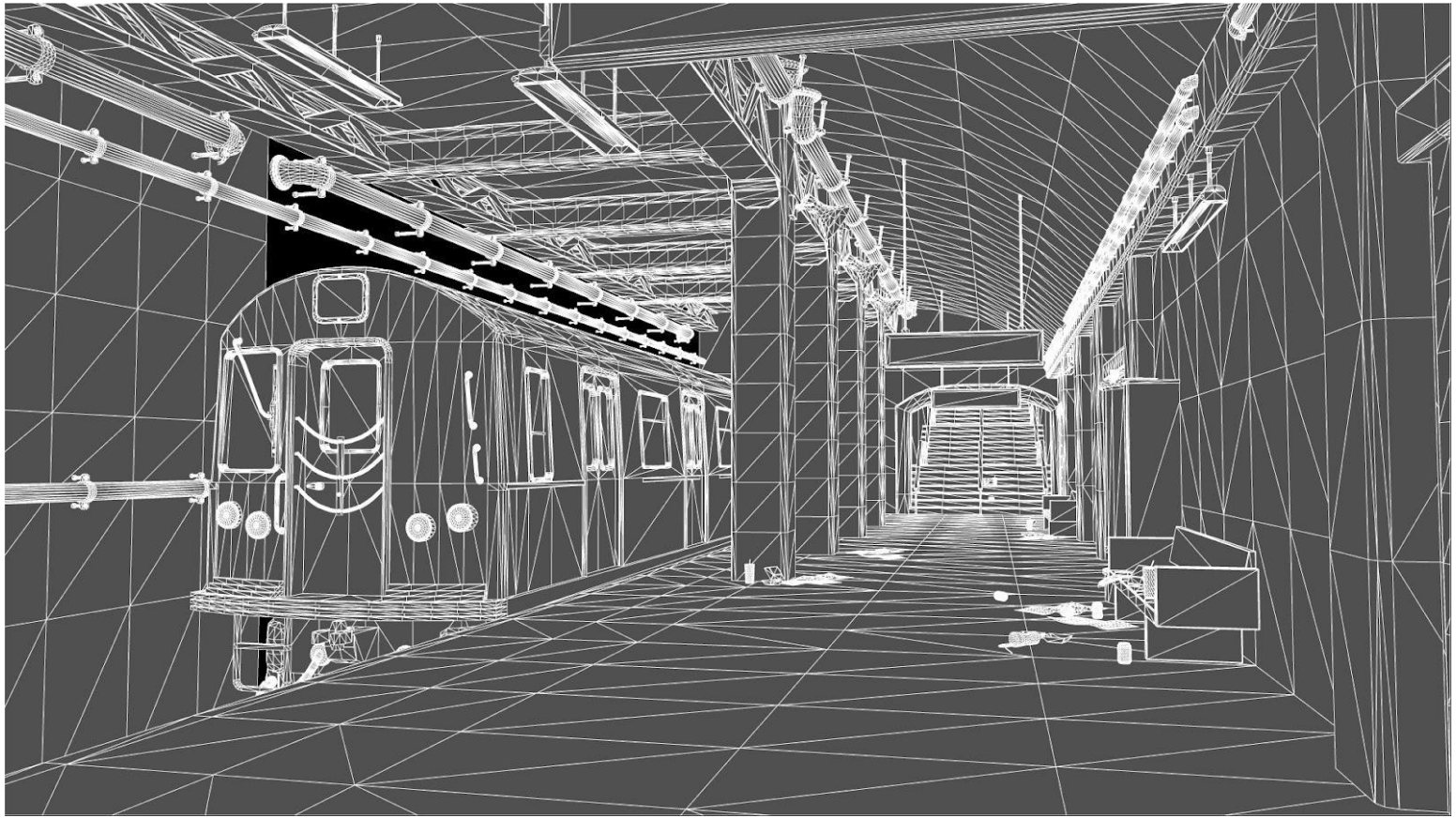
Out-of-Sample Generalization

Amir A. Soltani
Haibin Huang
Jiajun Wu
Tejas Kulkarni
Josh Tenenbaum

Motivation - 3D Vision Is Hard

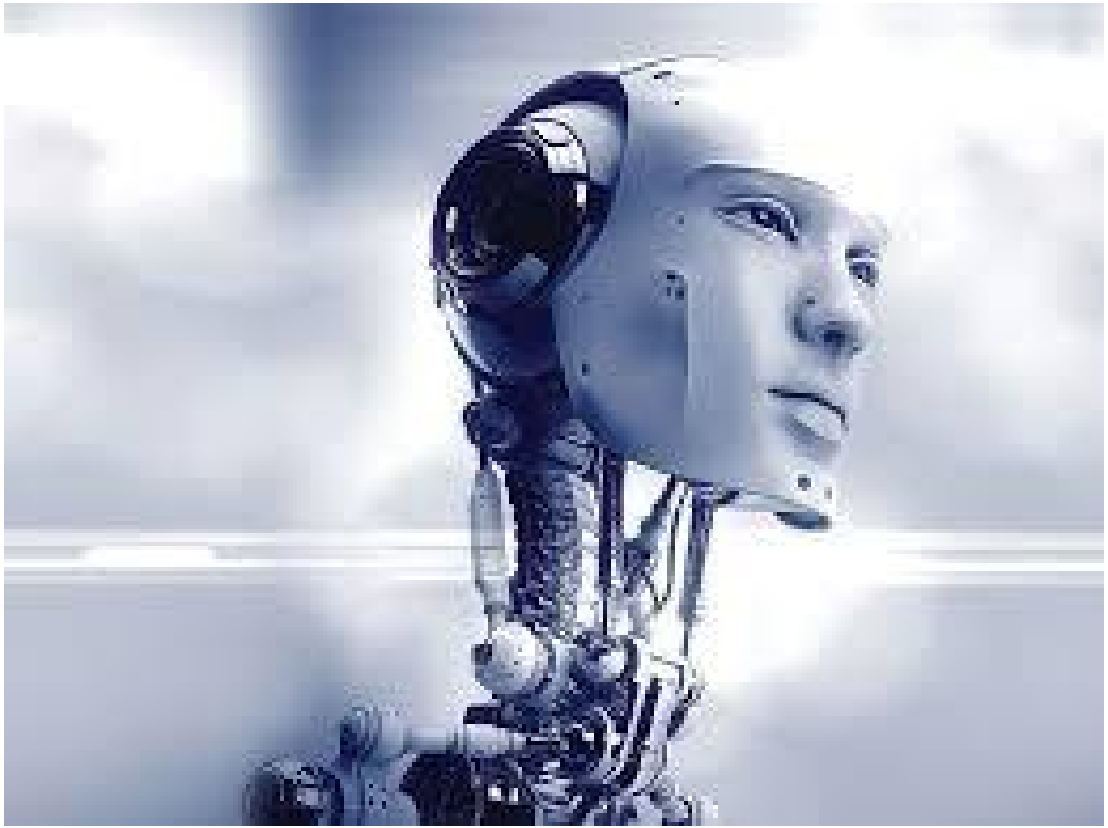
- Motivations
  - High-level
  - Technical
- Pipeline
  - Architectures
  - Training
- Results
  - Random & conditional sampling
  - Reconstruction
  - Representation analysis
- 3D Representations for Future AI Agents

Motivation - The World is in 3D



Motivation - The World is in 3D

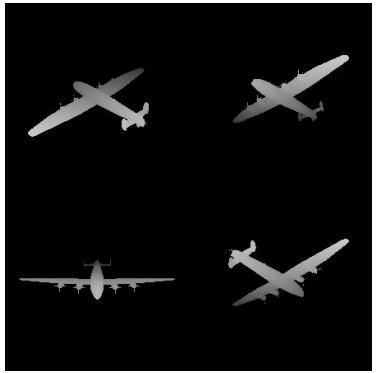
Motivation - The World is in 3D



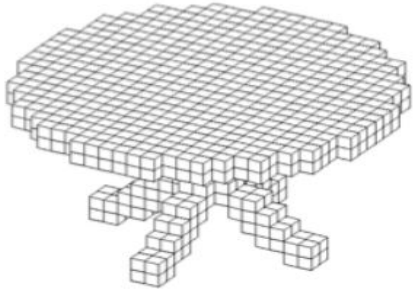
Motivation

- Computer Vision cannot rely on 2D images to solve 3D problems
- We need to have good 3D representations to solve inverse problems in 3D
- A generative model for 3D is a good starting point (A lot more needed though)
- Good progress has been made in the past 2 or 3 years
- Still, the choice of 3D representation is being debated
- Each representation has advantages and disadvantages
- So far there is not a good agreement on which representation to use
- We argue that multi-view representation is a better choice overall

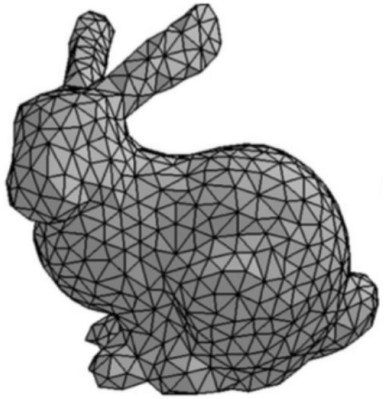
Motivation - Choice of Representation



Multi-view

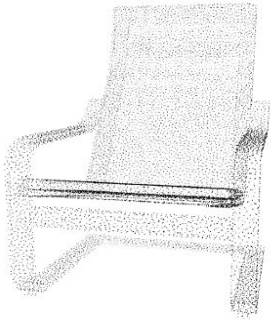


Voxels



Meshes

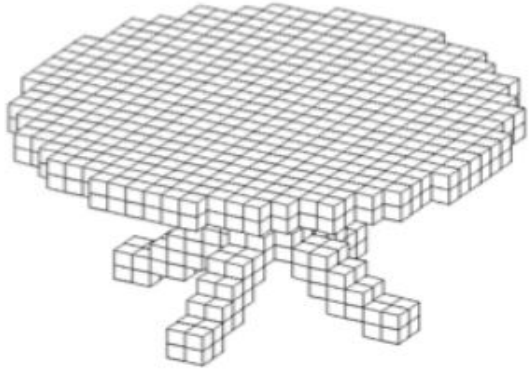
Template-based



Point clouds

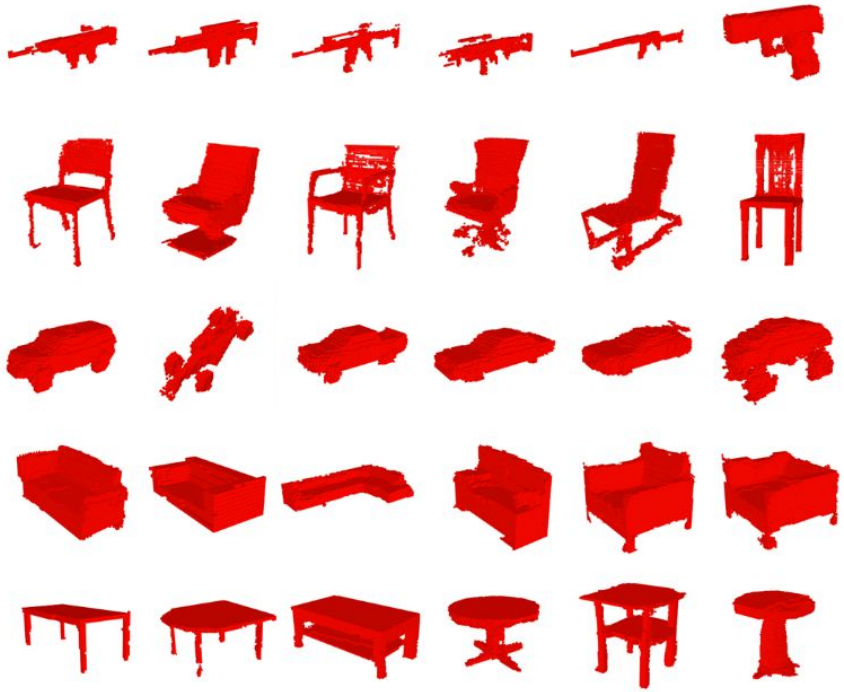
Choice of Representation - Voxels

- Computational complexity is very high (O^3) if used naively
- Cannot Model High-Res Shapes
- Details can easily get lost
- Highly sparse at higher resolutions
- Cannot model regular structures easily



Choice of Representation - Voxels

- Directly predicting high-res voxel-based outputs is very hard
- Highest res so far is 64 x 64 x 64
- One model trained per category



Wu et al, NIPS 2016

Choice of Representation - Point Clouds

- Things start to get mathematically-involved from here
  - The choice of loss function, non-differentiability issues etc
- Not obvious how many points to choose
  - Details Will Be Missing
- Not a lot of work done on point clouds so far

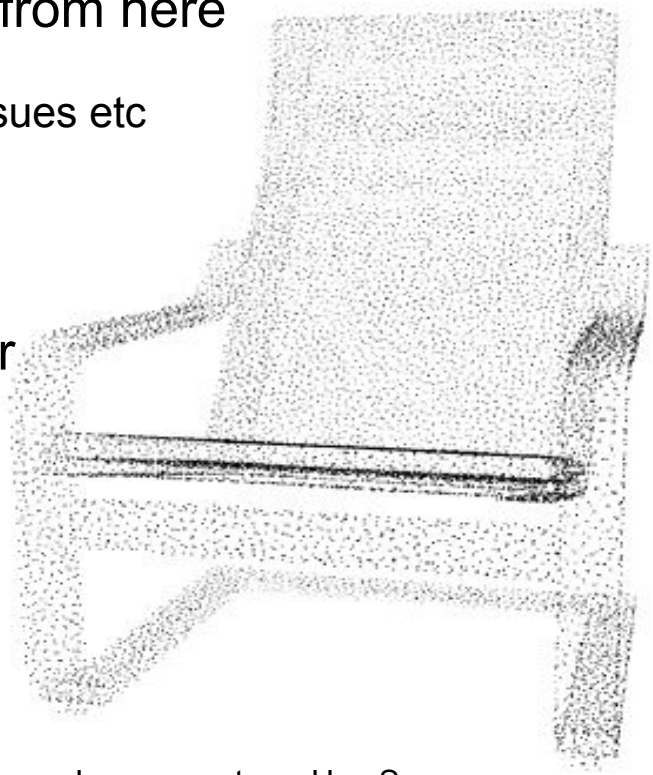
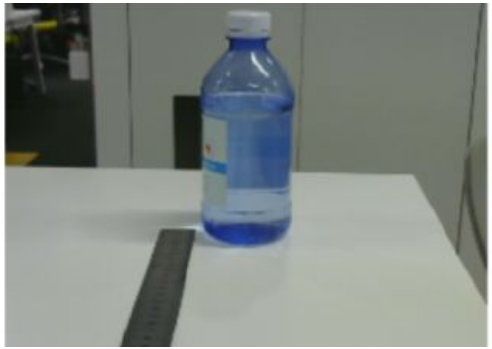
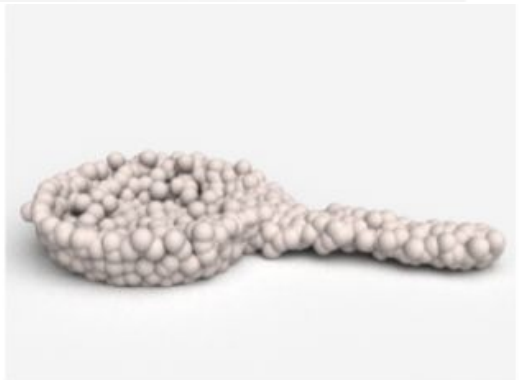


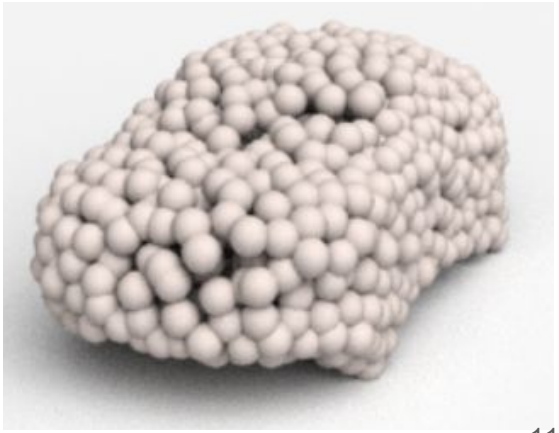
Image courtesy: Hao Su

Choice of Representation - Point Clouds

- Trained on ~200k shapes
- Used 5000 points to represent produced shapes







Su et al, CVPR 2017

Choice of Representation - Meshes

- Cannot directly apply out-of-the-box models on
- Need to build special kind of kernels for CNNs
- Mathematically Involved
- Can be seen as graphs as well

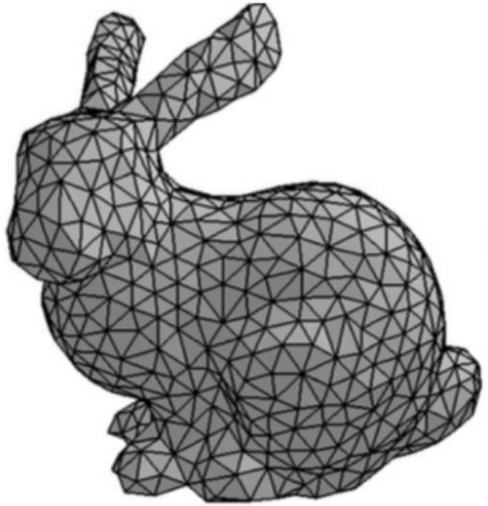


Image courtesy: Hao Su

Choice of Representation - Template-Based (CAD)

- Again, Not Able to Easily Apply Out-of-Box Models on
- Very Hard to Obtain Data
- Hard to Model Shapes Never Seen Before
- Offers Compositionality Intrinsically and Explicitly
  - Might be a Good Option for Learning Functionalities

Image courtesy: Haibin Huang

Choice of Representation - Why Multi-View?

- Multi-view representation is very lightweight
- Offers Flexibility (Depth Maps) and Eases the Computation Significantly
- Although 2D, Still Can Explicitly Model 3D Shapes
- Allows Generating Hi-Res, Detailed, Novel Objects
  - Without the machinery required for new voxel-based models
- Can easily apply out-of-the-box CNN models
- Not Mathematically Involved
- 2D images are widely available
  - No Doubt that it is Very Easy to Obtain 2D images or RGBD or just D

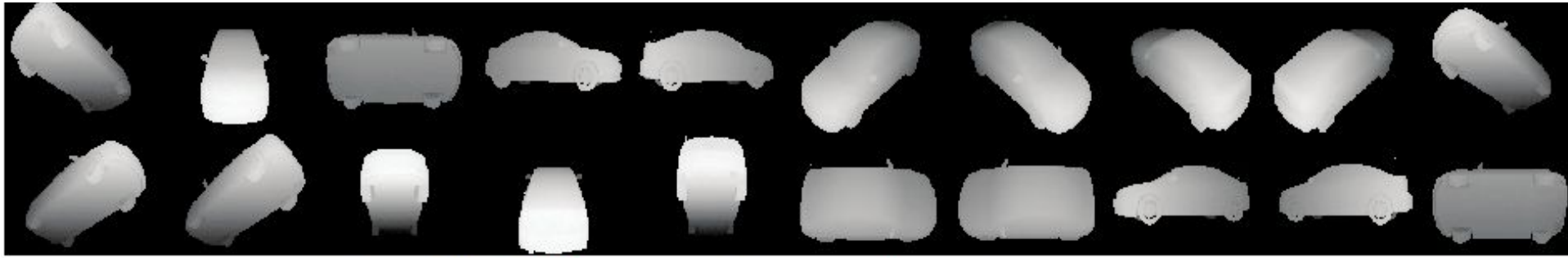
Choice of Representation - Multi-View

Our goals:

- Synthesize/Generate Hi-Res, Detailed and Novel Shapes
- Have Out-of-Sample Generalizability
  - A step forward towards obtaining 3D priors efficiently to solve inverse vision problems
- Obtain 3D representation from 2D
  - ([loosely] inspired from biological vision)
- Share the Same Representations For All Categories
- Combine generative and discriminative models
  - Obtaining hierarchical priors from the discriminator to ease learning new shapes
  - Not just building a deterministic mapping function
- Obtain explicit and consistent 3D representations
  - Not an implicit model for learning about 3D shapes (Zhou et al, Dosovitskiy et al)

Pipeline - Data Set

- Used ShapeNet Core
  - Contains Aligned, Normalized Shapes
  - ~37k for train, ~3k for test
- Render 20 views of depth maps
  - Camera Positions Fixed

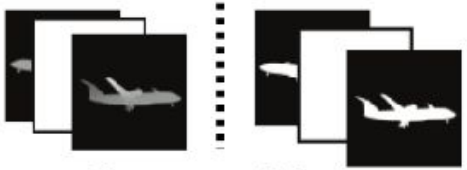


Pipeline - Architectures

- Train 3 Different **VAE** Architectures
  - AllVPNet: Train with All 20 Views
  - DropoutNet: Train with 2-5 Randomly Chosen Views
  - SingleVPNet: Train with 1 Randomly Chosen View
- *Z* Layer Has 100 Nodes for Unconditional and 40 for Conditional
  - 40 nodes to implicitly enforce the network focus on modeling variabilities
- L1 Loss Function is Used During Training

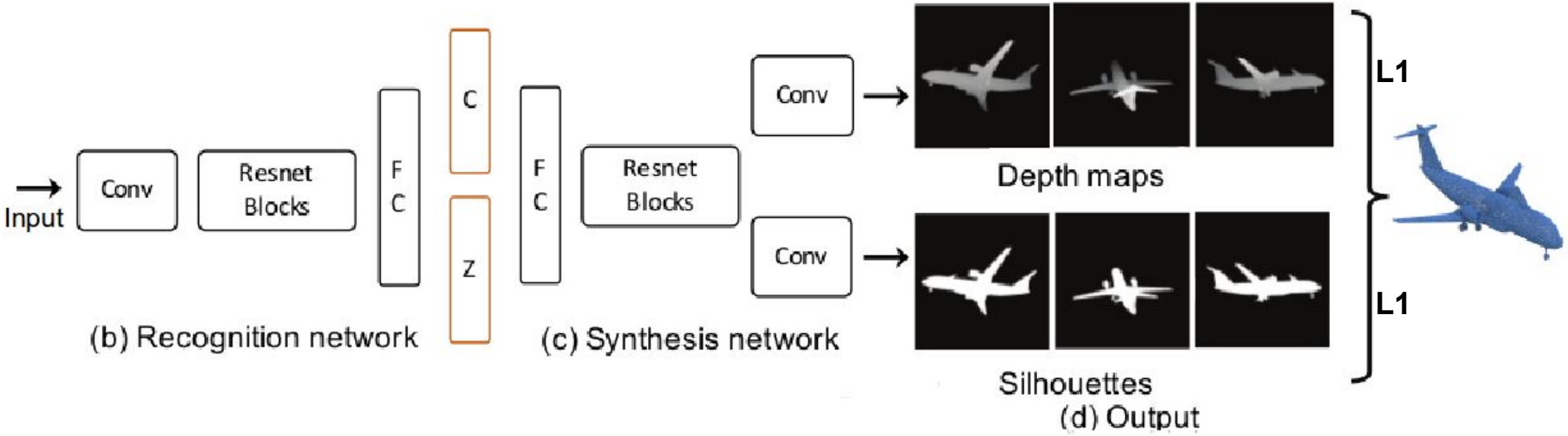
Pipeline - Architectures

AllVPNet

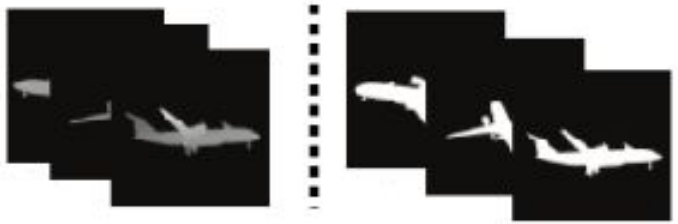


DropoutNet

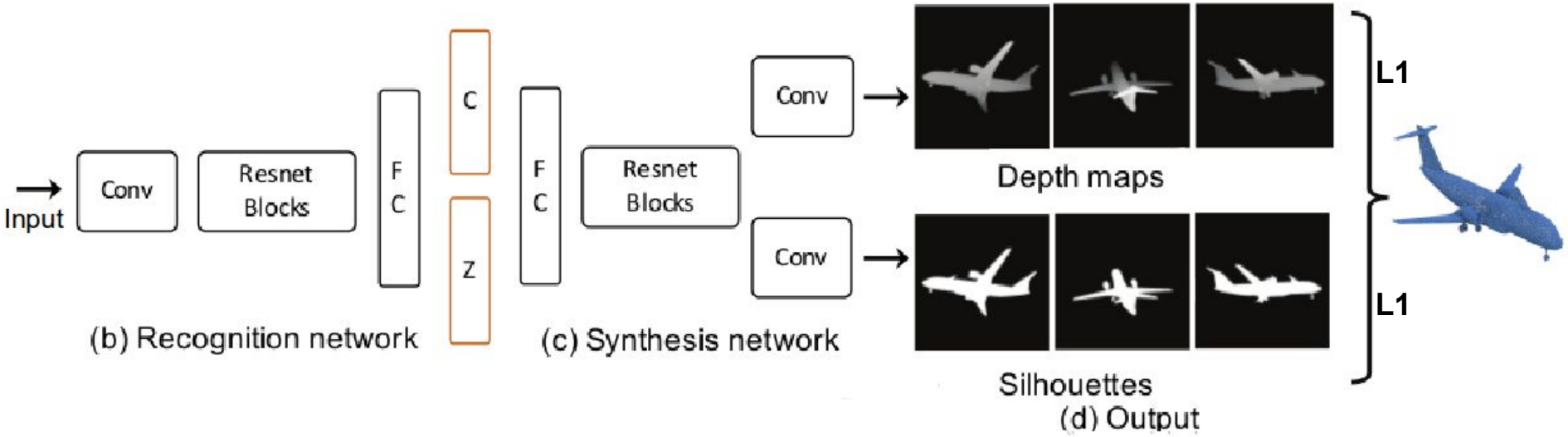
SingleVPNet



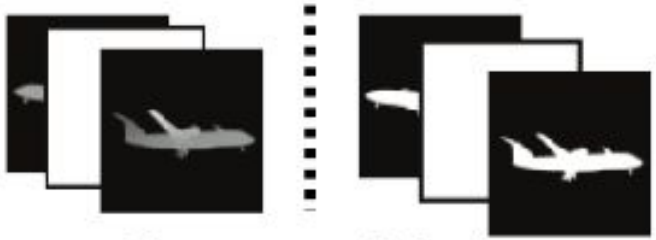
Pipeline - Architectures



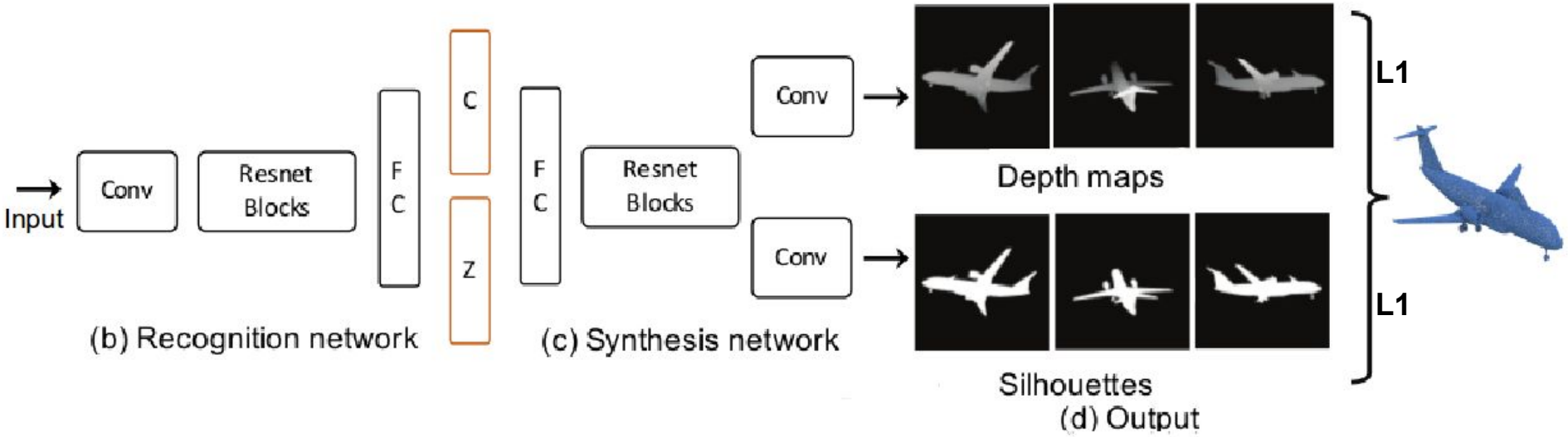
AllVPNet



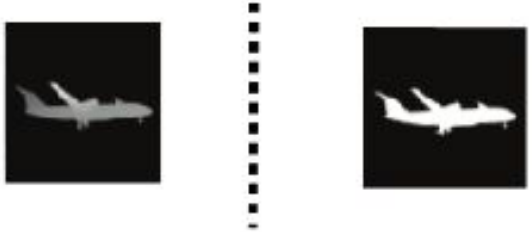
Pipeline - Architectures



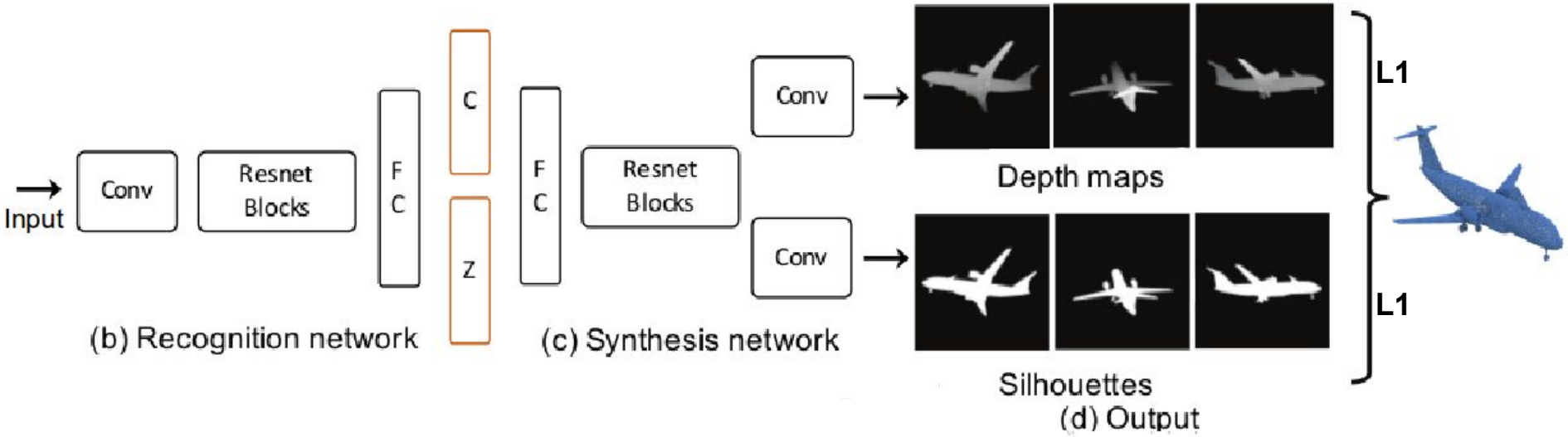
DropoutNet



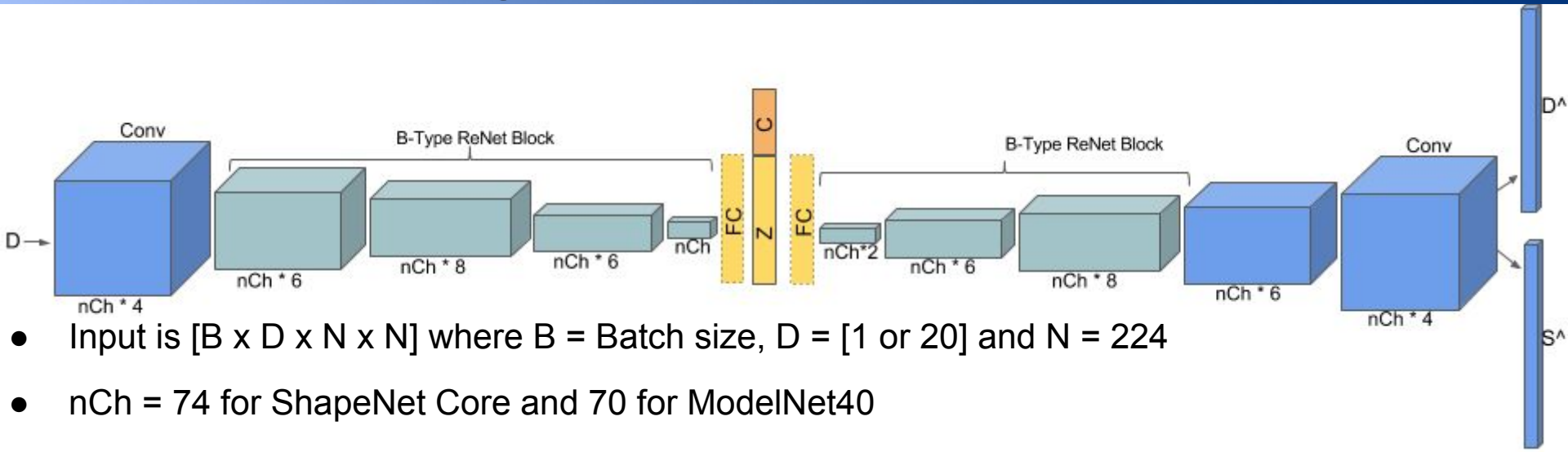
Pipeline - Architectures



SingleVPNet

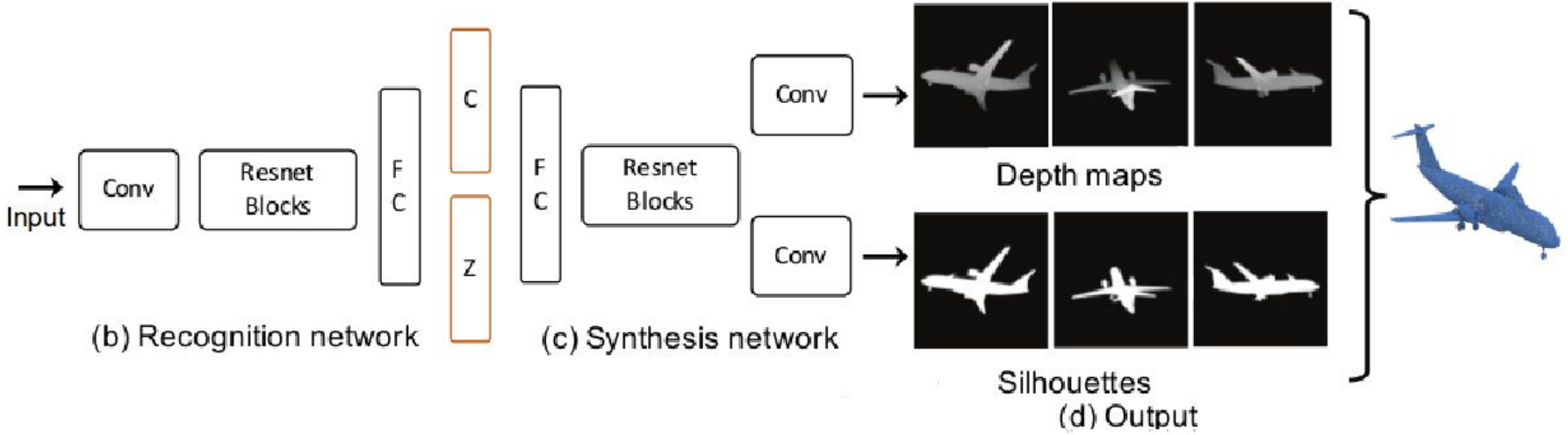


Pipeline - Training



- Input is [B x D x N x N] where B = Batch size, D = [1 or 20] and N = 224
- nCh = 74 for ShapeNet Core and 70 for ModelNet40
- No backpropagation through C from decoder when training conditional models
- Used ADAM for optimization
- Learning rate starts at 8.5 x 10^-5 and anneals at ~0.98
  - Learning rate is multiplied by 0.35 on epoch 18 and each 6 epochs until epoch 30
- Batch size is 4
  - increase to 6 on epoch 20 and 8 on epoch 40
- KLD gradient magnitude is 220, 180 and 130 for AllVPNet, DropoutNet and SingleVPNet

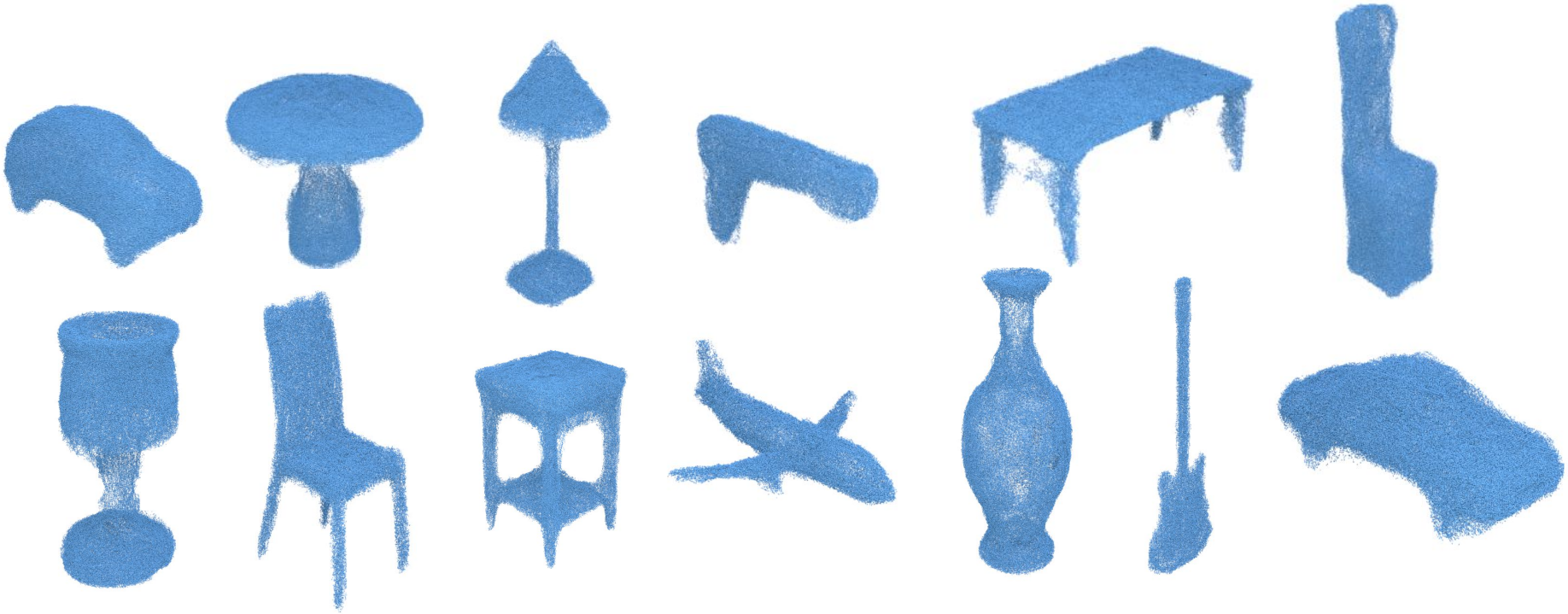
Pipeline - 3D Reconstruction



- Deterministic Function is Used to Generate the Final 3D Point Cloud
- Number of Points is Between ~30k to ~400k depending on Shape Complexity
  - Not fixed
- Used empirical distribution for sampling
  - both unconditional and conditional
- Code available at **github.com/Amir-Arsalan/Synthesize3DviaDepthOrSil**

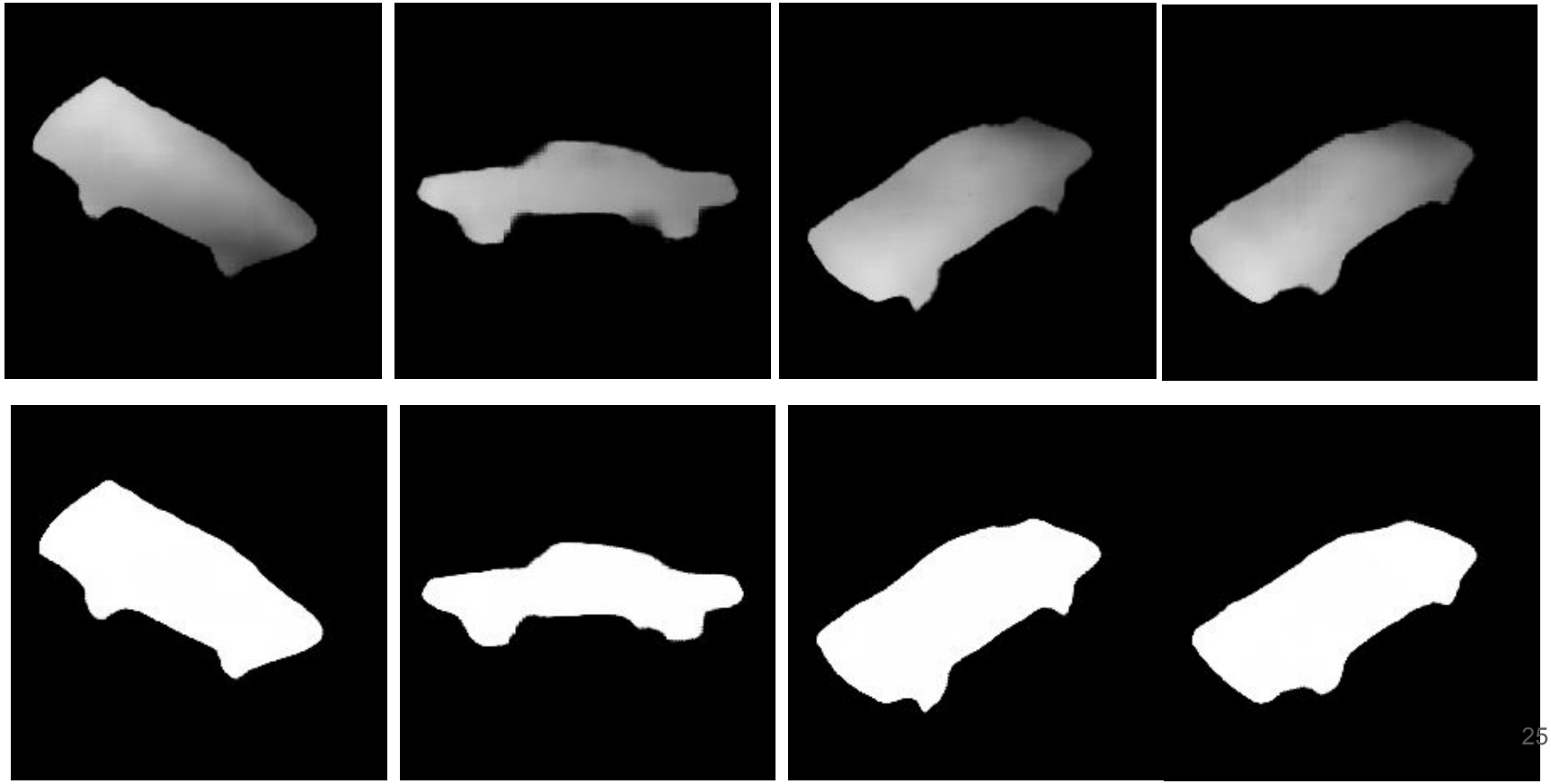
Results - Sampling

Random Sampling



Results - Random Sampling

Random Samples - 4 Depth Maps and Silhouettes



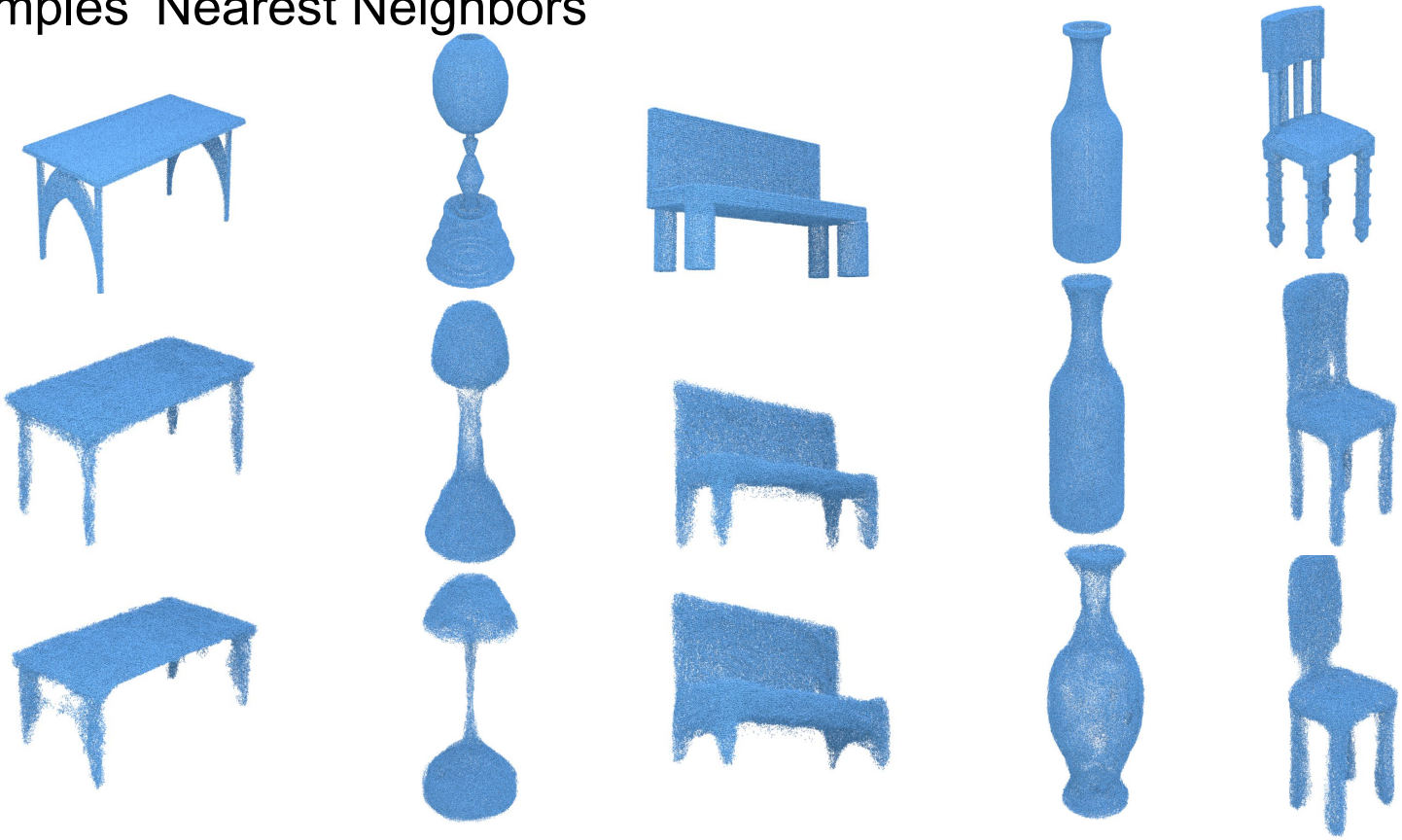
Results - Random Sampling

Random Samples' Nearest Neighbors

Training set

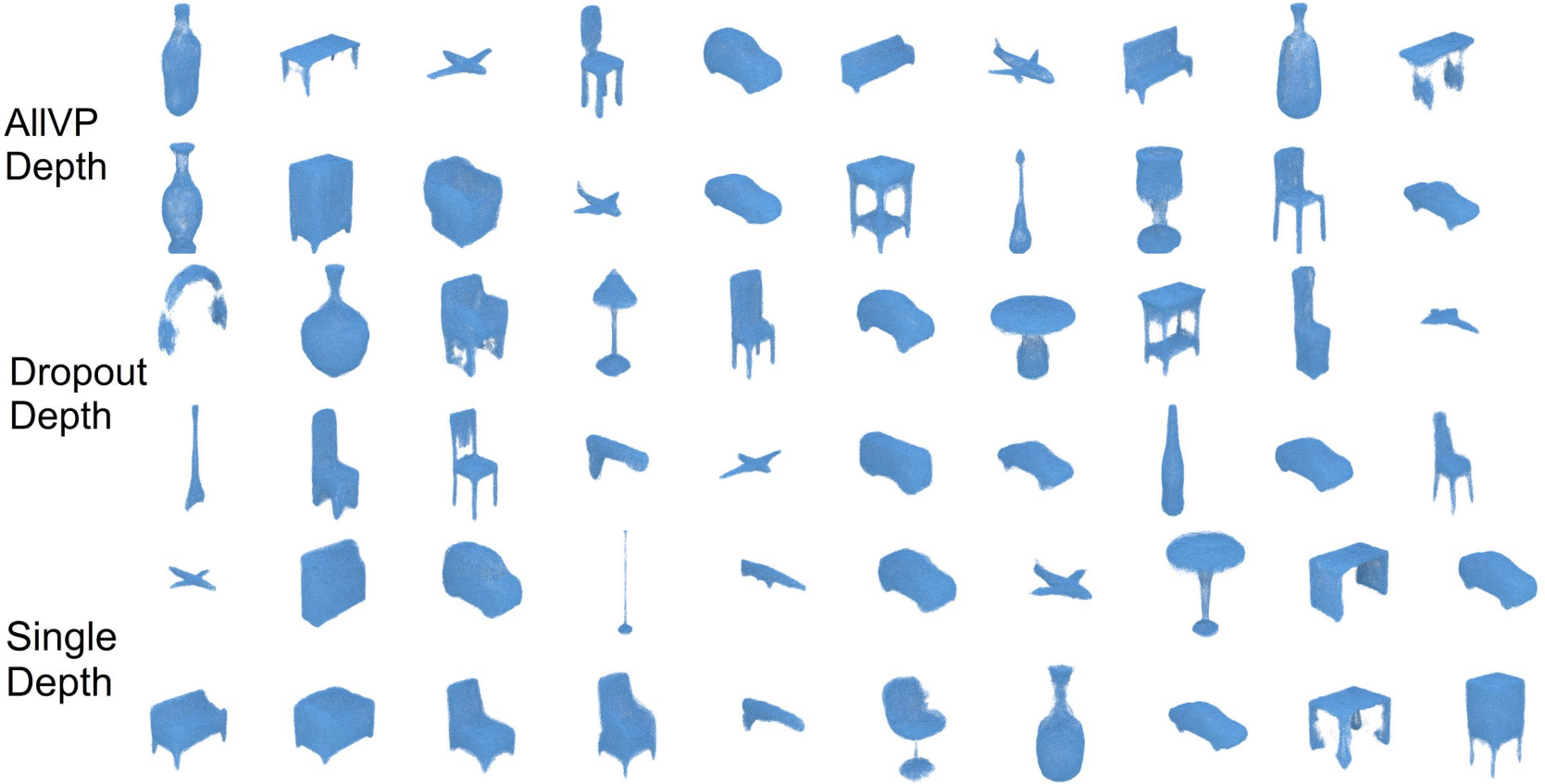
Training Sample Reconstruction

Random Sample



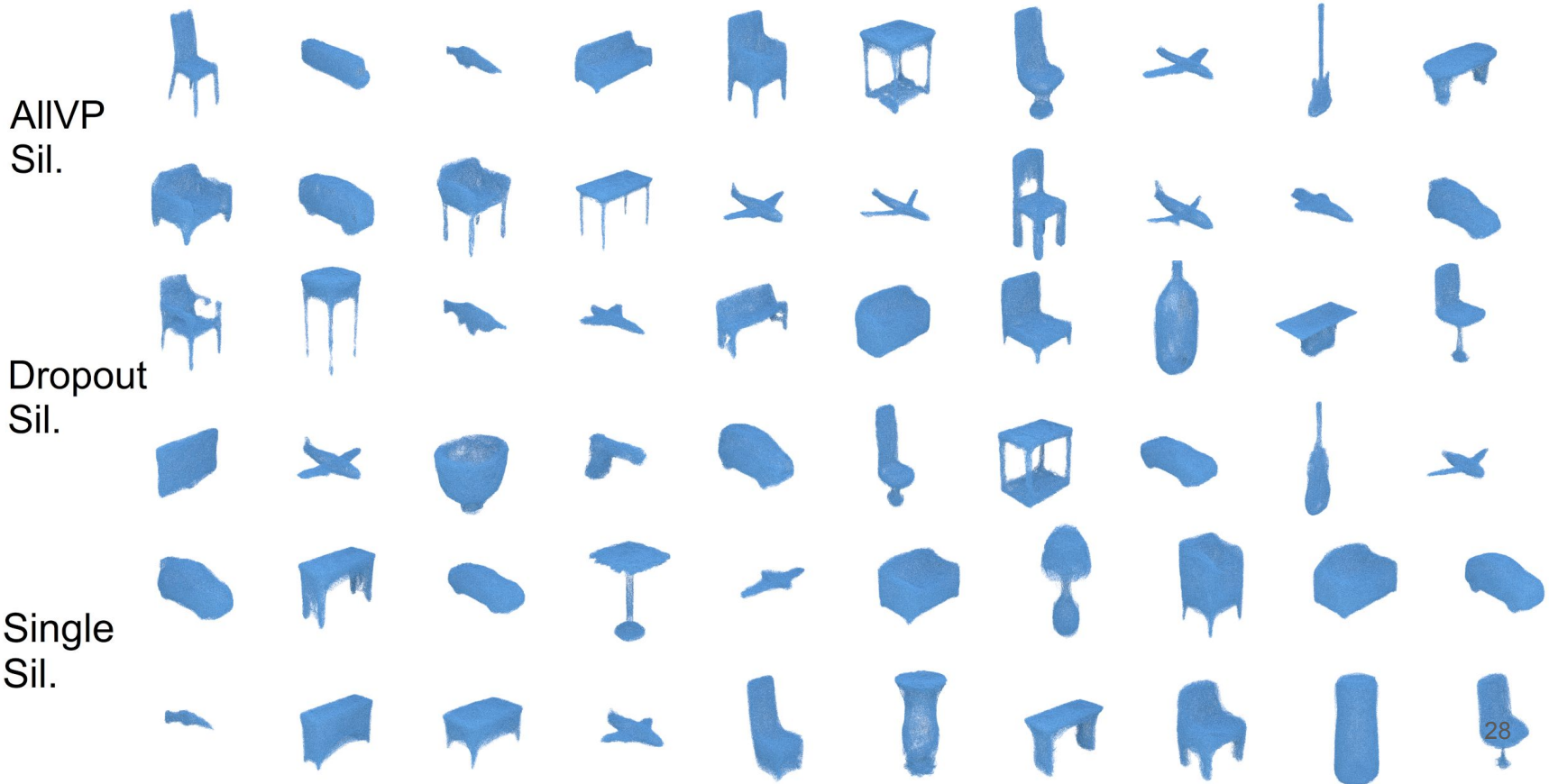
Results - Random Sampling

Not biased samples despite having ~22k shapes from 5 categories in training set



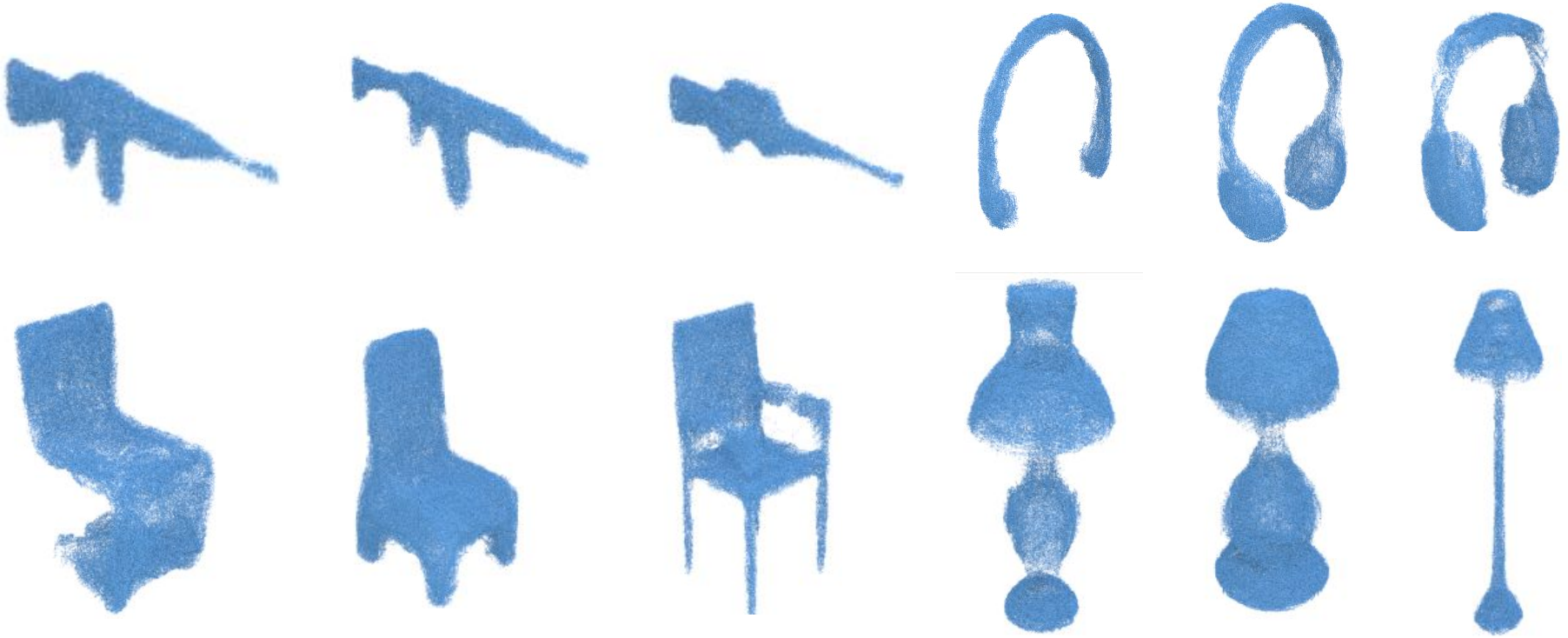
Results - Random Sampling

More Random Samples from Networks Trained on Silhouettes



Results - Conditional Sampling

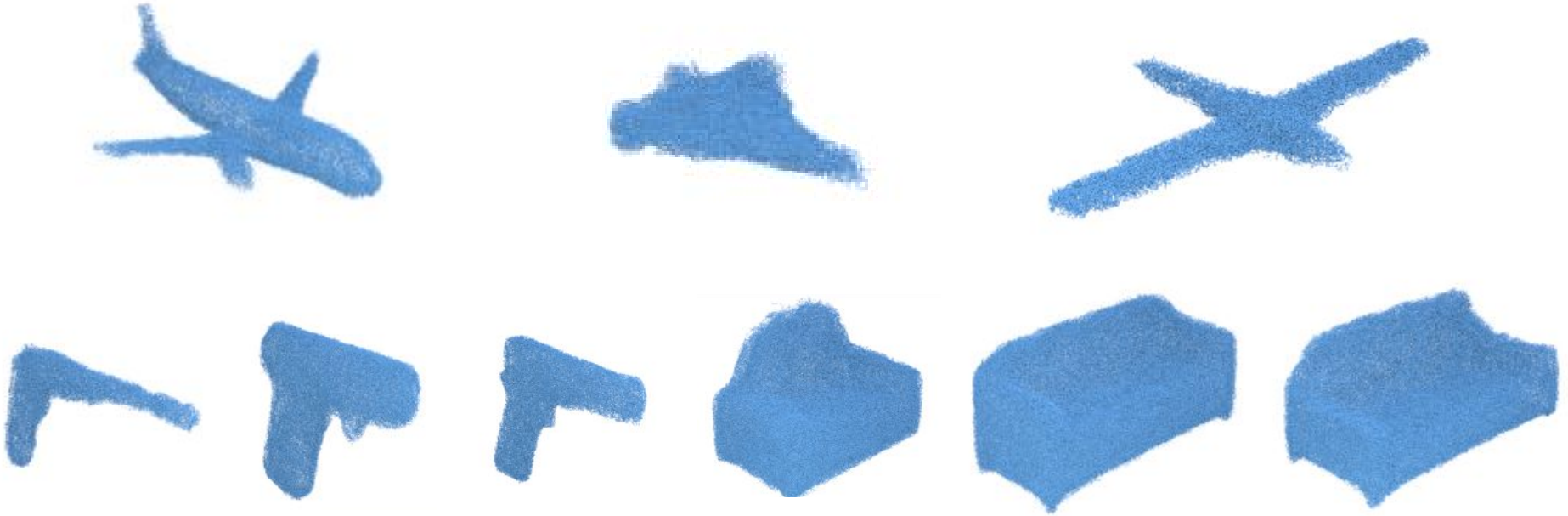
Good “Headphone” samples despite having less than 60 samples in training set



Results - Conditional Sampling

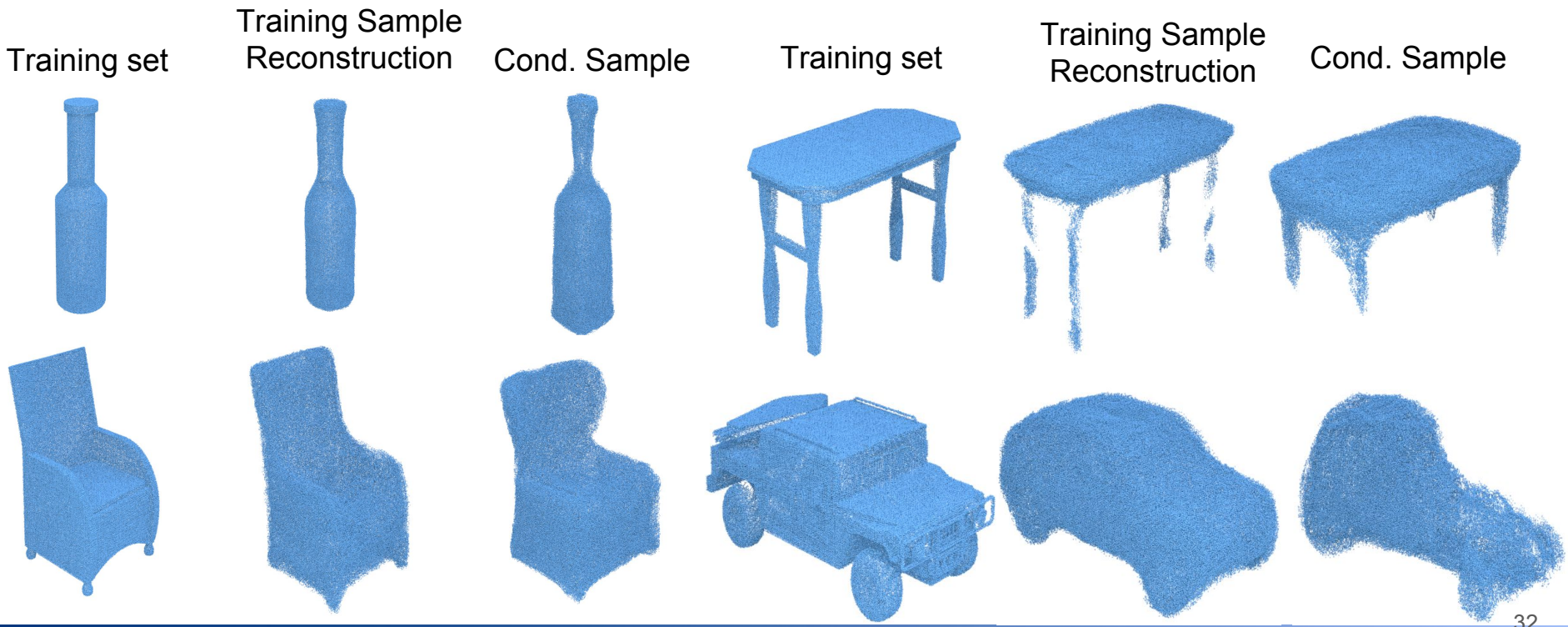
Results - Conditional Sampling

More Conditional Samples



Results - Conditional Sampling

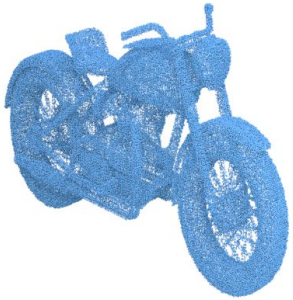
Nearest Neighbors



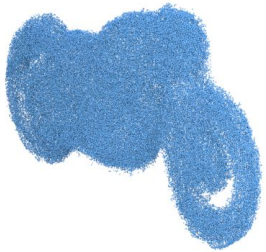
Results - Conditional Sampling

Obtaining Good-Looking Samples In Low-Data Regimes

Training set



Training Sample Reconstruction



Cond. Sample

Results - 2D → 3D Reconstruction

Results - Classification

Classification, Reconstruction Error

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	$65.3\% \pm 0.3$

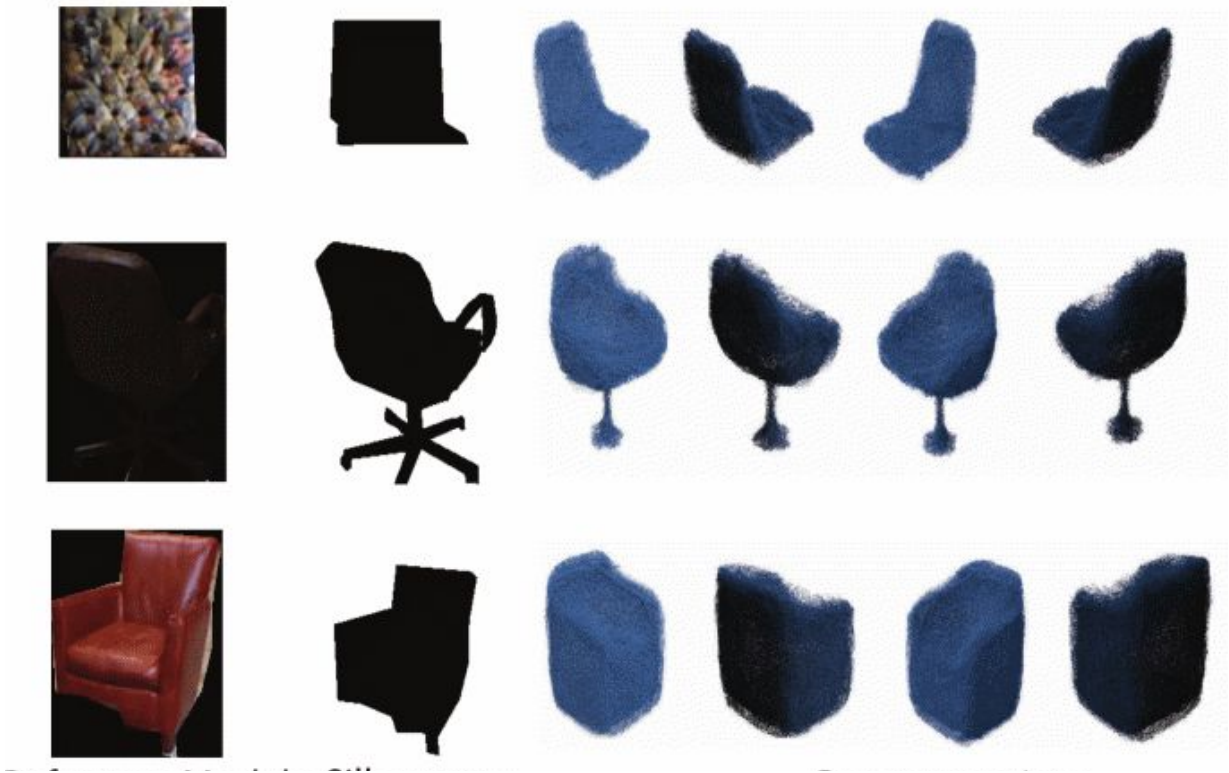
Results - Reconstruction

Out-of-Sample Generalization:

- Only Used **SingleVPNet** Models
- Put Silhouettes/Depth Maps into 224 x 224 canvases
- Images Scaled to Fit
- Camera Pose Not Fixed Anymore
- Different Size and Orientation
- NYUD and Silhouettes from the Internet

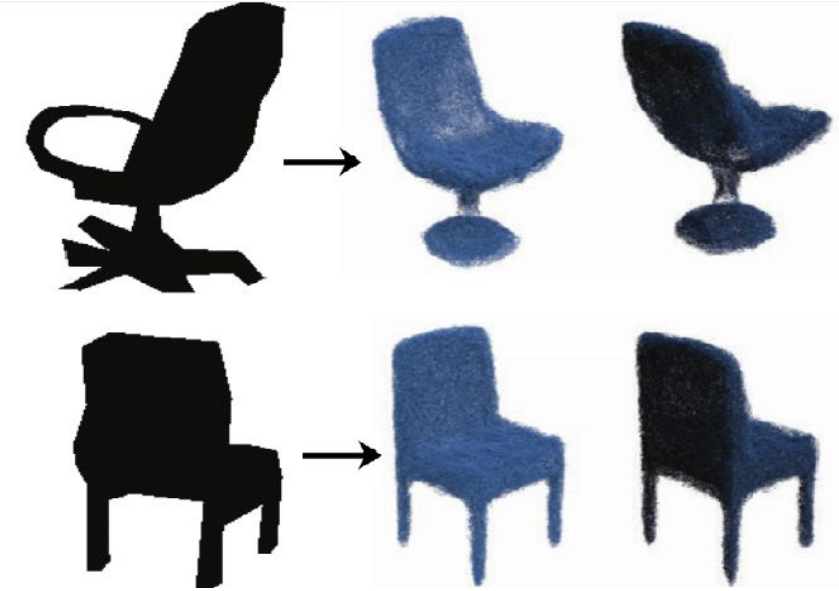
Results - Reconstruction

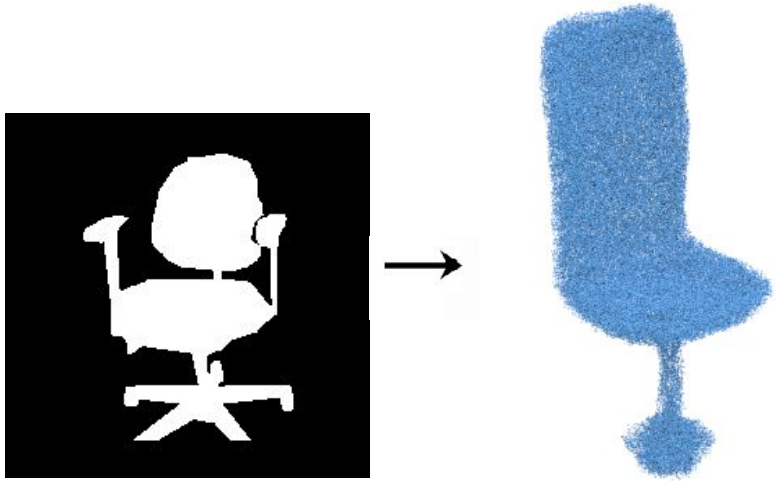
Out-of-Sample Generalization (NYUD)



Results - Reconstruction

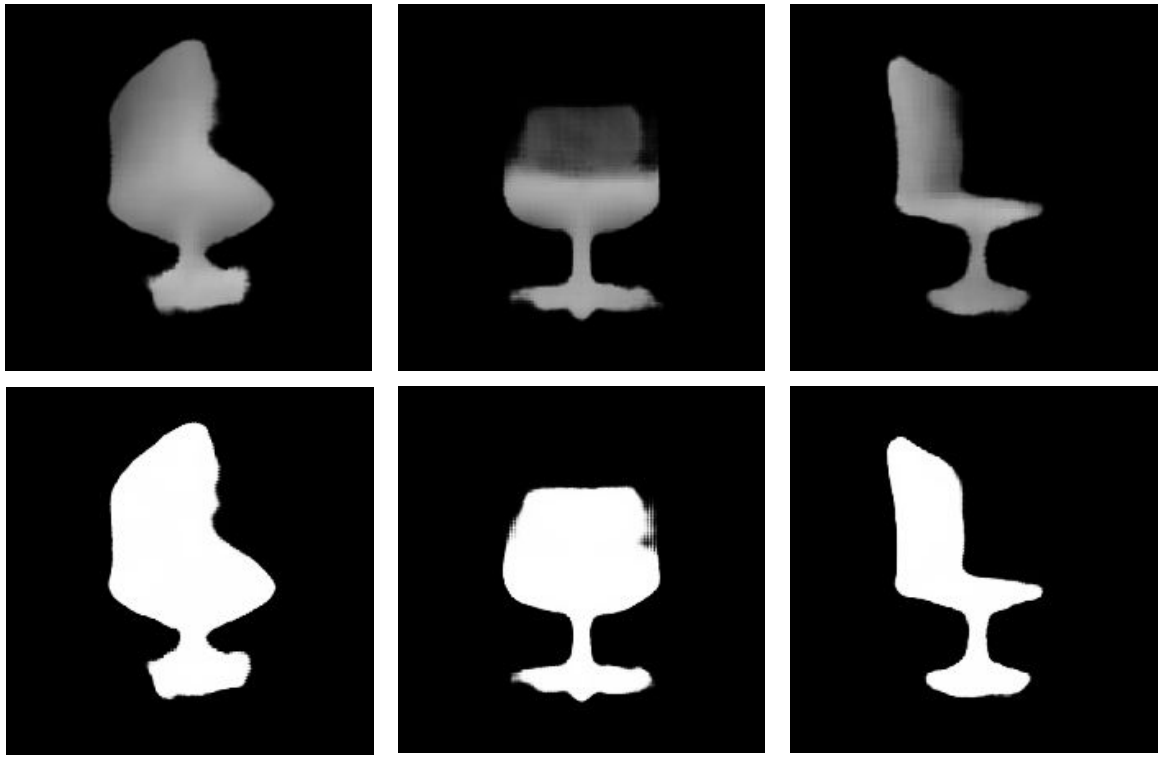
Out-of-Sample Generalization (Uncond. SinlgeVPNet - NYUD Silhouettes)





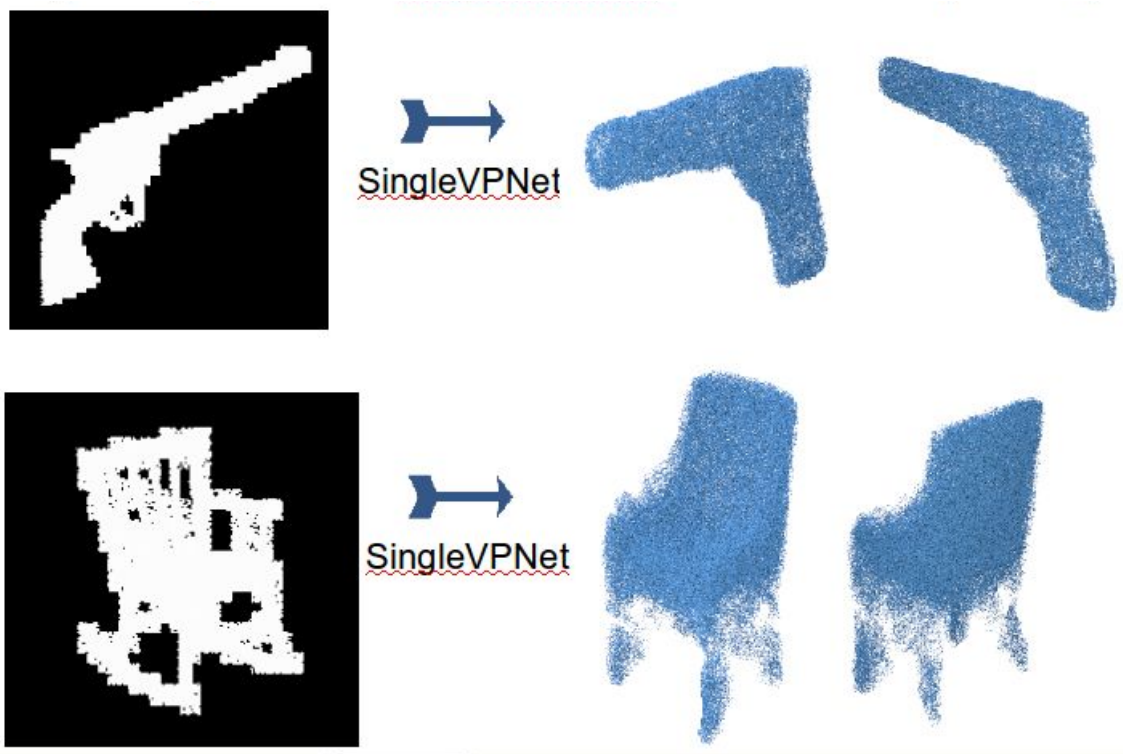
Results - Reconstruction

Out-of-Sample Generalization (Uncond. SinlgeVPNet - NYUD Silhouettes)



Results - Reconstruction

Out-of-Sample Generalization (Silhouettes From Web)



Results - Representation Analysis

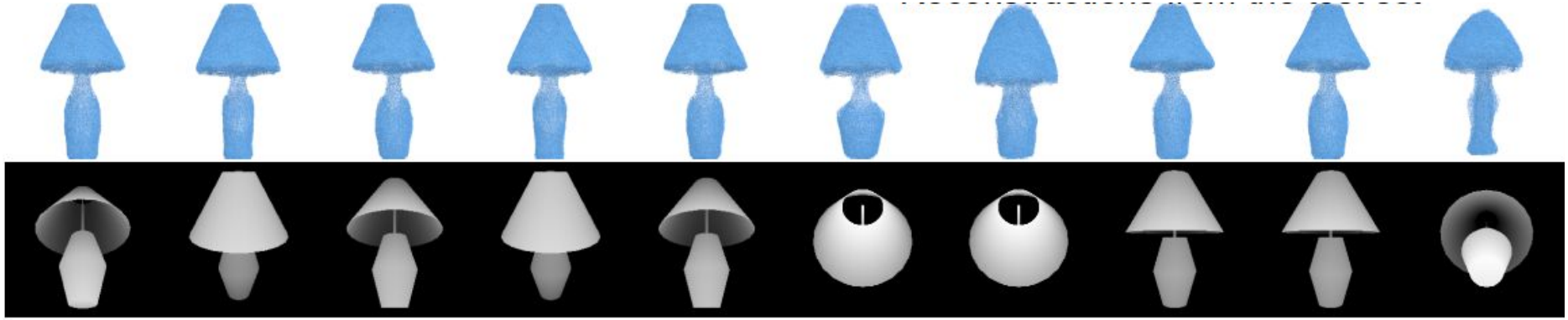
Consistent Representation

- Naturally Would Like to Get The Same Shape Across All Views
- Intuitively-Thinking, Uncertainty is Actually Part of Consistency
  - Not possible to obtain the same shape from ambiguous views
- Obtaining Good Priors Is Important!

Results - Analysis

Consistent Representation (1st row: Class. Acc., 2nd row: IoU)

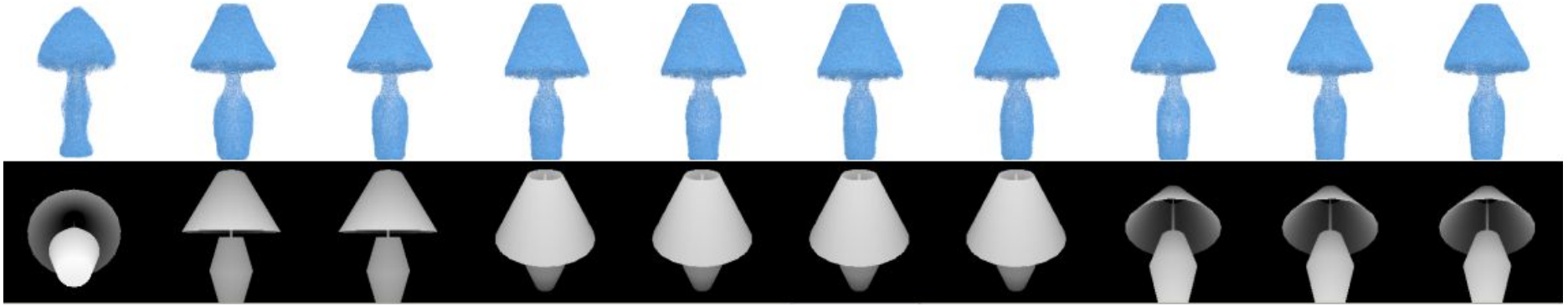
1	2	3	4	5	6	7	8	9	10
83.5	84.0	82.1	81.5	80.0	80.7	80.6	85.3	83.8	79.4
73.3	70.2	69.4	69.7	68.8	67.2	67.4	736	73.2	68.0



Results - Analysis

Consistent Representation (1st row: Class. Acc., 2nd row: IoU)

11	12	13	14	15	16	17	18	19	20
79.1	85.1	83.7	85.7	85.5	84.8	85.0	84.7	85.9	84.1
67.9	73.0	73.0	73.6	73.2	73.4	73.6	73.3	73.4	73.3

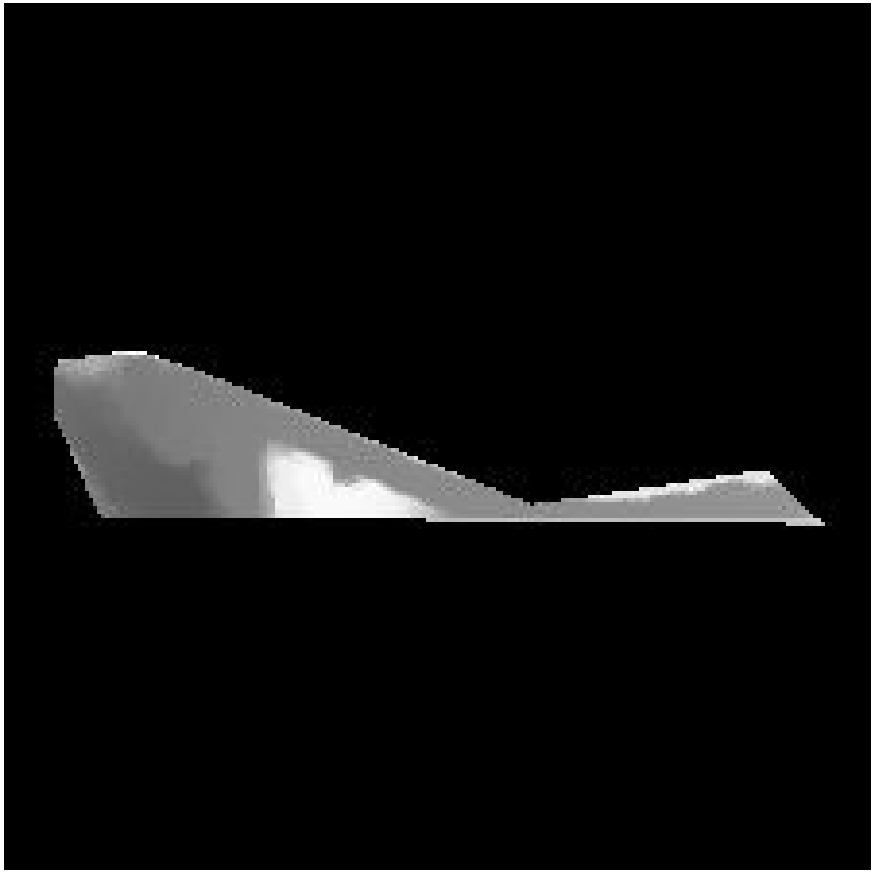


Results - Analysis

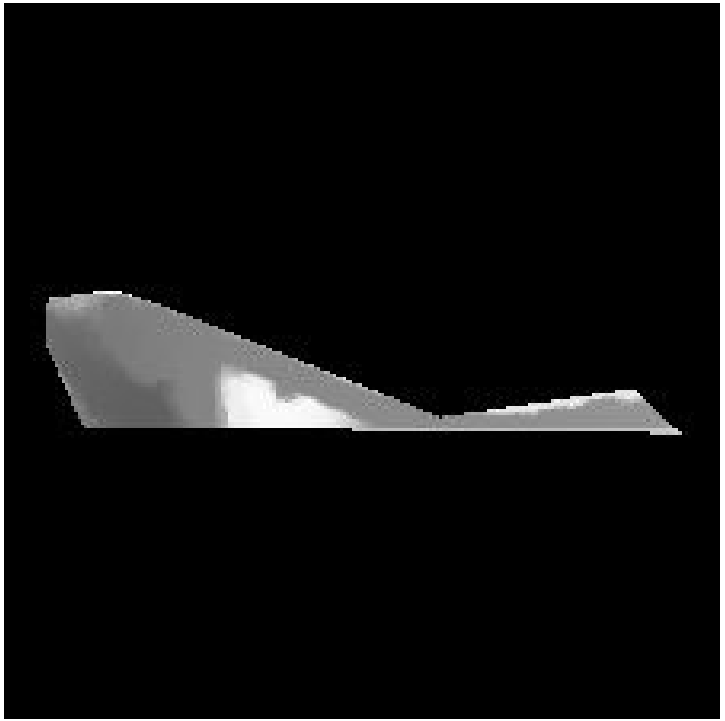
Priors Matter!

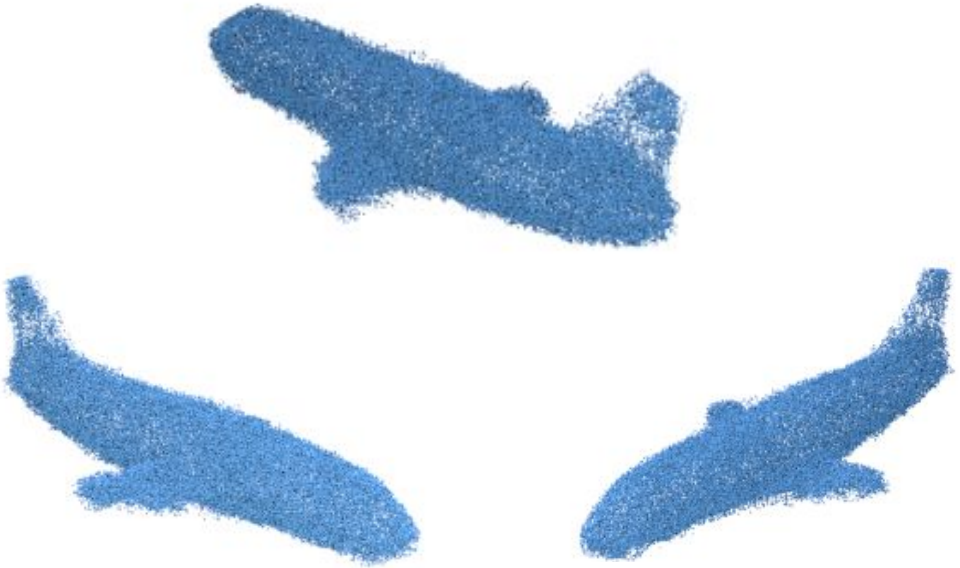
Results - Analysis

What 3D shape is this?



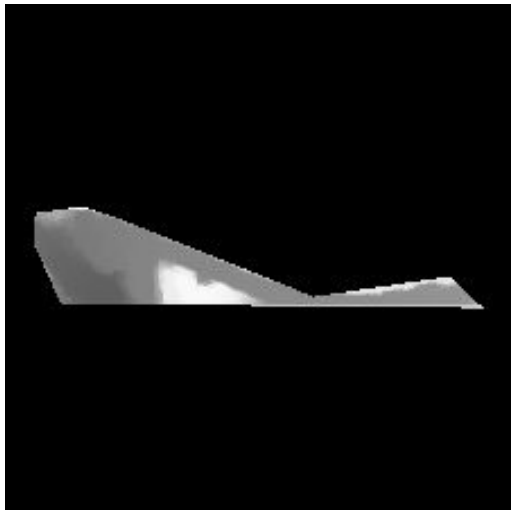
Results - Analysis

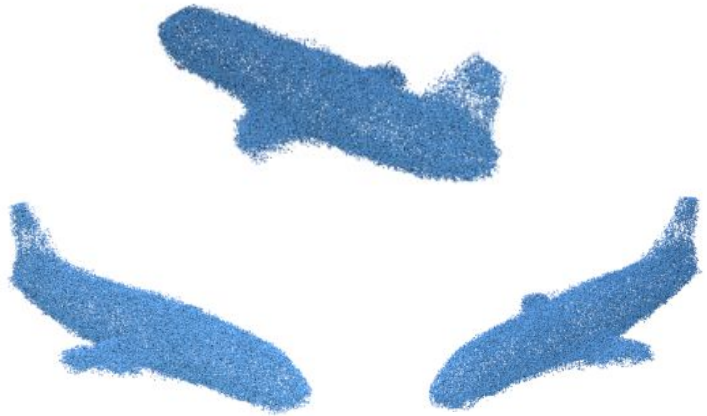




Results - Analysis

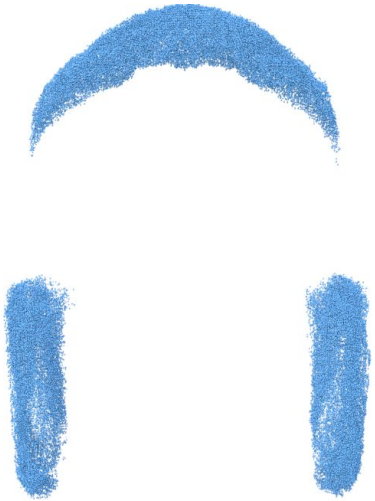
- Model's Category Prediction: "**airplane**"
- Quite meaningful and intuitive
- Obtaining good, inductive biases is hard but helps a lot for learning!
- The prior can be viewed as a hierarchical prior over Z

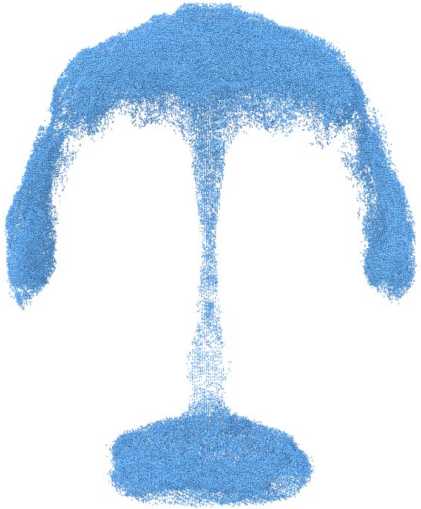


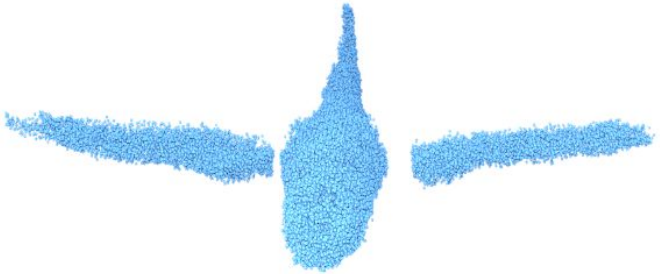


Results - Analysis

Implicitly Learning About Parts

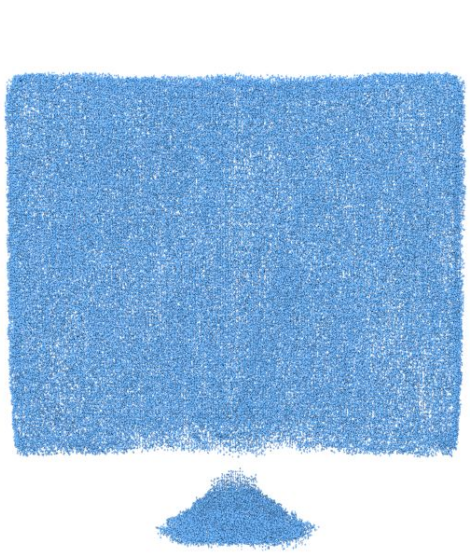


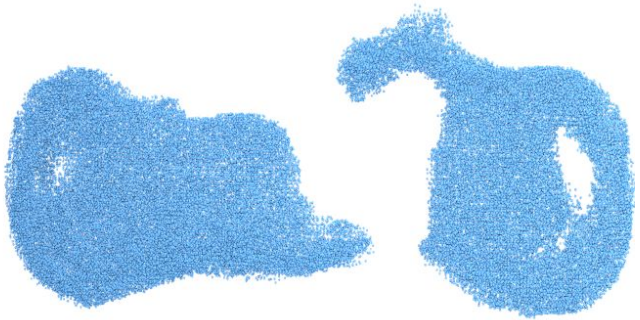


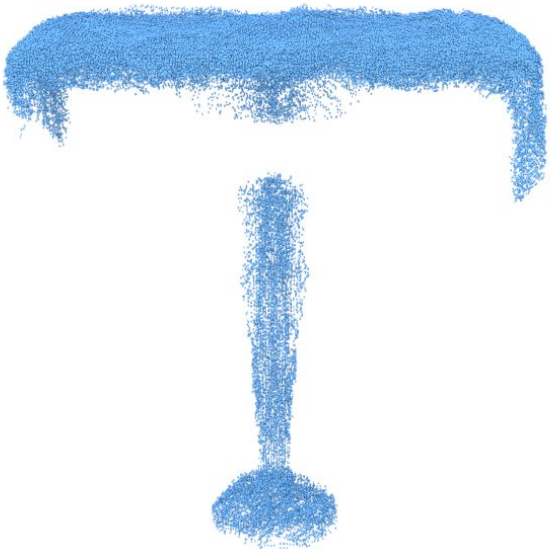


Results - Analysis

Implicitly Learning About Parts







Conclusion

- We showed an effective paradigm for learning 3D shapes using multiview representation
- Samples obtained look realistic, novel and detailed
- Higher resolution that previous works: could be scaled up to even higher
- Out-of-sample generalization is attainable via good generative models + meaningful priors
- Enough inductive bias via hierarchical priors effectively helps in low data regimes
- Consistency in representation is appealing and needed for interesting 3D vision
- Parts can be learned implicitly. It is hard to explicitly learn parts for real-word tasks
  - Learning where the variabilities come from in data implicitly gives clues for parts

Future Directions and Challenges

- Current data sets are not very effective to learn about 3D vision
  - In some ways similar to 2D data sets: MNIST, CIFAR, ImageNet
- 3D shapes are the end product of a lot of underlying processes, mainly physics
- 3D shapes are composed of things like material, mass, etc
- Affordances play a major role in solving inverse vision and planning problems
- No model as of now learns to build 3D shapes with affordances and physical priors
- Purely data-driven approaches do not get us to where we want to be
- To meaningfully interact with 3D shapes we need to do more! (data sets etc)
- Learning fast, and accurate physics engines might be another good starting point

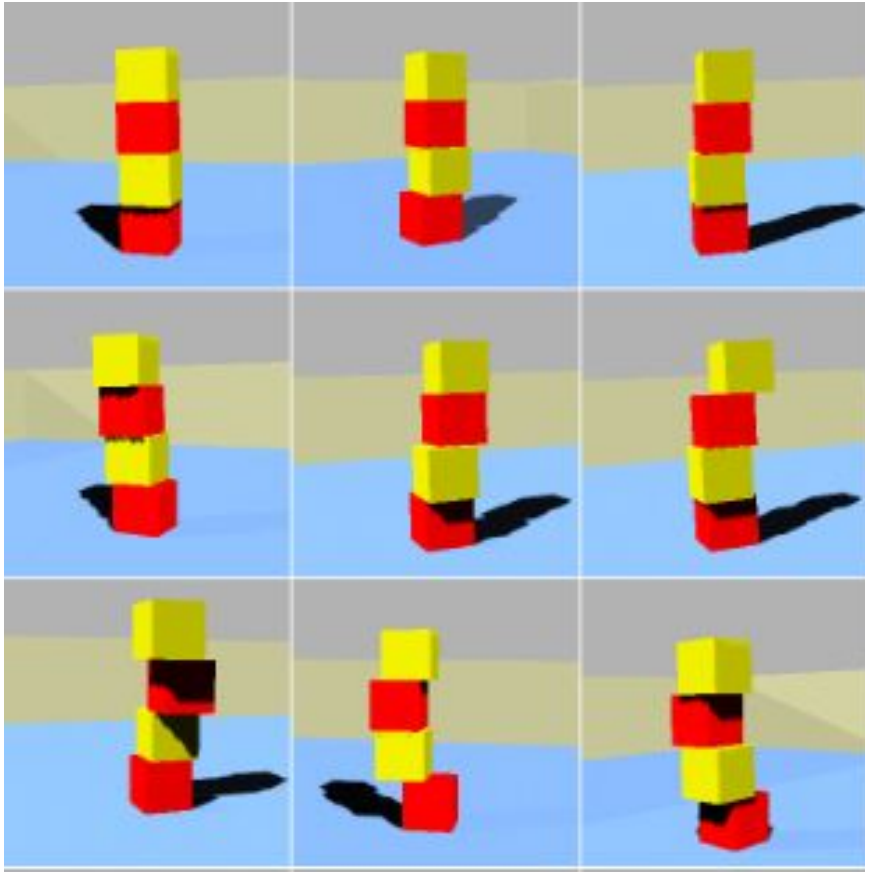
3D Representations for AI Agents

Motivation

- Understanding the world around us through perception for future AI agents
- End goal: building models of the environment
  - A very simple example: learn about shapes and their physical relationships
- Explicit knowledge of 3D objects and their relationships induces [good] priors
- Current RL models do not take the advantage of explicit vision models
- 3D representations can be a good starting point but still not enough
- Physics, affordances, how to manipulate, relations with objs, functionalities
- Eventually helps for inverse vision/planning

Motivation

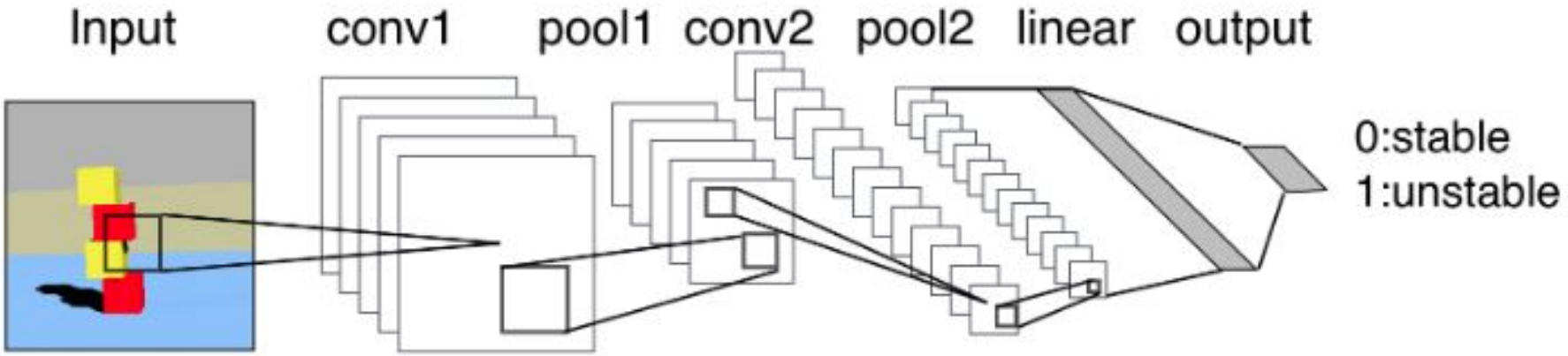
A simple scenario in 2D



Wu et al, CogSci 2016

Motivation

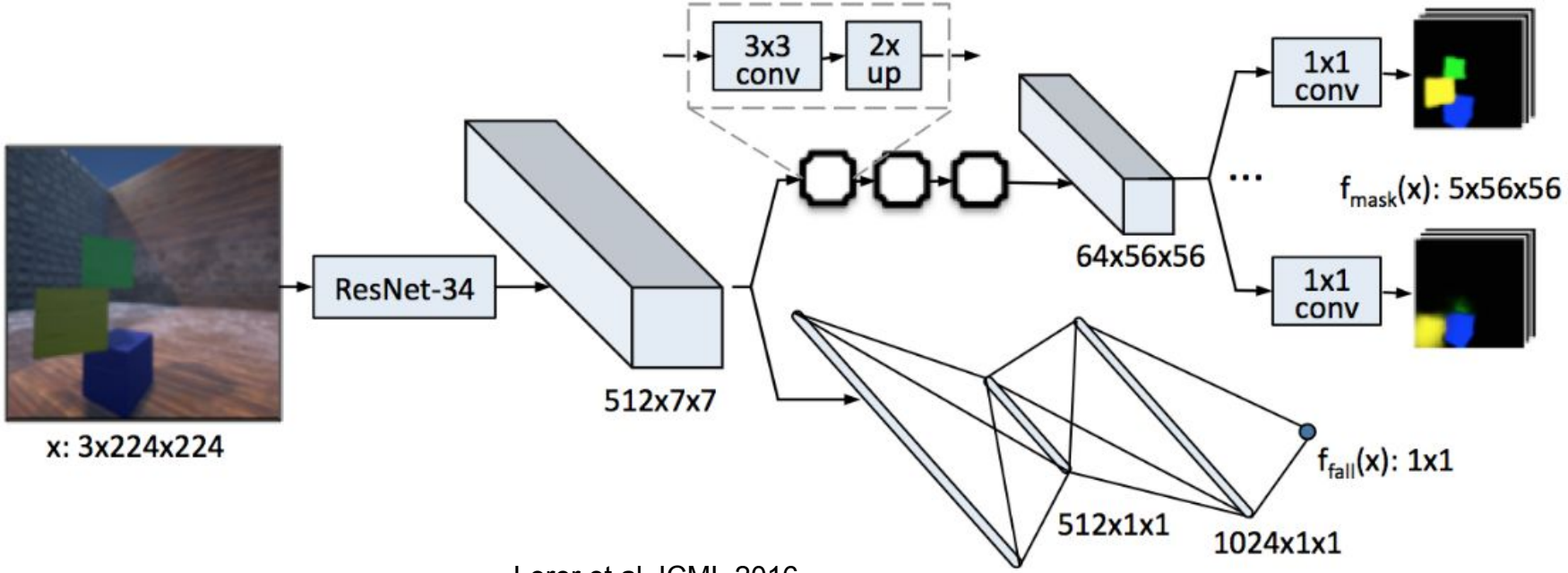
- Building simple mapping functions is not enough



Wu et al, CogSci 2016

Motivation

- Implicit 2D generative models help a little bit more

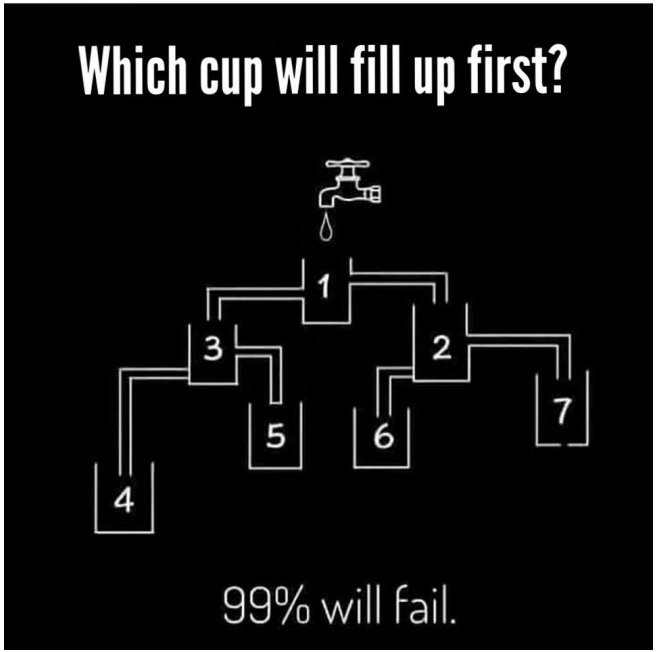


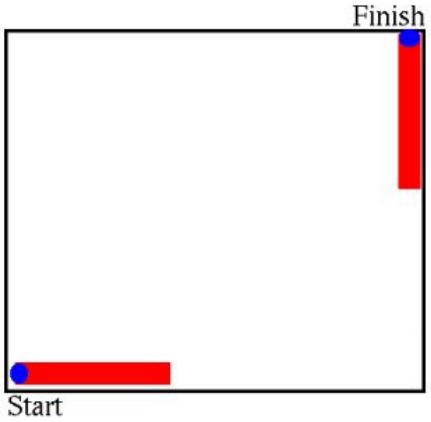
Lerer et al, ICML 2016

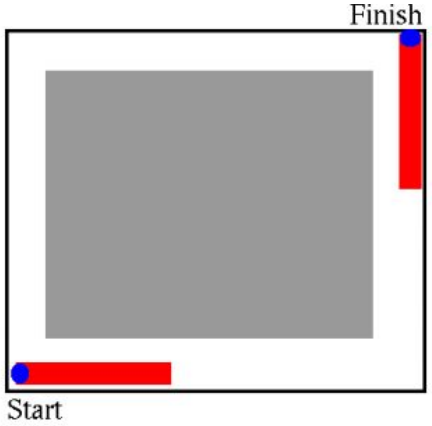
Motivation - Inverse Planning in 2D

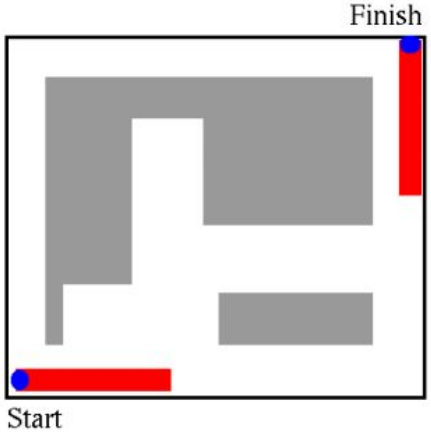
How to avoid physical constraints?

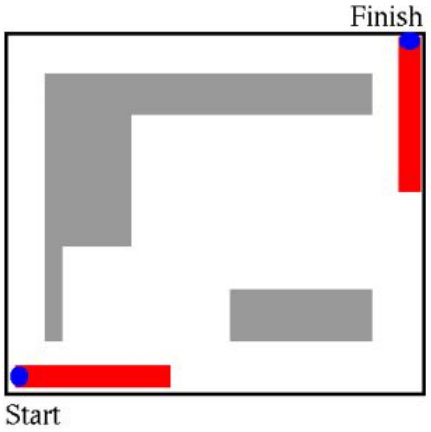
How to use 3D priors for solving 2D tasks?





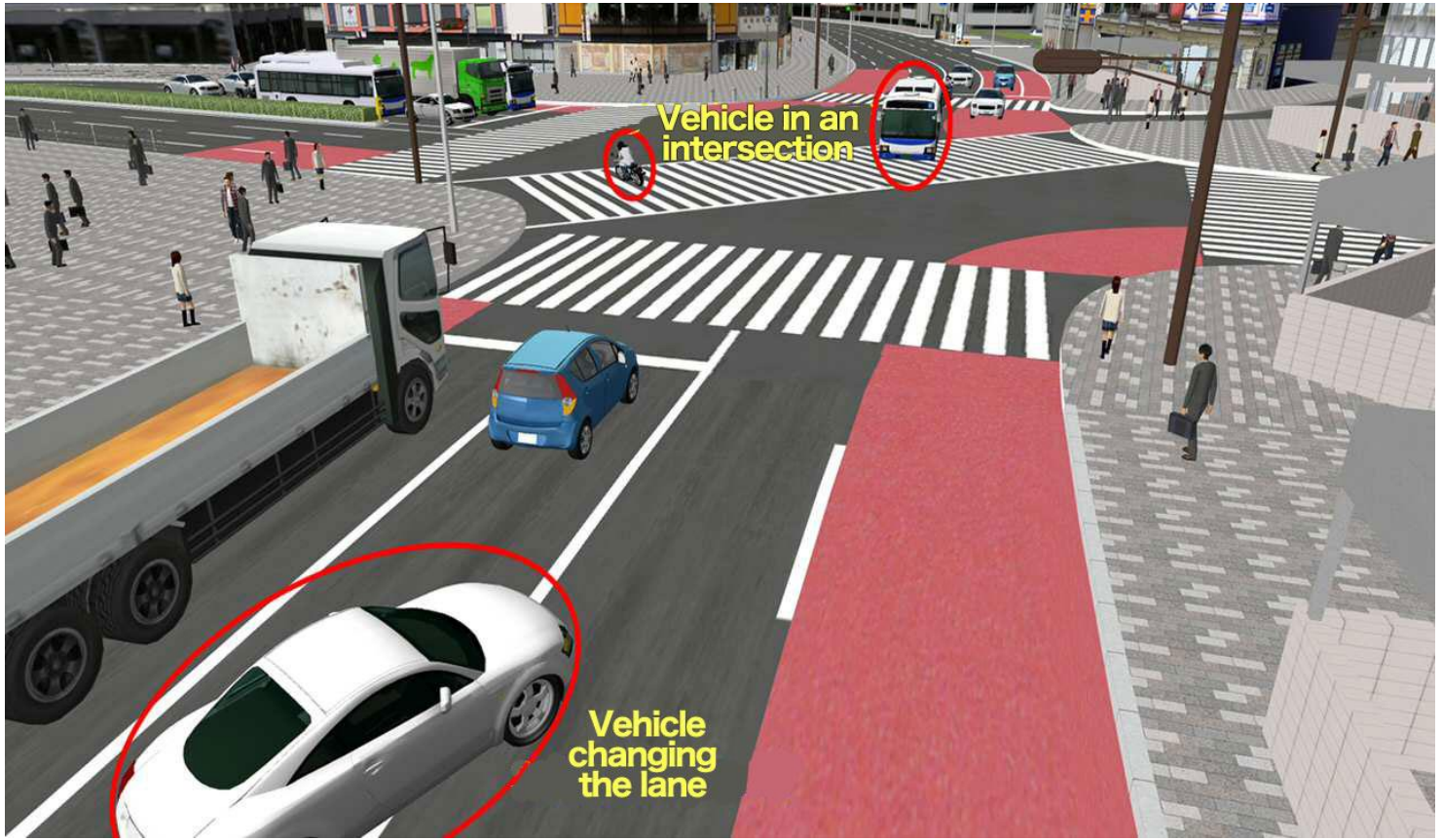




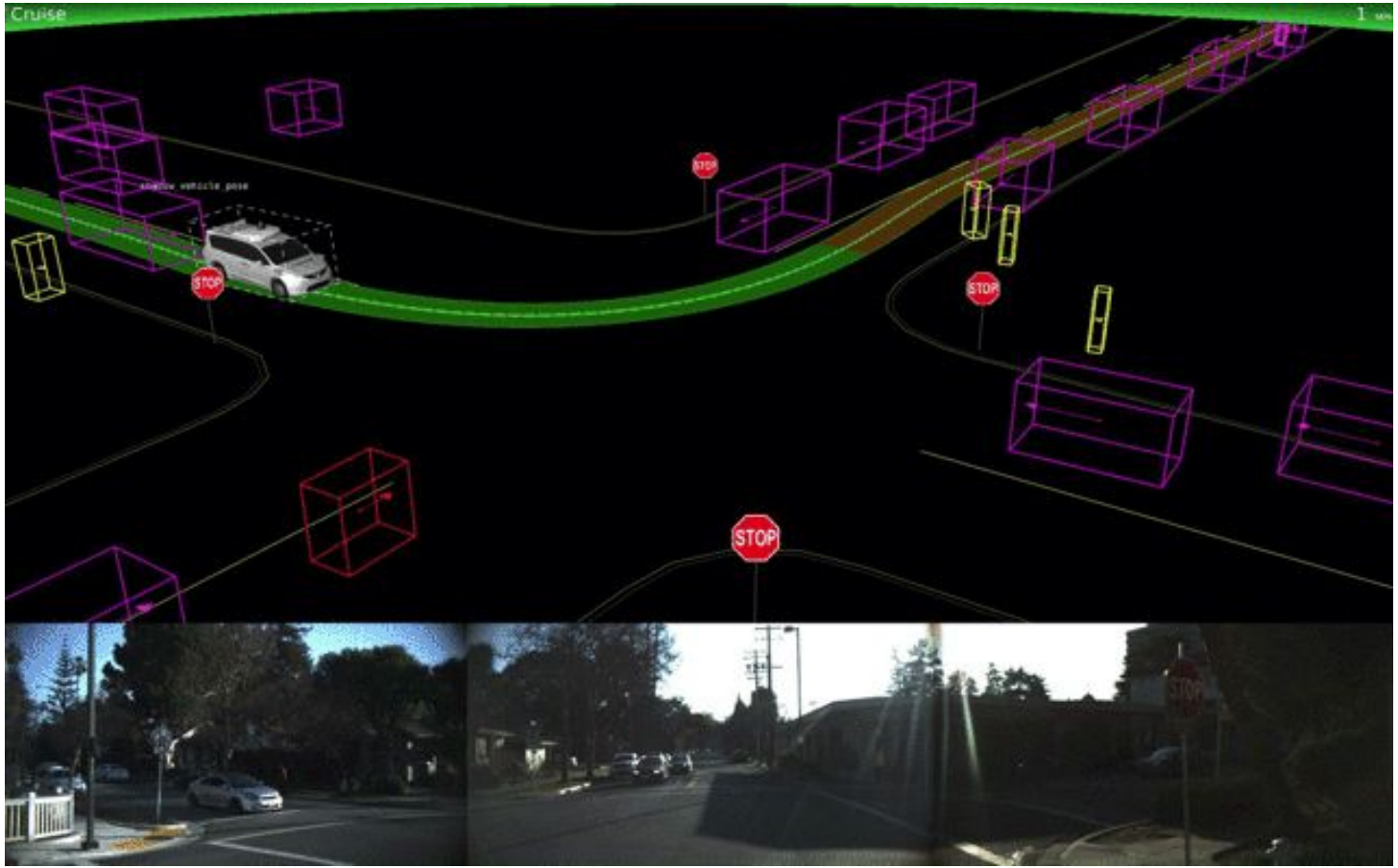


Building 3D Models of the World

Building 3D Models of the World

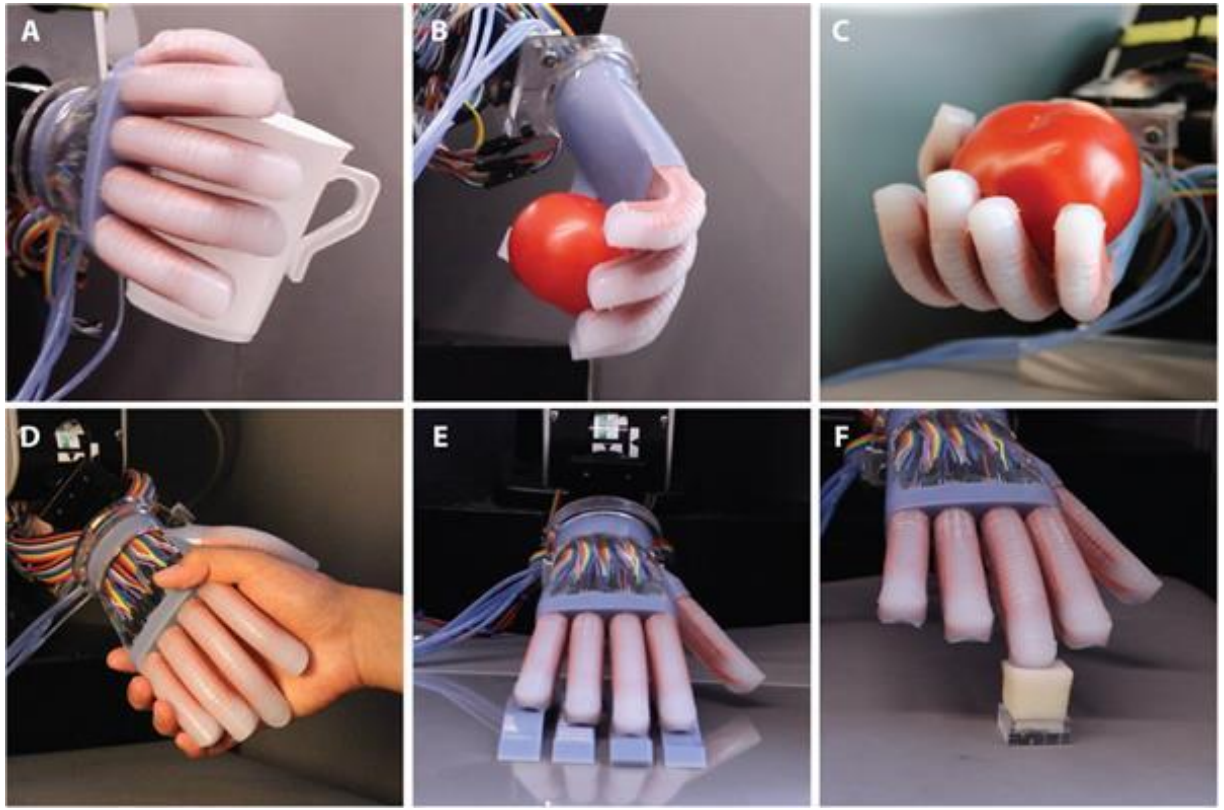


Building 3D Models of the World



Building 3D Models of the World

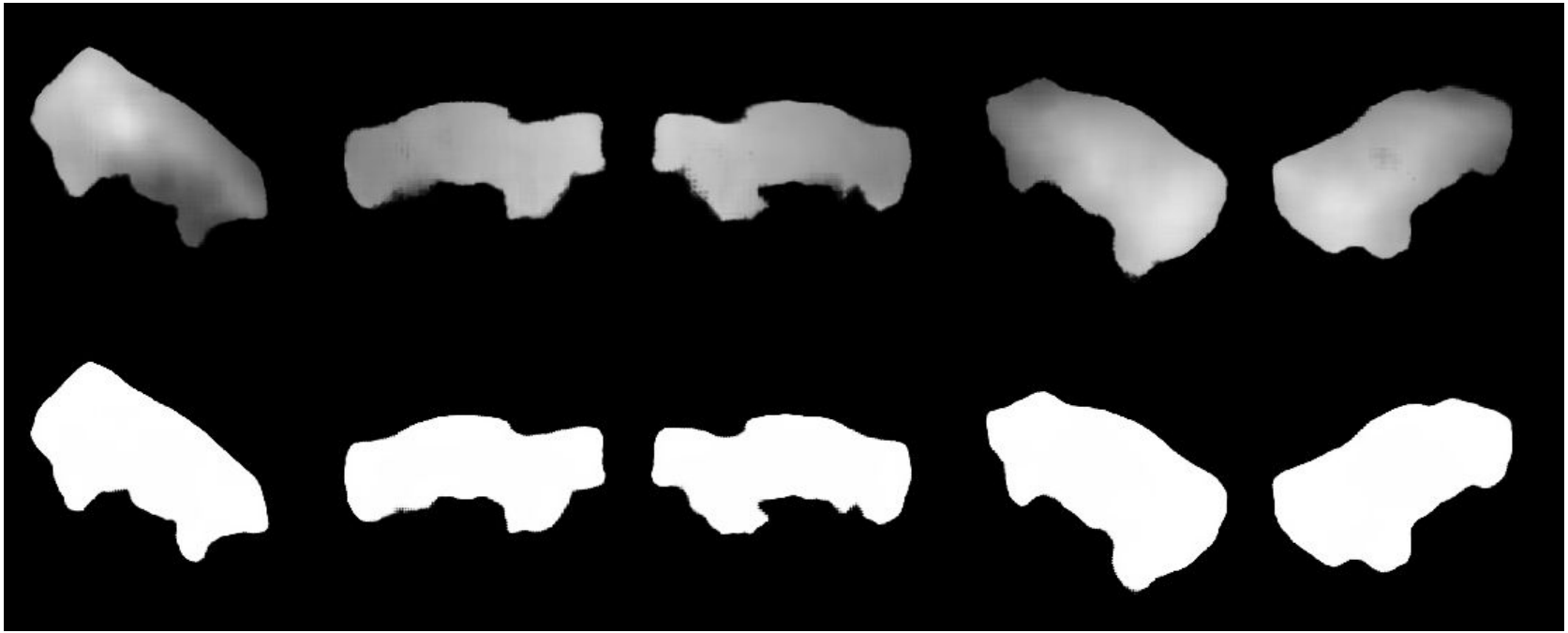
Object manipulation



Thank you!

Results - Conditional Sampling

Conditional Samples



Results - Classification, Recon. Err.

- The Goal Is Not to do Classification or Recon. But to Have Hierarchical Priors
- Strong Regularization

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

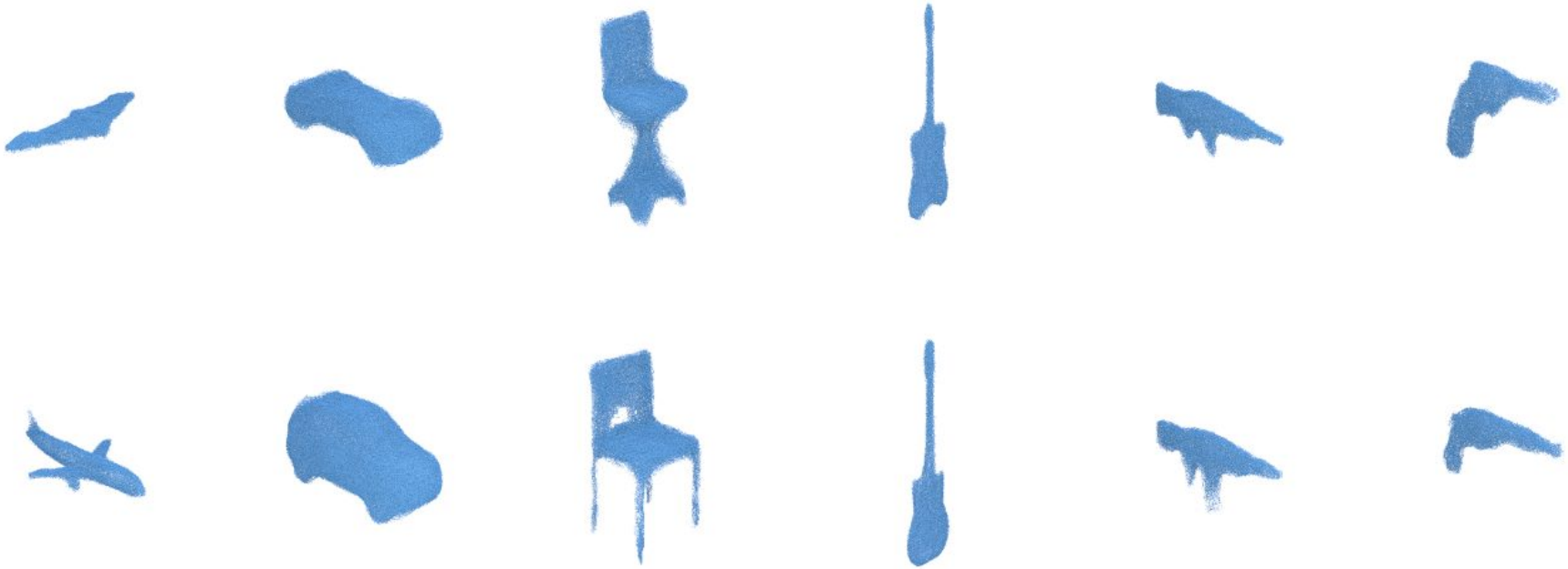
Results - IoU

IoU numbers for ShapeNet Core

	AllVP		Dropout		SingleVP	
	Depth	Sil.	Depth	Sil.	Depth	Sil.
Uncond.	81.5	80.8	77.0	76.3	71.1	68.0
Cond.	81.4	81.2	77.0	76.1	71.7	68.8

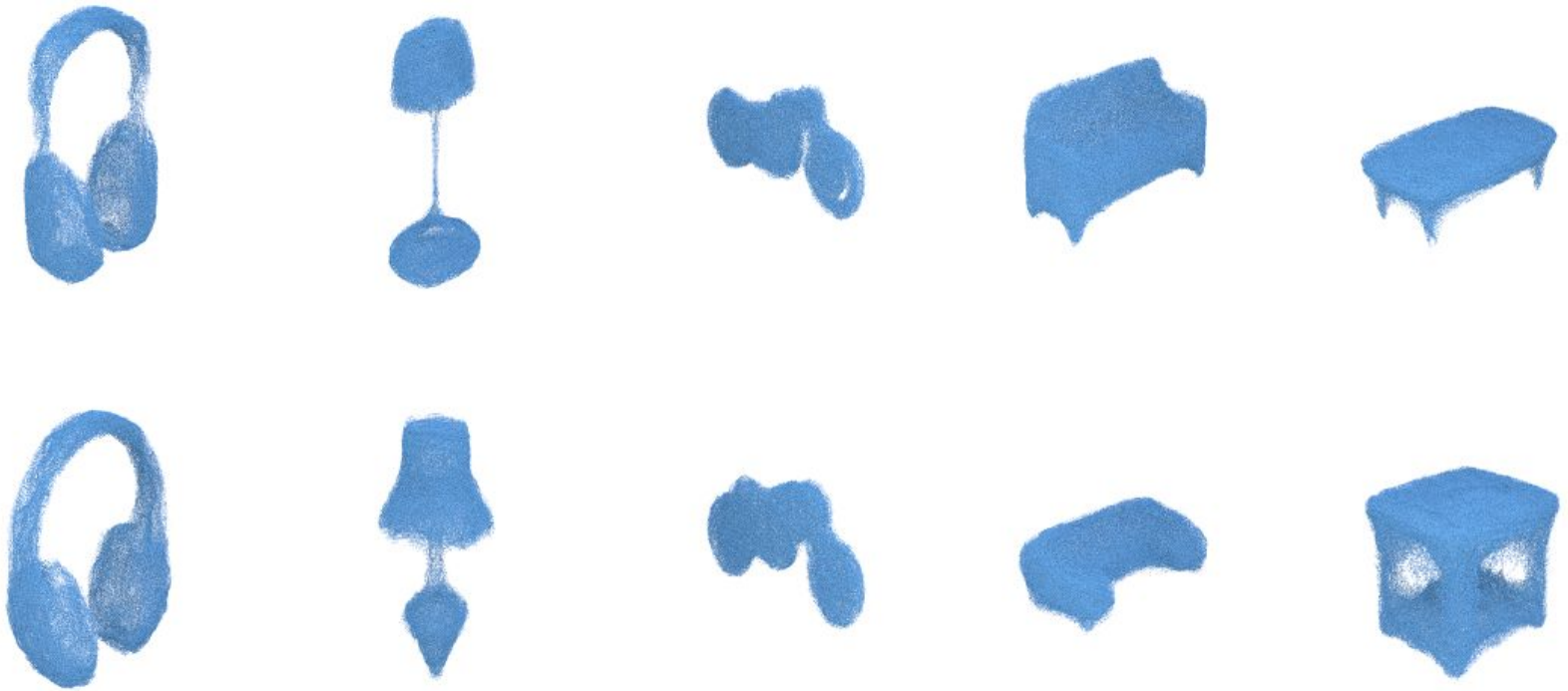
Results - Conditional Sampling

More Conditional Samples



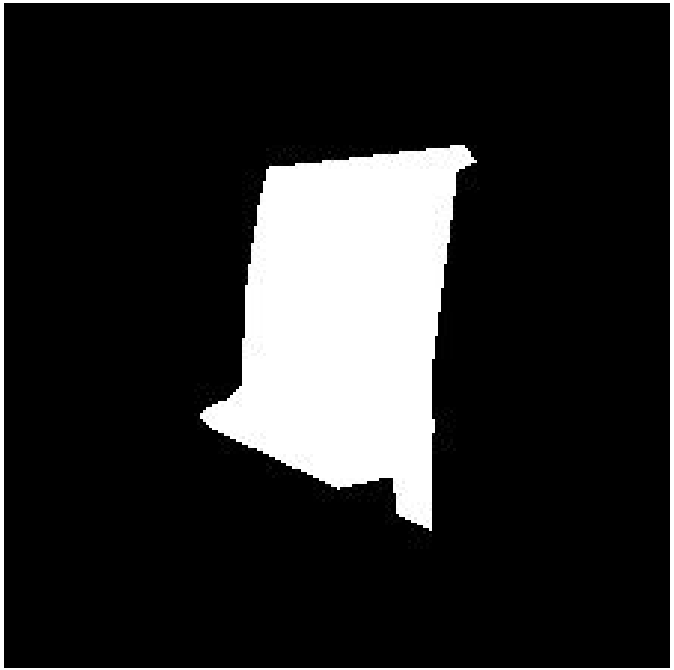
Results Conditional Sampling

More Conditional Samples

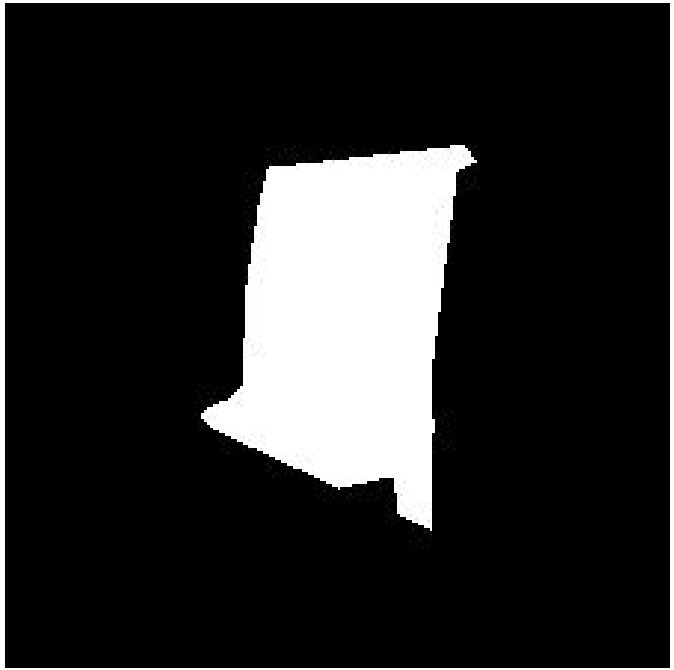


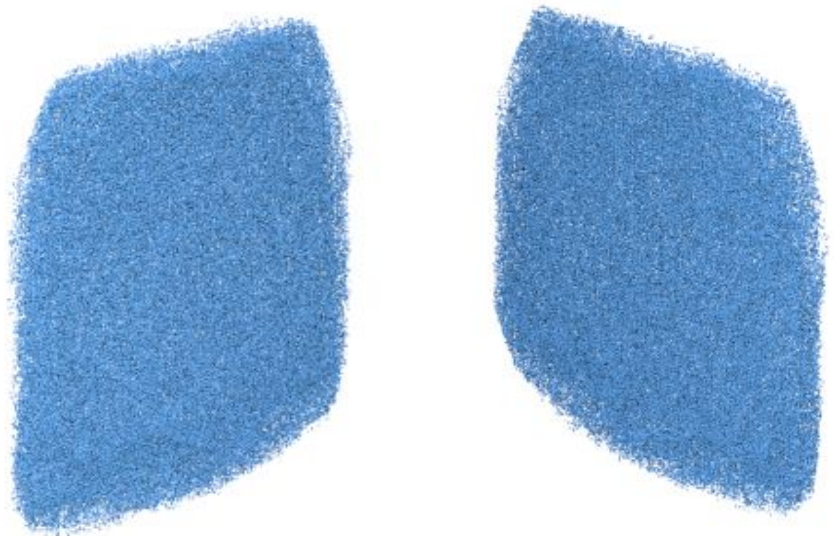
Results - Analysis

What about this?

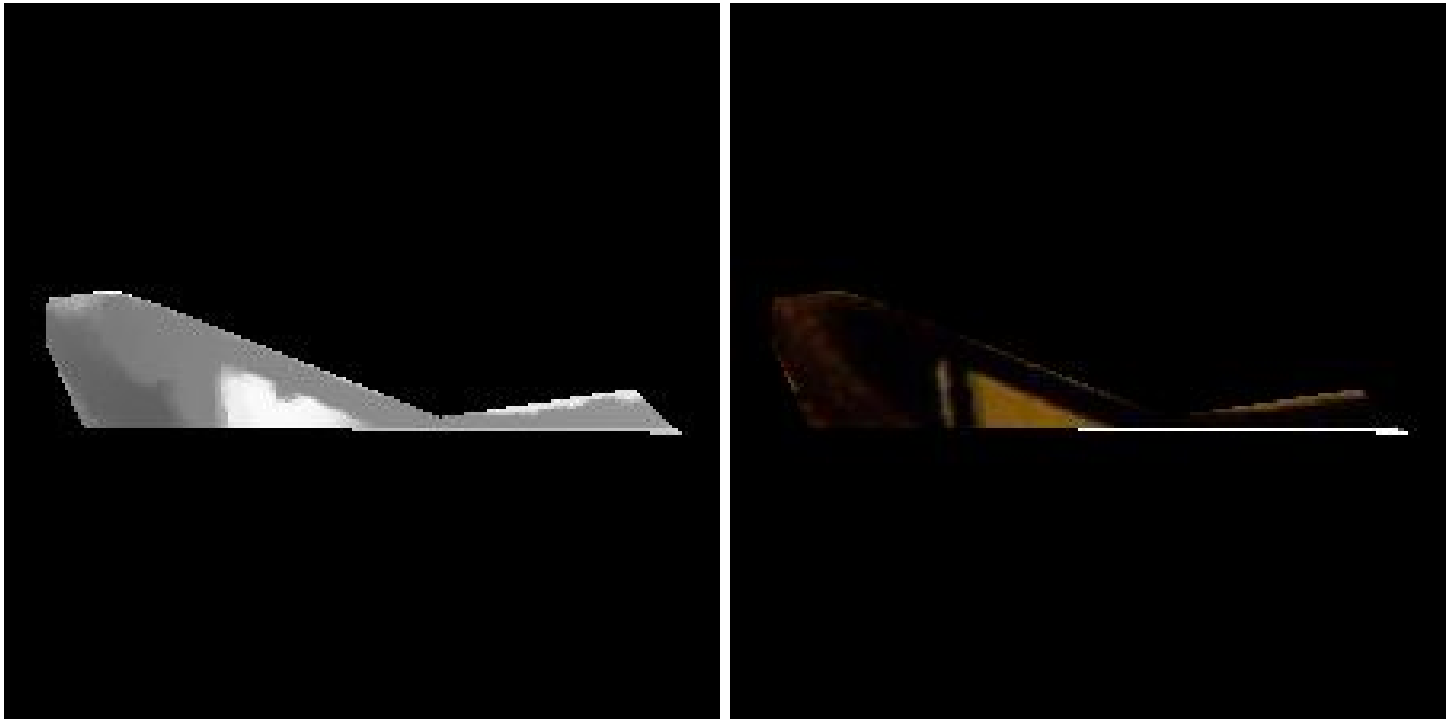


Results - Analysis

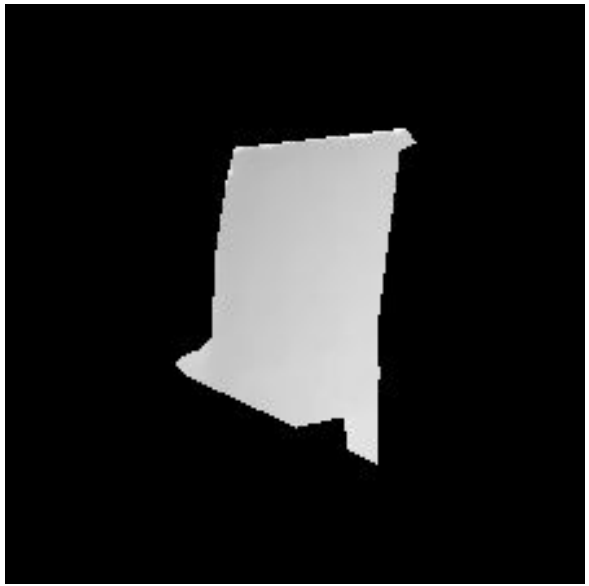


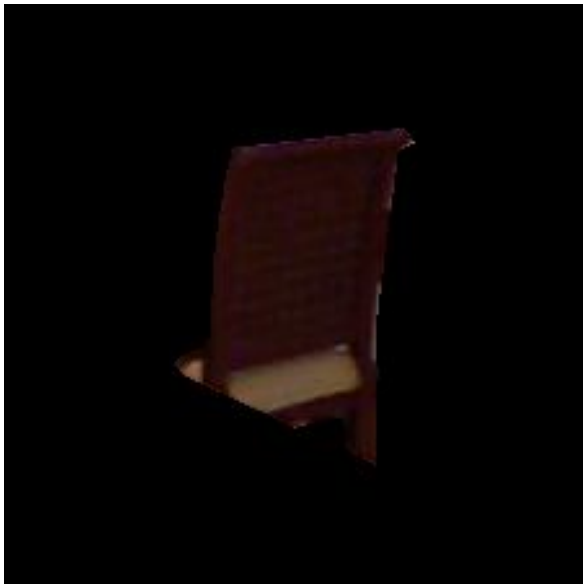


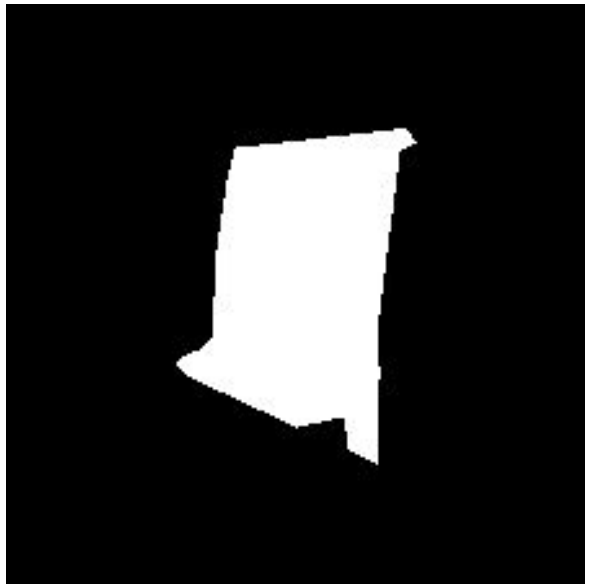
Results - Analysis



Results - Analysis







Building 3D Models of the World

Object manipulation