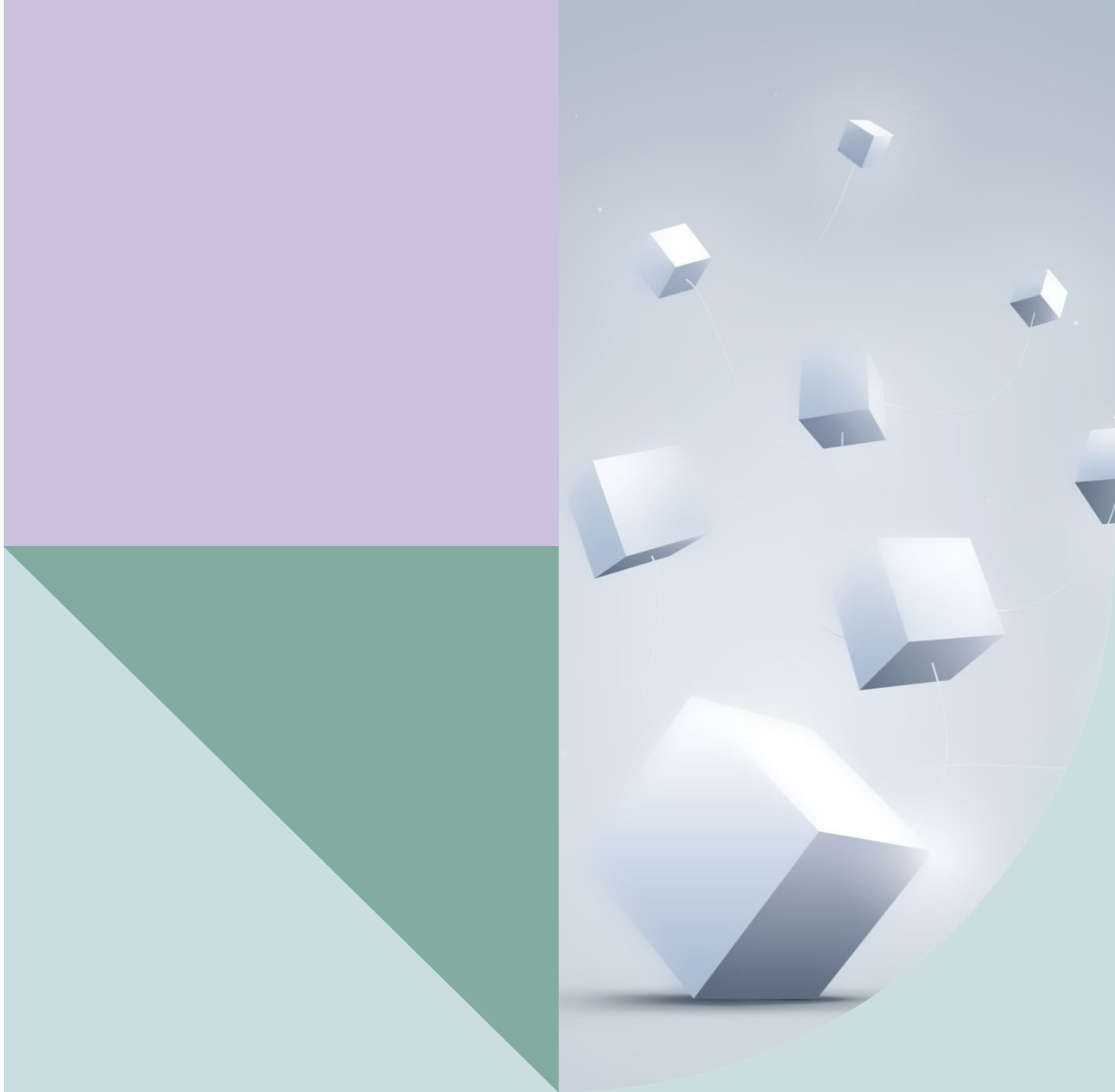
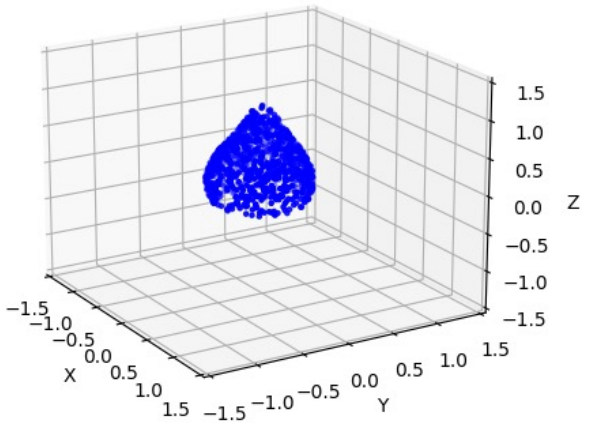
Shape-Shifters: Editing 3D Data with Language

Presenters: Abhinav Narayan Harish,
Laura Roettges,
Kevin Macauley, Keshav Sharan



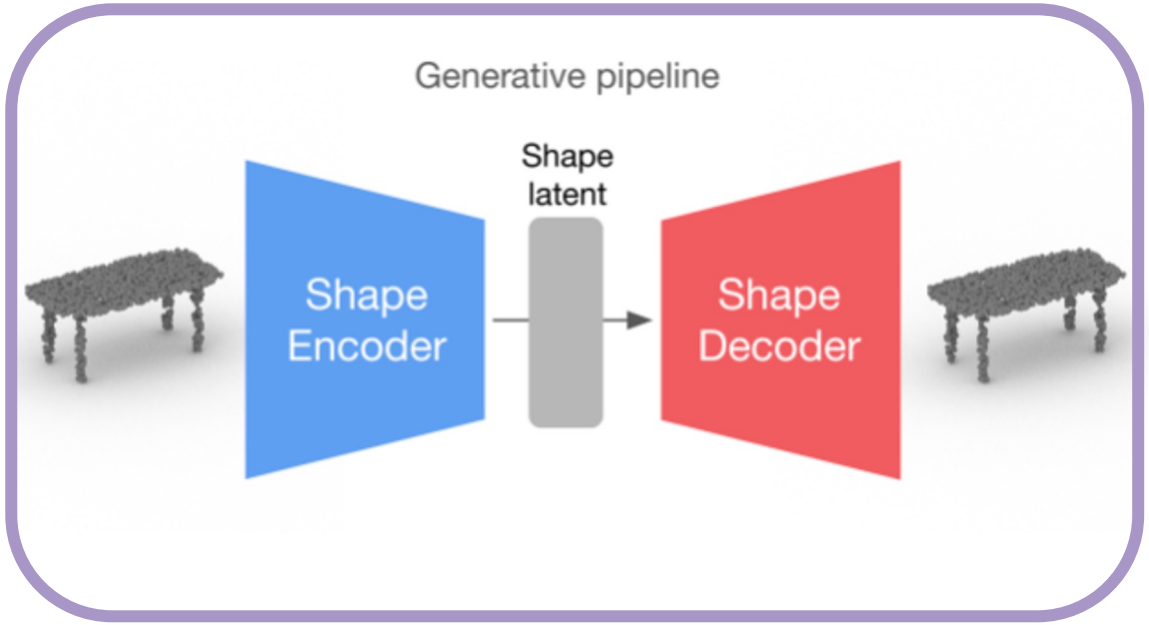
High level idea:
Use existing frameworks to manipulate 3D shapes with language inputs rather than elaborate editing software.

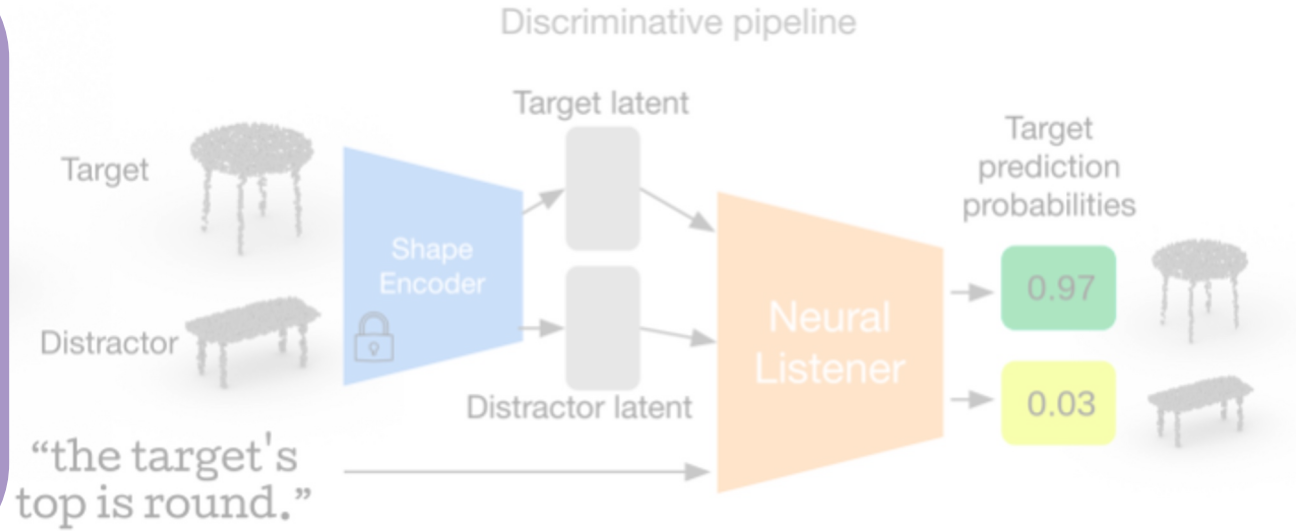


Sample stl & point cloud representation of a vase generated with the Creaform Handyscan 700 from the Makerspace

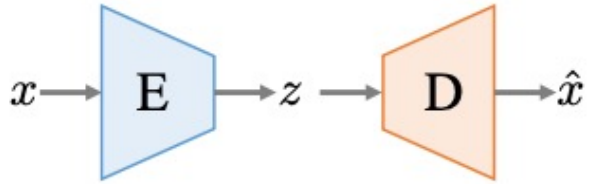
Can we 'fix' this by just sayings something like:
"The base is round and closed"

1st Stage: Generation & Discrimination



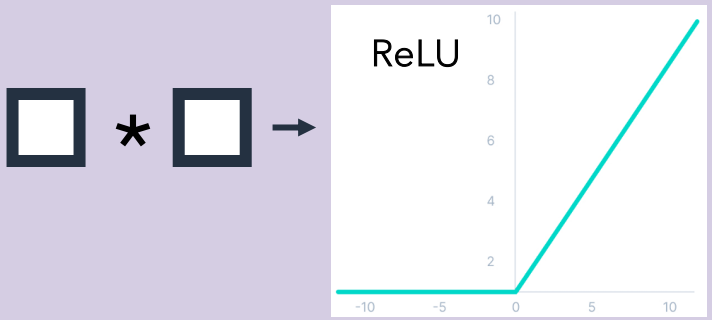


PC-AE



2048 x 3 matrix
(aka the point cloud)

- 1-D convolutional Layers
- Each layer followed by ReLU and batch-normalization layer



& batch normalization

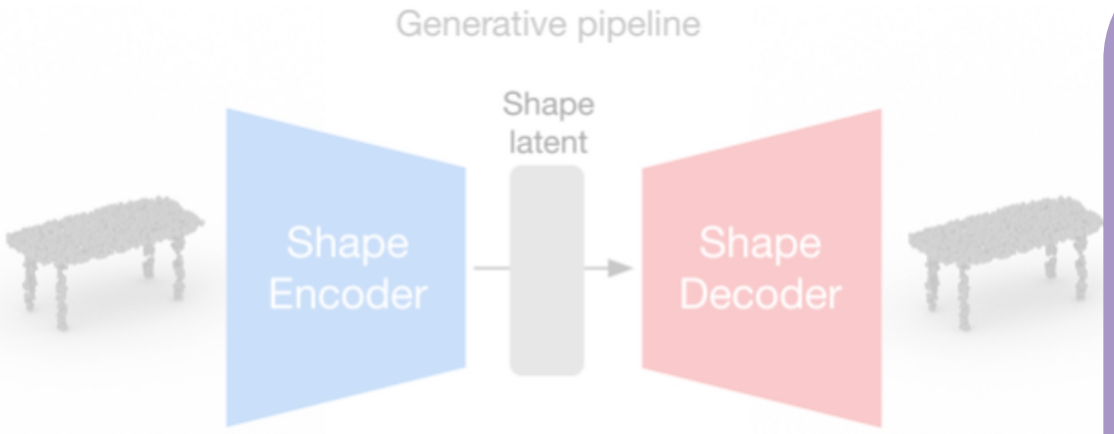
output is passed to a feature-wise maximum to produce 256 - dimensional vector

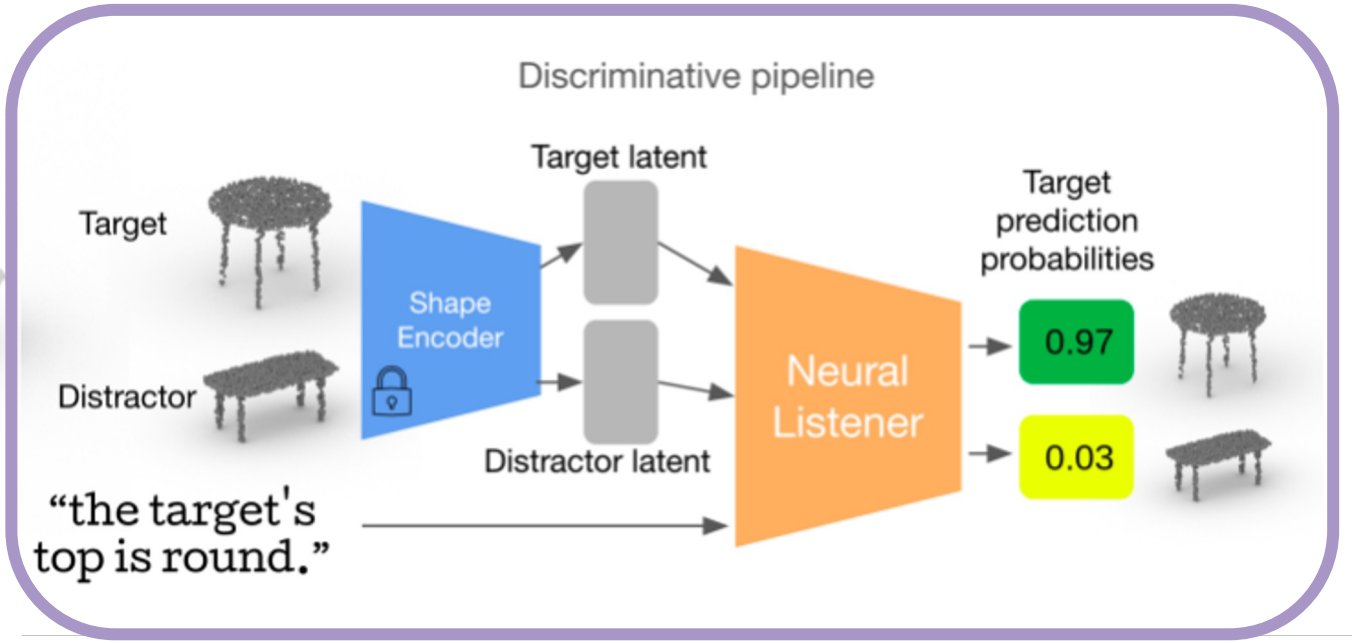
256-D vector
(aka our latent shape)

Decoder: 3 fully connected layers, 1st 2 w/ ReLU applied

2048 x 3 matrix

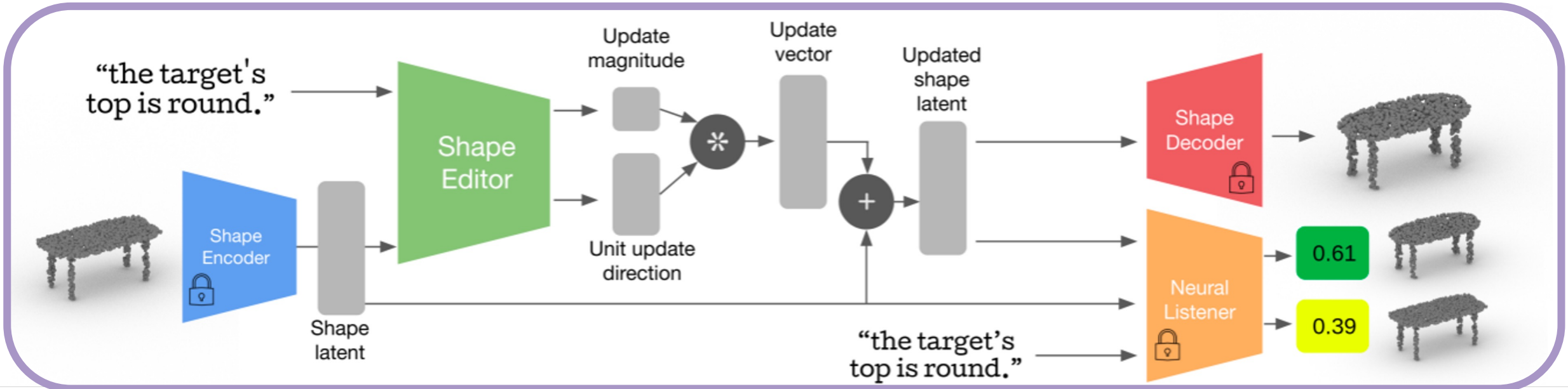
1st Stage: Generation & Discrimination





1st Stage: Generation & Discrimination

2nd Stage: Edit Prediction & Decoding



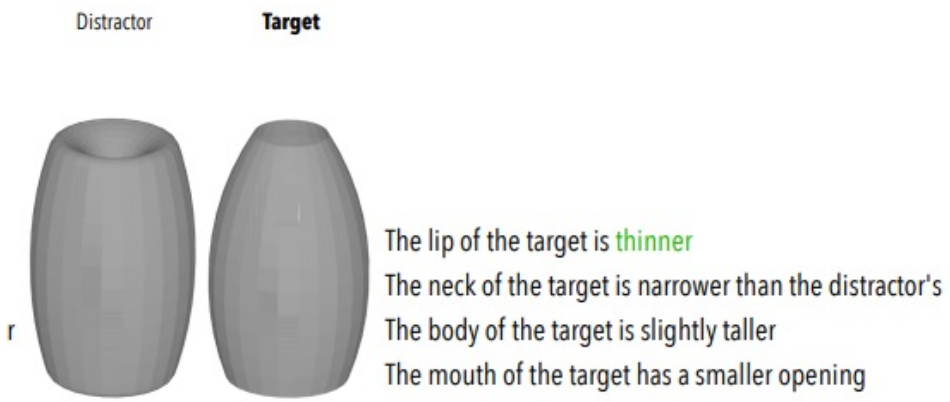
Experiment Design:

- Goal: Study the robustness of the change it 3d under varying input/output scenarios
- *Experiment-1:* Strength of ChangeIt3D to gaussian noise on point clouds
- *Experiment-2:* Robustness of ChangeIt3D to varying language instructions
- *Experiment-3:* How good is ChangeIt3D at Part removal?

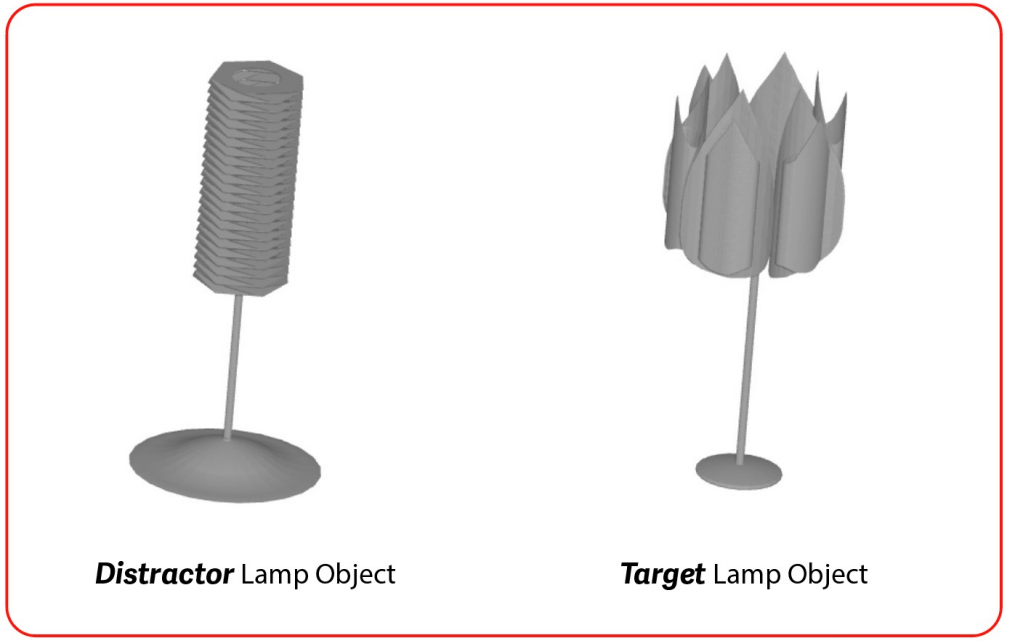
Dataset

- Shape-Talk Dataset: 536k utterances and 36k shapes

- Each utterance captures a contrast between two shapes: A *distractor* and a *target*



***Communication Context** (Easy)*



Utterances:

Each word represents a ***token***

Most apparent difference

Saliency 0: ***The target has a distinctive top shaped like a pointy crown***

Saliency 1: ***The target has a smaller base***

Lesser distinctive differences

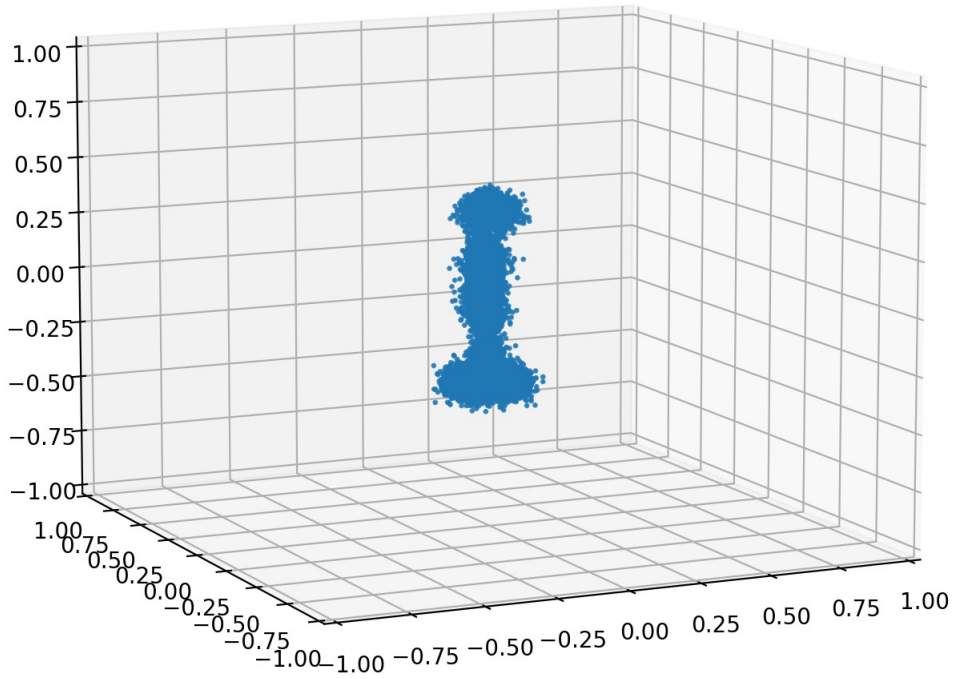
Saliency 2: ***The target has a longer pole.***

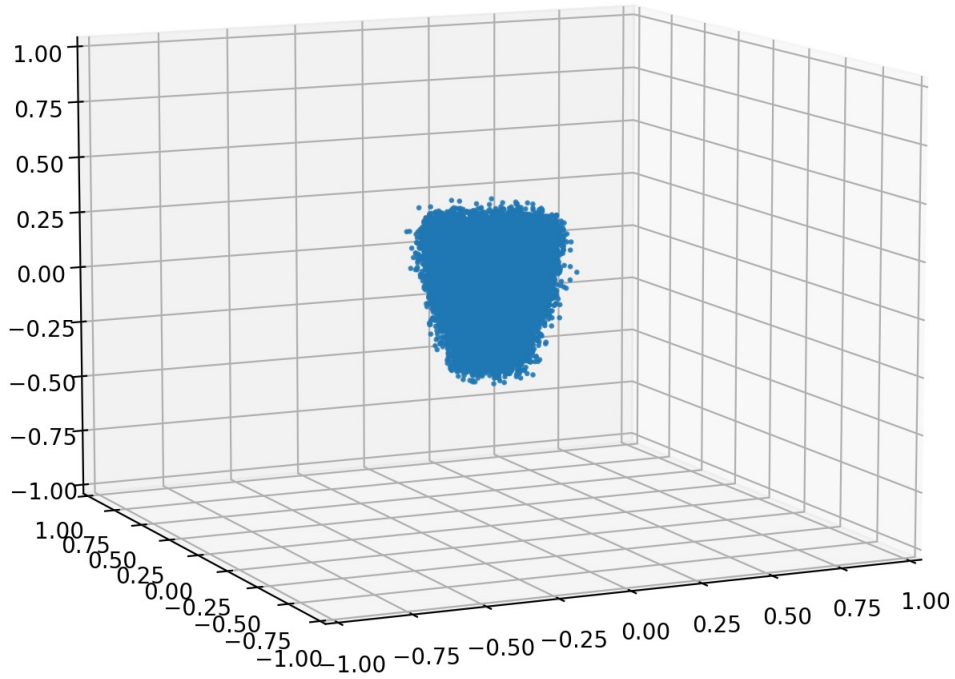
...

Experiment 1 – Evaluating performance on noisy point clouds

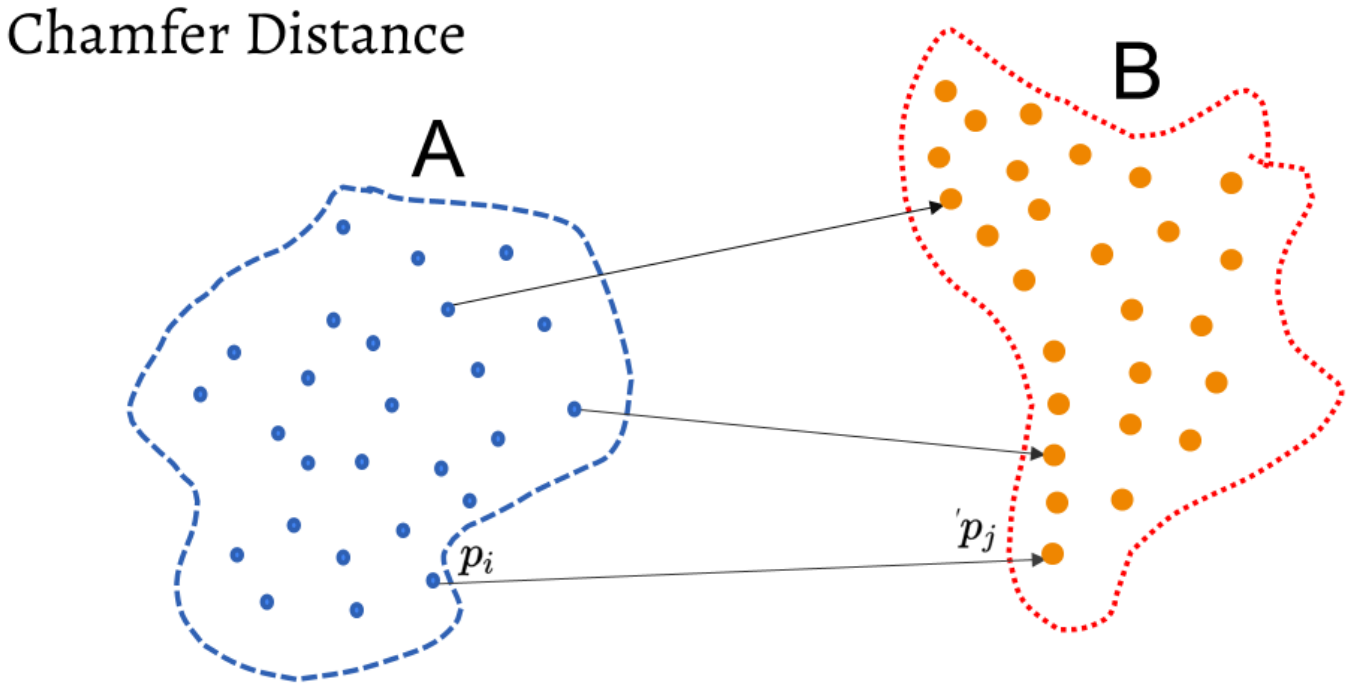
Dataset prep

- 3 classes (lamp, bottle, mug)
- 100 noisy point clouds (per class)
- sigma values : 10 values from (0.003, 0.03)





Evaluation Metric



$$d_{AB} = \sum_{p_i \in A} \min \|p_i - p_j\| \, \forall p_j \in B$$

$$d_{BA} = \sum_{p_j \in B} \min \|p_i - p_j\| \, \forall p_i \in A$$

Evaluation

Chamfer Distance - is used as a measure of dissimilarity between two sets of points, with a lower distance indicating greater similarity.

- Averaged over 100 point clouds/class

Noise Sigma	Lamp	Bottle	Mug
0.003	0.00552	0.00150	0.00099
0.005	0.00539	0.00144	0.00093
0.008	0.00514	0.00131	0.00084
0.01	0.00494	0.00123	0.00071
0.02	0.00383	0.00089	0.00061
0.025	0.00312	0.00082	0.00058
0.03	0.00276	0.00078	0.00056

Experiment 2 – evaluation on language differentiation

- tall vs lanky
- short vs squat
- circular vs curved

…

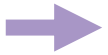
Dataset prep

- Across 3 classes: lamp, chair, & vase modified language for 20 utterances each
- Changes to language instructions

  "shorter" => "less high"

  "thinner" => "slim"

The seat is less wide.
The legs of the chair are thicker and the wheels are bigger
The legs are smaller.
The legs are very thin in shape.
Its back rest and seat are made from thin, curved slats.
the seat is very long, but thin
The target's back is taller.
The legs are shorter.
The arms of the chair are shorter and wider
The legs of the chair are much thinner.
The seat is more narrow.
The spindles are thinner
Its seat is thinner and not arched
it has a large round pedestal
The back of the chair is shorter.
it has four tall, thin legs
The seat is thinner, almost like a sheet of metal
The back is slightly narrower.
The back posts are round.
it has thinner and taller legs



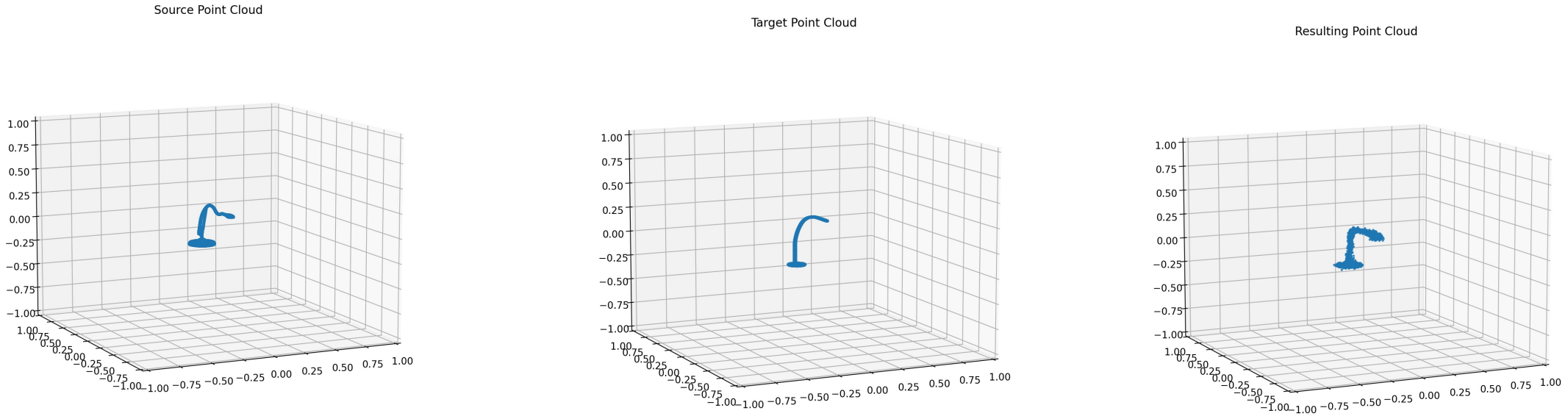
the seat is more condensed
the legs of the chair are more girthy and wheels more substantial
the legs are less sizeable
the legs are slight in shape
its back rest and seat are made from slim bent slats
the seat is very extensive but skinny
the back of the target towers more
the legs are less high
the arms of the chair are less extensive and more thick
the legs of the chair are much more svelte
the seat is less wide
the spindles are slim
the seat is thin and not curved
it has a big curved pedestal
the back of the chair is more stubby
it has four high skinny legs
the seat is skinny almost like a sheet of metal
the back is a little more skinny
the back posts are curved
it has more lanky legs

Chamfer Distance - is used as a measure of dissimilarity between two sets of points, with a lower distance indicating greater similarity.

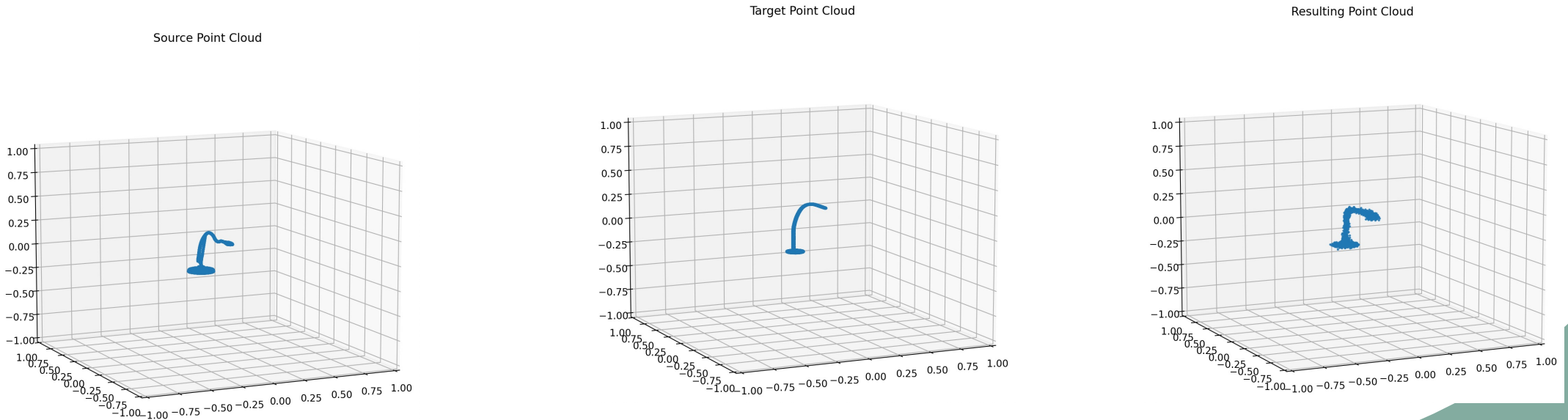
Evaluation

Experiment	Lamp	Chair	Vase
Baseline	0.00417	0.00257	0.00266
Modified Language	0.00459	0.00263	0.00288

Original: The long support arm is a smooth curve

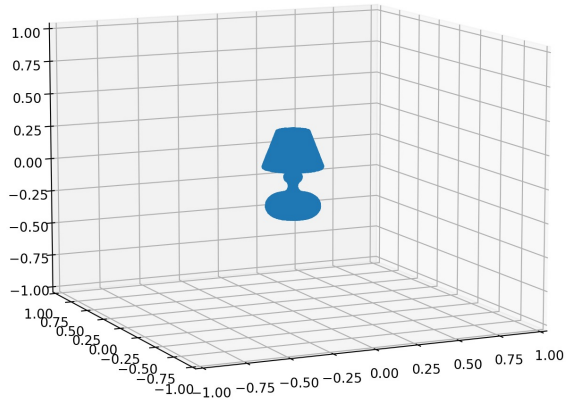


Modified: The long support arm is sleek

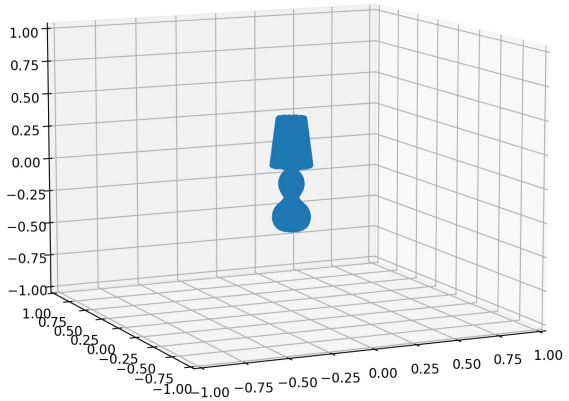


Original: It narrower body

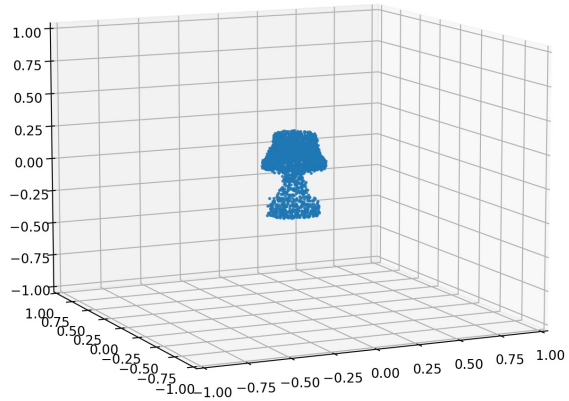
Source Point Cloud



Target Point Cloud

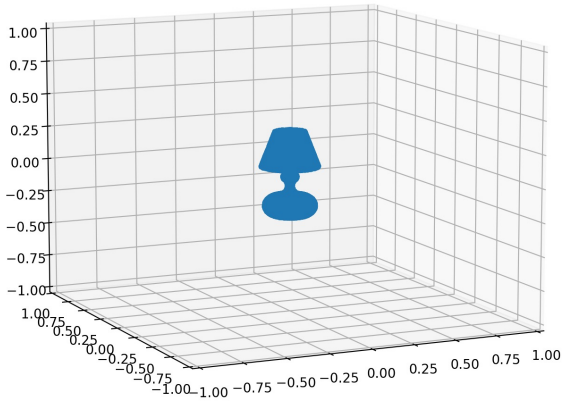


Resulting Point Cloud

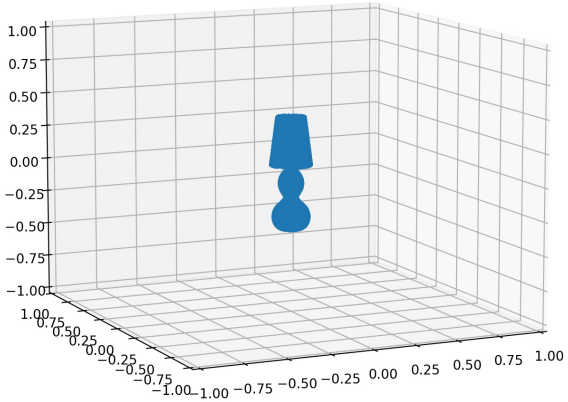


Modified: it has a thin body

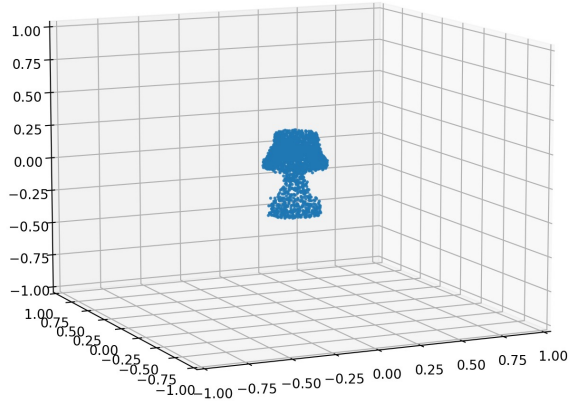
Source Point Cloud



Target Point Cloud

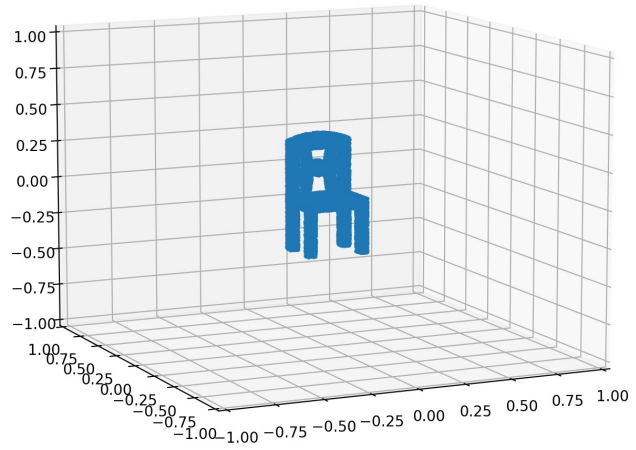


Resulting Point Cloud

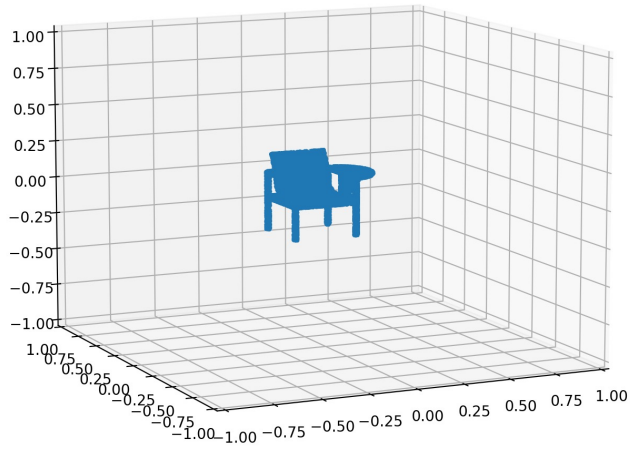


Original: Its back rest and seat are made from thin, curved slats

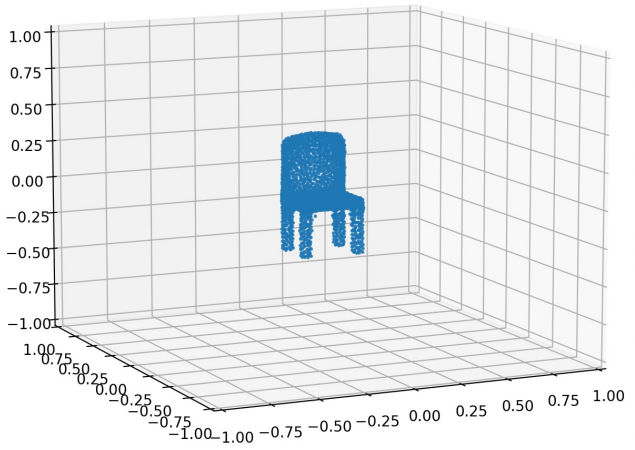
Source Point Cloud



Target Point Cloud

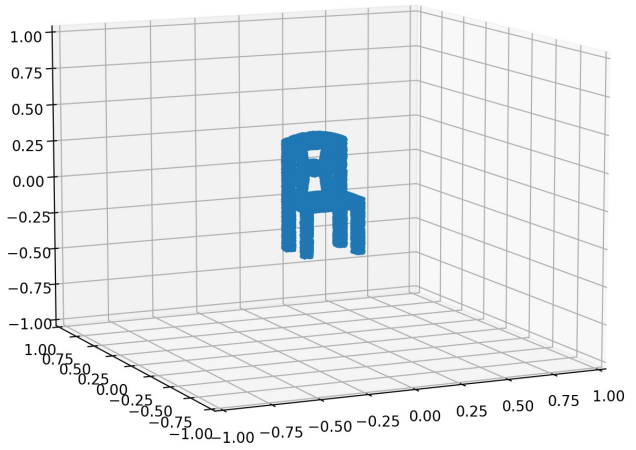


Resulting Point Cloud

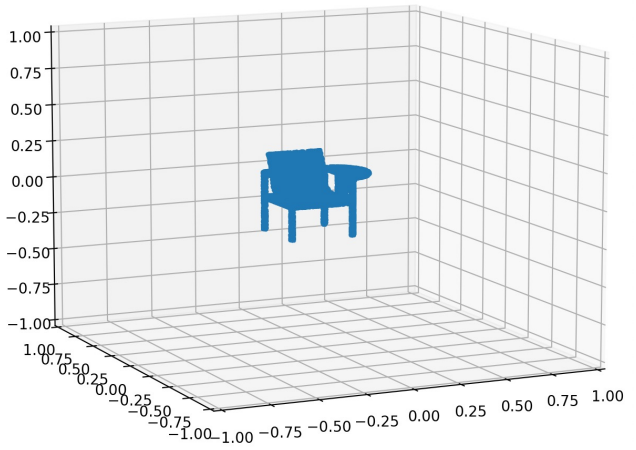


Original: its back rest and seat are made from slim bent slats

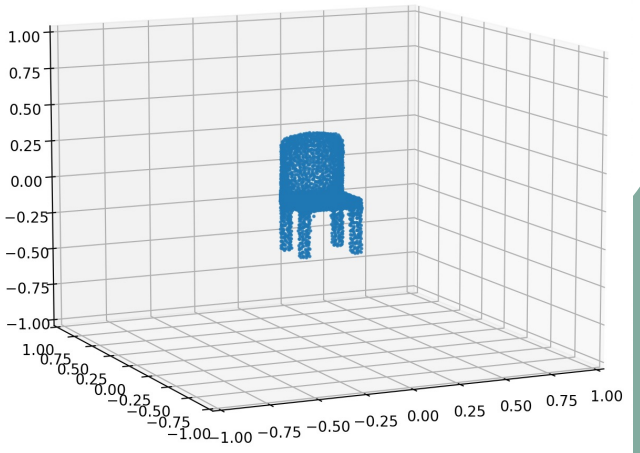
Source Point Cloud

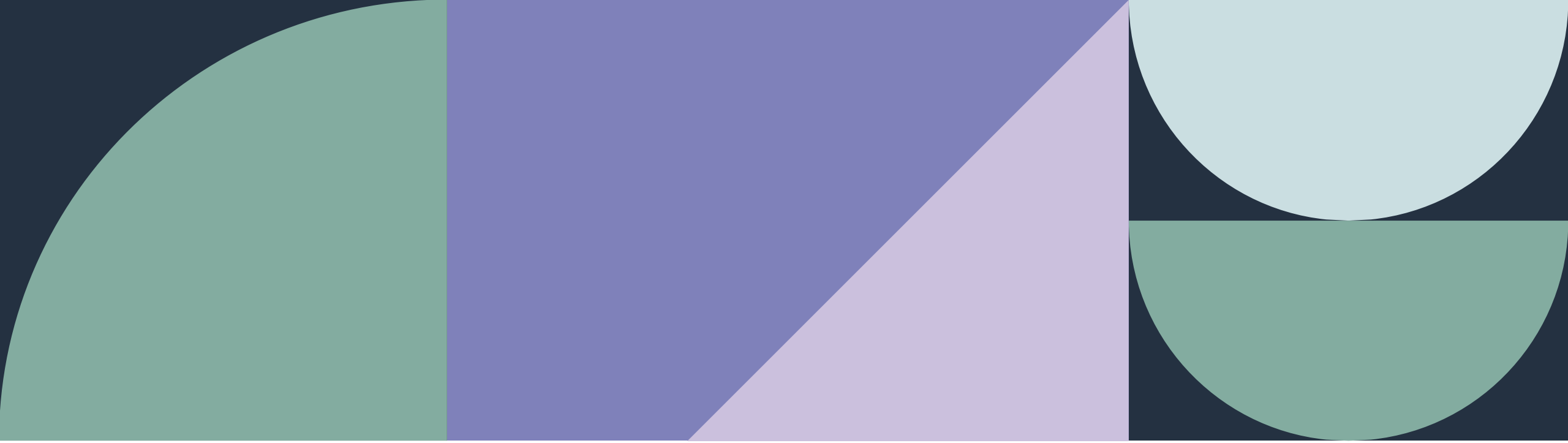


Target Point Cloud



Resulting Point Cloud





Experiment 3 – Shape Segmentation

Dataset Prep

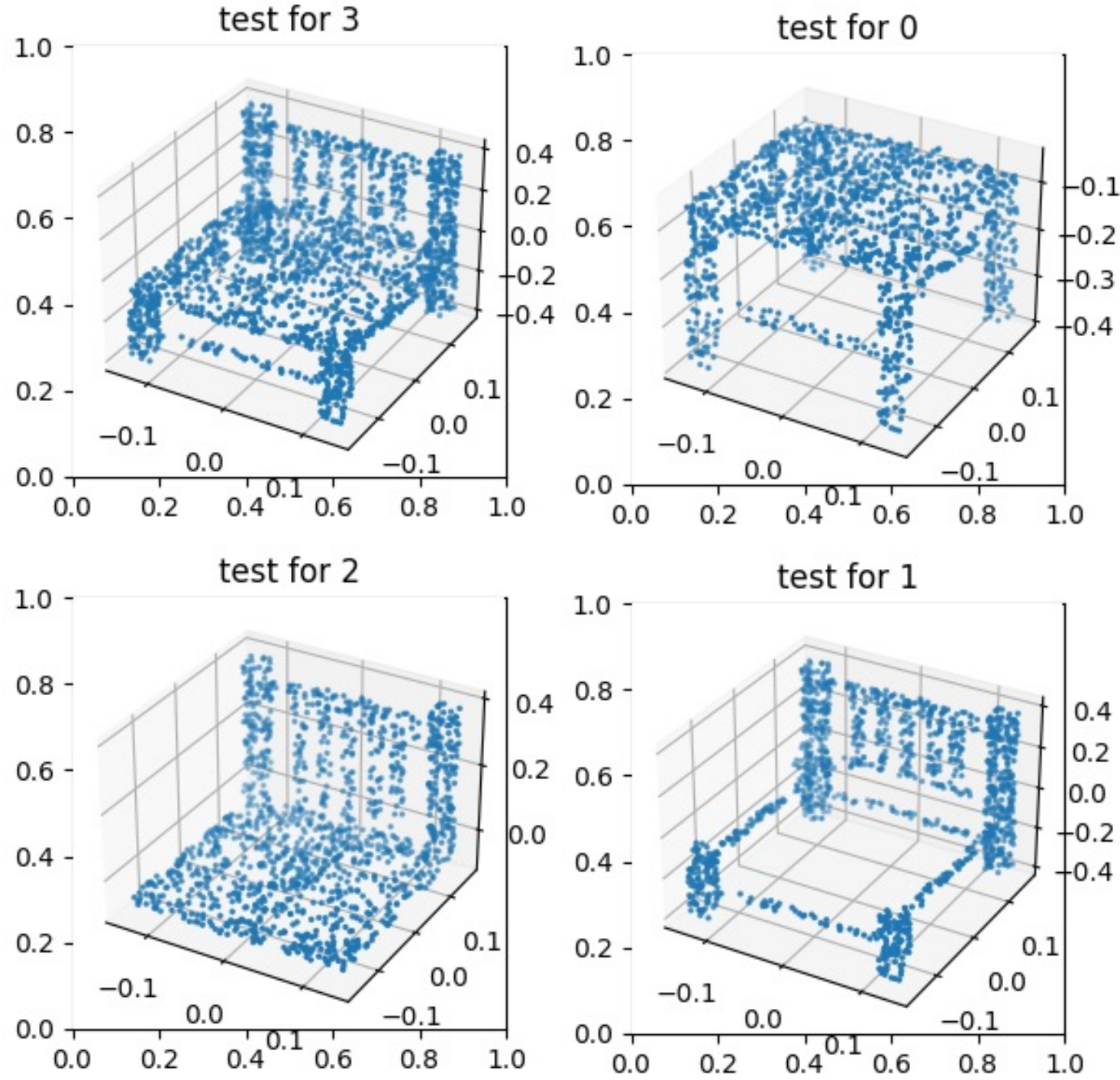
- Created a dataset of 140 shapes
- Used a pretrained network that classified points
- Manually separated all points for

  0 = back

  1 = seat

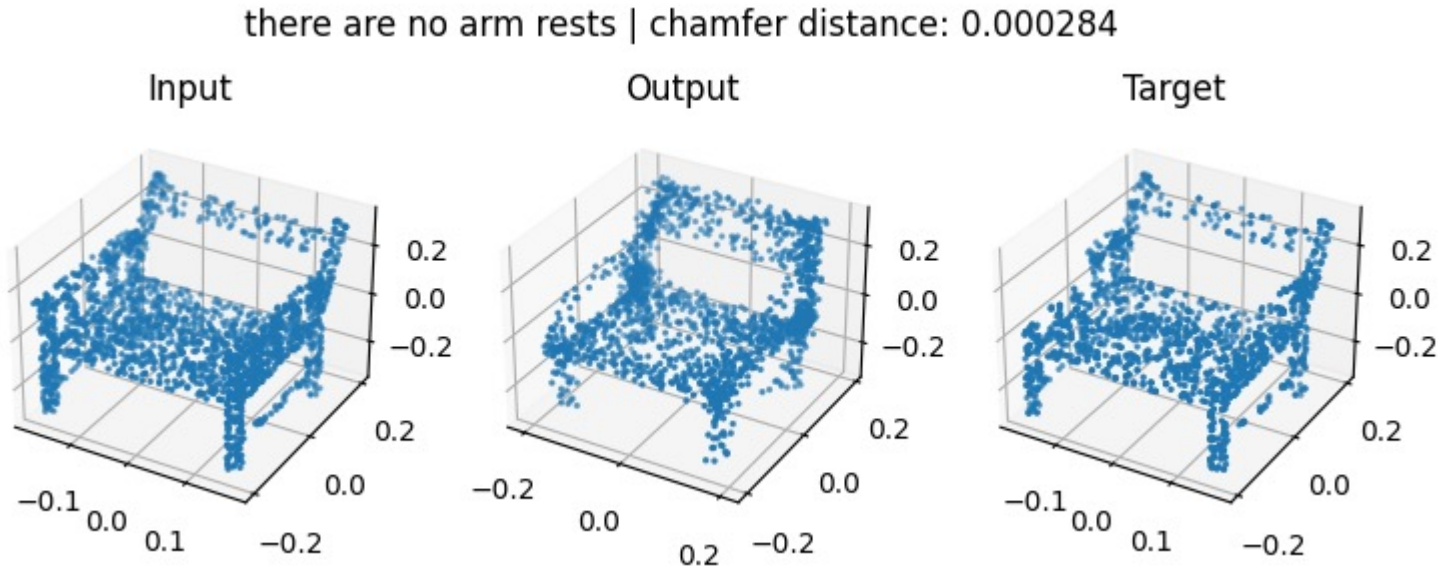
  2 = legs

  3 = arms

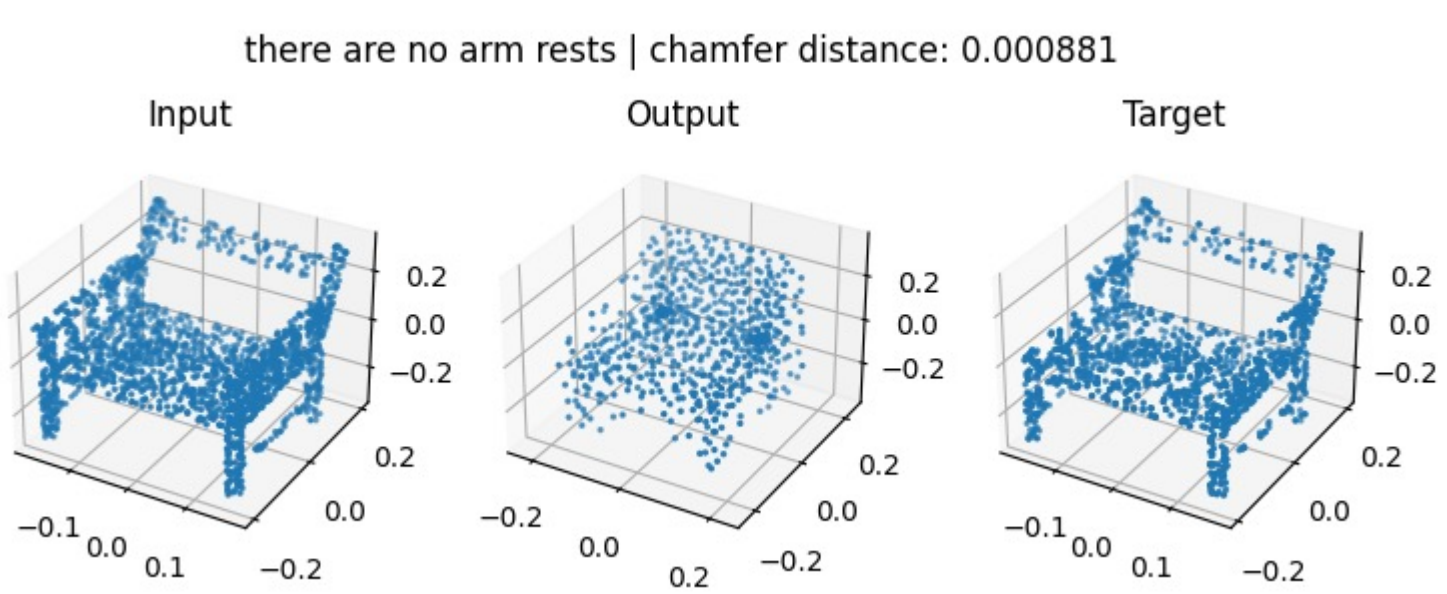


Results: Shape Segmentation with the Chair Class

ChangeIt3d

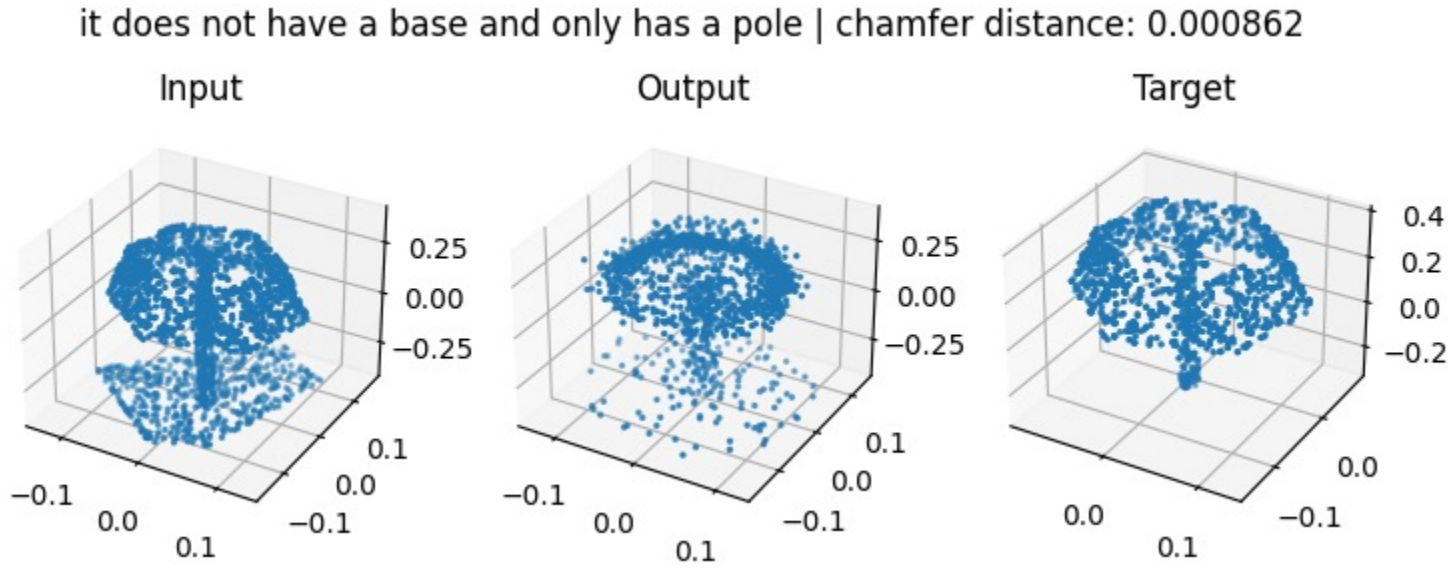


Monolithic

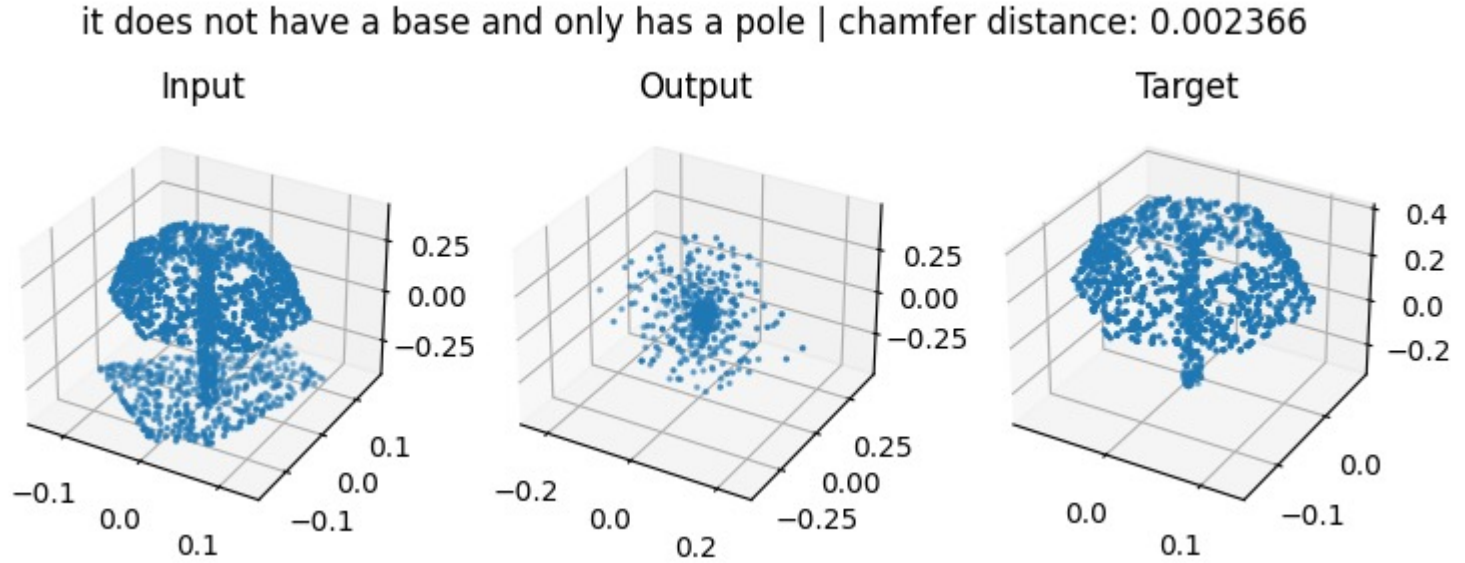


Results: Shape Segmentation with the Lamp Class

ChangeIt3d



Monolithic



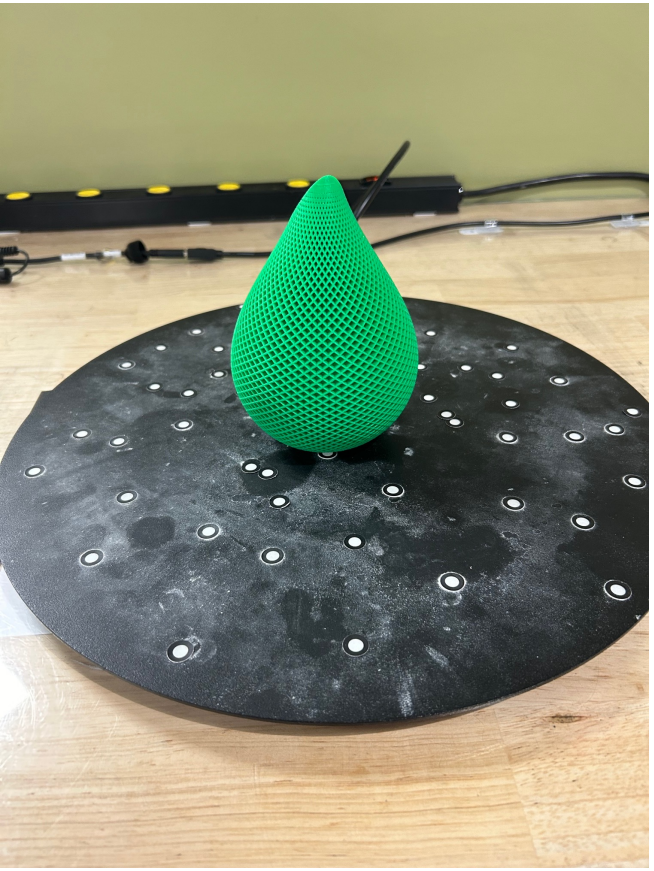
Results: Shape Segmentation Chamfer Distances

Class	Monolithic Chamfer Distance	Changeit3d Chamfer Distanc e
Chair	0.000831	0.000252
Lamp	0.001484	0.000305

Experiment 4 – Real World Data

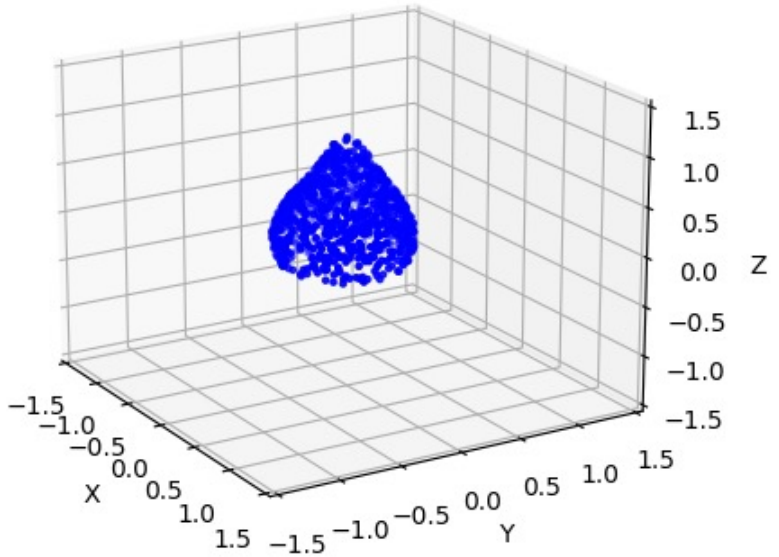
Real World Data Preparation

Physical Model

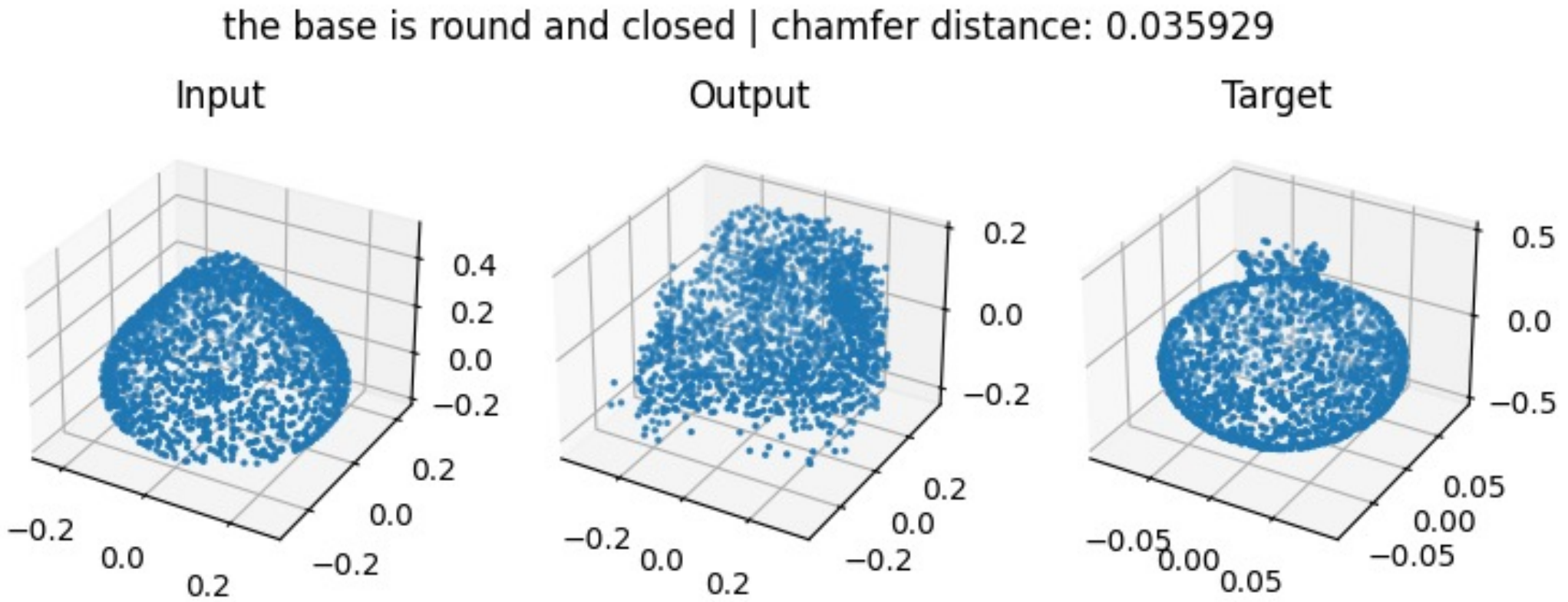


.stl File

Point Cloud Representation



Results: Shape Editing 3D Scan Data on Changeit3D



Conclusion and Next Steps

- Model is robust to changes within Shape-talk but struggle to generalize to point clouds outside the dataset

- Model good benefit from part recognition, i.e. Legs and arms look similar

References

[1] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, Zero-1-to-3: Zero-shot one image to 3d object, 2023. arXiv: 2303.11328 [cs.CV].

[2] A. Abdelreheem, I. Skorokhodov, M. Ovsjanikov, and P. Wonka, Satr: Zero-shot semantic segmentation of 3d shapes, 2023. arXiv: 2304.04909 [cs.CV].

[3] P. Achlioptas, I. Huang, M. Sung, S. Tulyakov, and L. Guibas, "ShapeTalk: A language dataset and framework for 3d shape edits and deformations," in Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[4] A. Radford, J. W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, 2021. arXiv: 2103.00020 [cs.CV].

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, Hierarchical text-conditional image generation with clip latents, 2022. arXiv: 2204.06125 [cs.CV].

[6] A. Sanghi, H. Chu, J. G. Lambourne, et al., Clip-forge: Towards zero-shot text-to-shape generation, 2022. arXiv: 2110.02624 [cs.CV].

[7] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, Zero-shot text-guided object generation with dream fields, 2022. arXiv: 2112.01455 [cs.CV].

[8] R. Fu, X. Zhan, Y. Chen, D. Ritchie, and S. Sridhar, Shapecrafter: A recursive text-conditioned 3d shape generation model, 2023. arXiv: 2207. 09446 [cs.CV].

[9] P. Achlioptas, J. Fan, R. Hawkins, N. Goodman, and L. Guibas, "ShapeGlot: Learning language for shape differentiation," in International Conference on Computer Vision (ICCV), 2019.

[10] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, Learning representations and generative models for 3d point clouds, 2018. arXiv: 1707.02392 [cs.CV].

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, Dec. 2015, Microsoft Research. "arXiv": 1512.03385 (cs.CV).

[12] Z. Chen and H. Zhang, Learning implicit fields for generative shape modeling, 2019. arXiv: 1812.02822 [cs.GR].

[13] A. X. Chang, T. A. Funkhouser, L. J. Guibas, et al., "Shapenet: An information-rich 3d model repository," CoRR, vol. abs/1512.03012, 2015. arXiv: 1512.03012. [Online]. Available: http://arxiv.org/abs/1512.03012.

[14] Z. Fang, X. Li, X. Li, S. Zhao, and M. Liu, Modelnet-o: A large-scale synthetic dataset for occlusion-aware point cloud classification, 2024. arXiv: 2401.08210 [cs.CV].

[15] K. Mo, S. Zhu, A. X. Chang, et al., "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," CoRR, vol. abs/1812.02713, 2018. arXiv: 1812.02713. [Online]. Available: http://arxiv.org/abs/1812.02713.

[16] P. Achlioptas, I. Huang, M. Sung, S. Tulyakov, and L. Guibas, Changeit3d: Language-assisted 3d shape edits and deformations, Supplemental Material, 2020.

[17] J. Koo, I. Huang, P. Achlioptas, L. J. Guibas, and M. Sung, "Partglot: Learning shape part segmentation from language reference games," in Conference on Computer Vision and Pattern Recognition (CVPR), 2022

Questions?