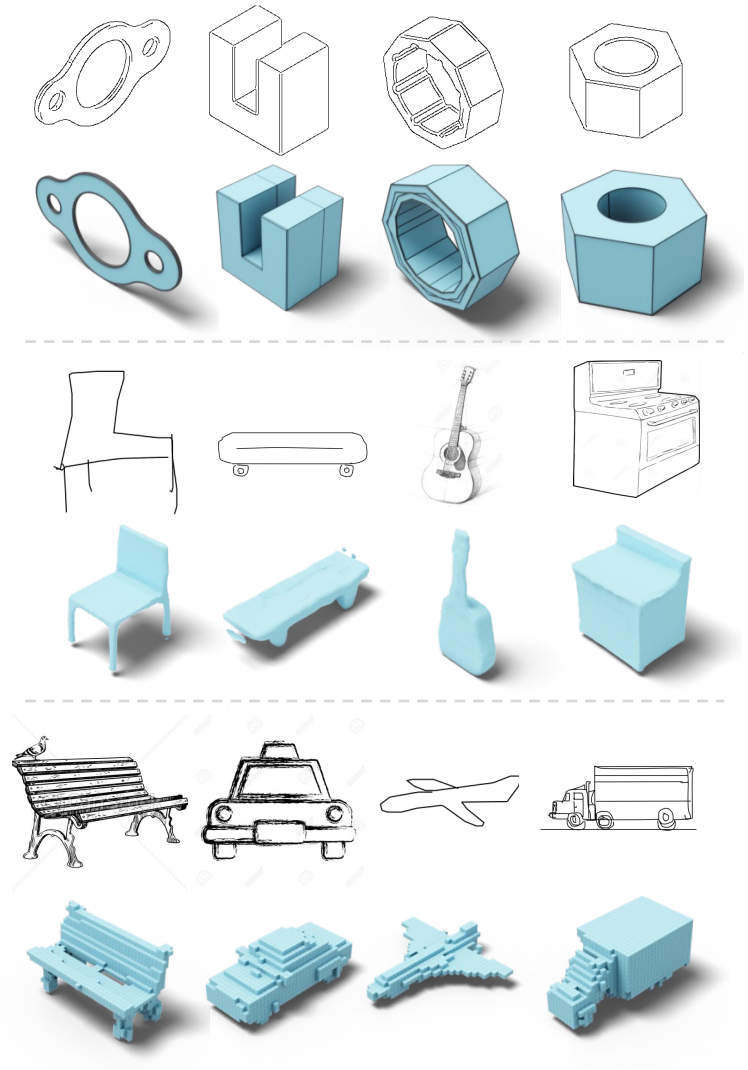
AUTODESK

Sketch-A-Shape

Zero-Shot Sketch-to-3D Shape Generation



Arianna Rampini
BIRS Workshop
July 12th 2023

AUTODESK

Aditya Sanghi

Pradeep Kumar Jayaraman

Joseph Lambourne

Hooman Shayani

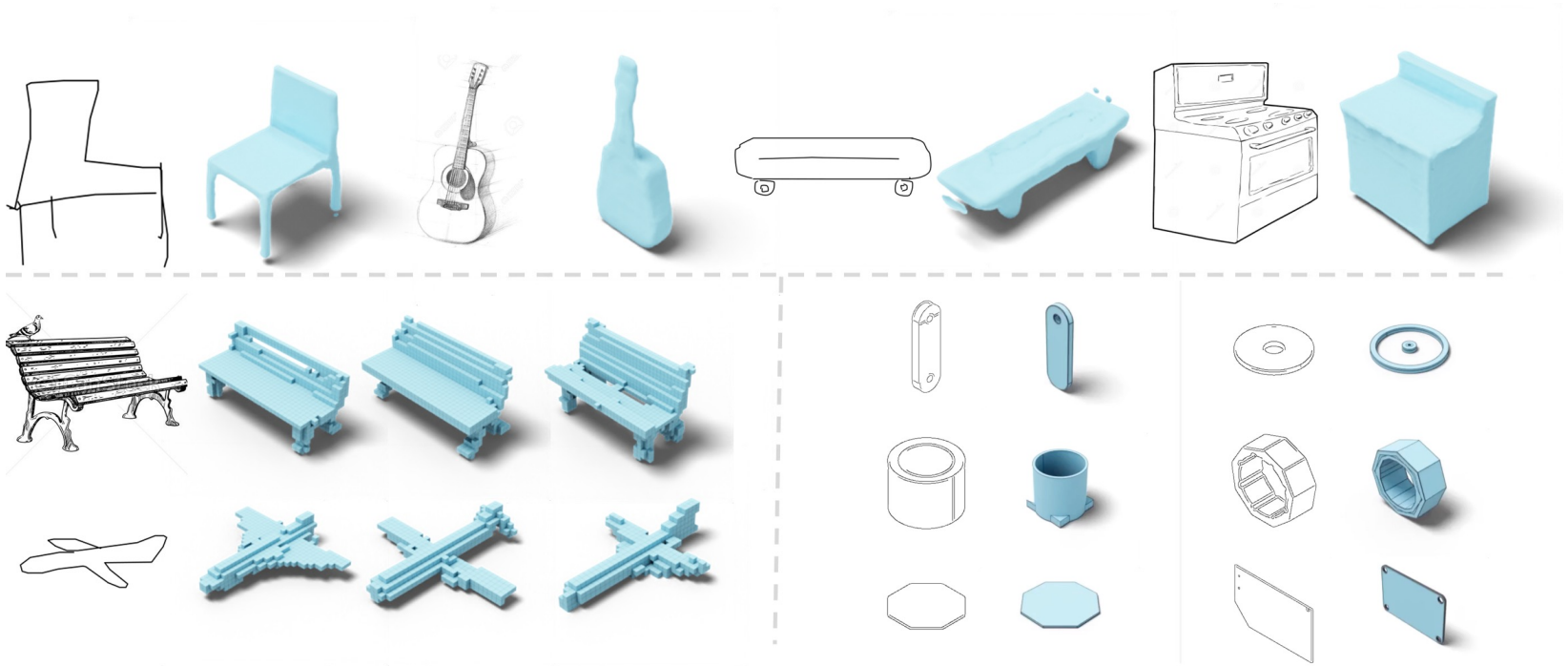
Saeid Asgari Taghanaki

Evan Atherton

How do humans sketch objects? – Mathias Eitz, James Hays and Marc Alexa, SIGGRAPH 2012

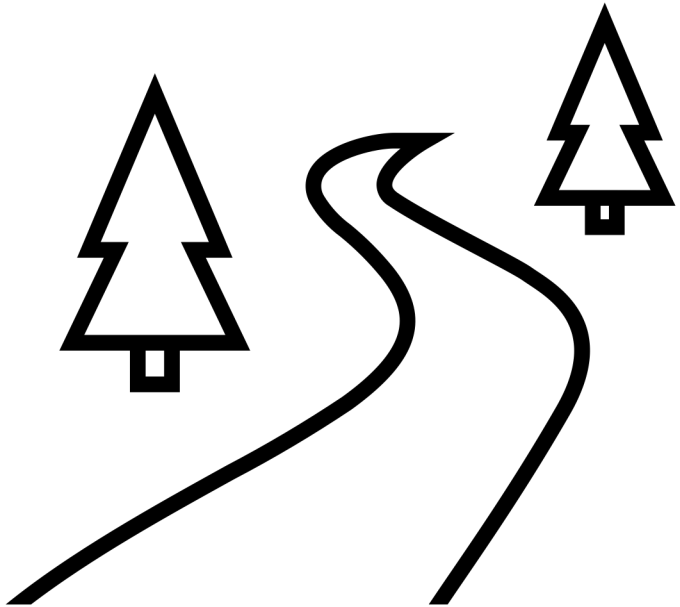

IT'S "DIGNITY!"

Sketch to 3D



Outline

- Previous methods
- Our idea
- Results
- Analysis
- Future work

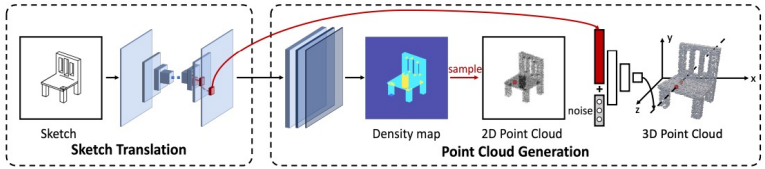


Previous approaches

SketchSampler: Sketch-based 3D Reconstruction via View-dependent Depth Sampling

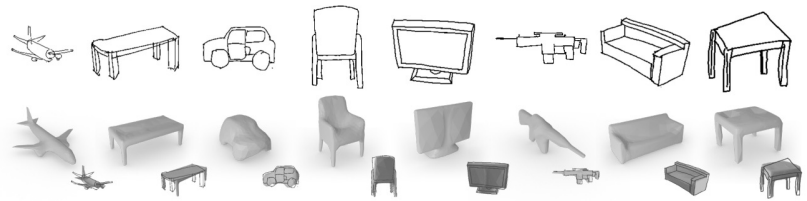
Chenjian Gao[1], Qian Yu[1⋆], Lu Sheng[1], Yi-Zhe Song[2], and Dong Xu[3]

[1] School of Software, Beihang University
{gaochenjian, qianyu, lsheng}@buaa.edu.cn
[2] SketchX, CVSSP, University of Surrey
y.song@surrey.ac.uk
[3] Department of Computer Science, The University of Hong Kong
dongxudongxu@gmail.com



Sketch2Model: View-Aware 3D Modeling from Single Free-Hand Sketches

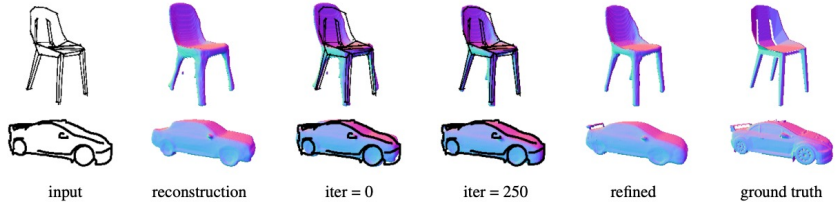
Song-Hai Zhang* Yuan-Chen Guo Qing-Wen Gu
BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing
shz@tsinghua.edu.cn, guoyc19@mails.tsinghua.edu.cn, gqw17@mails.tsinghua.edu.cn



Sketch2Mesh: Reconstructing and Editing 3D Shapes from Sketches

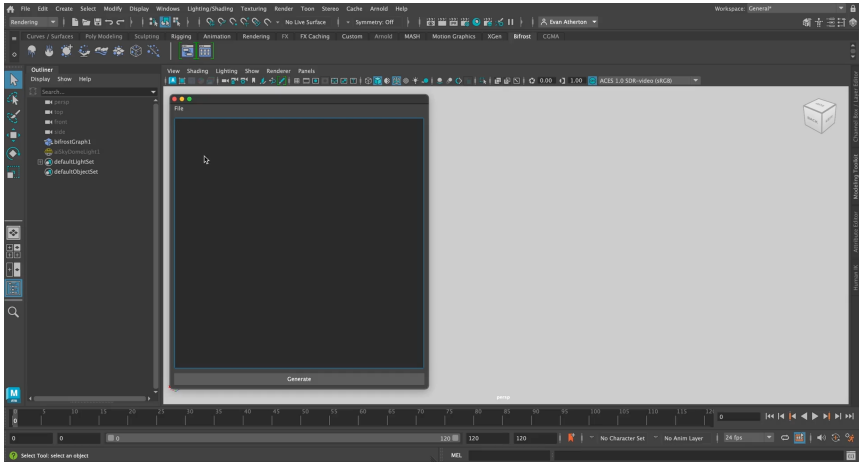
Benoit Guillard*, Edoardo Remelli*, Pierre Yvernay, Pascal Fua

CVLab, EPFL
name.surname@epfl.ch



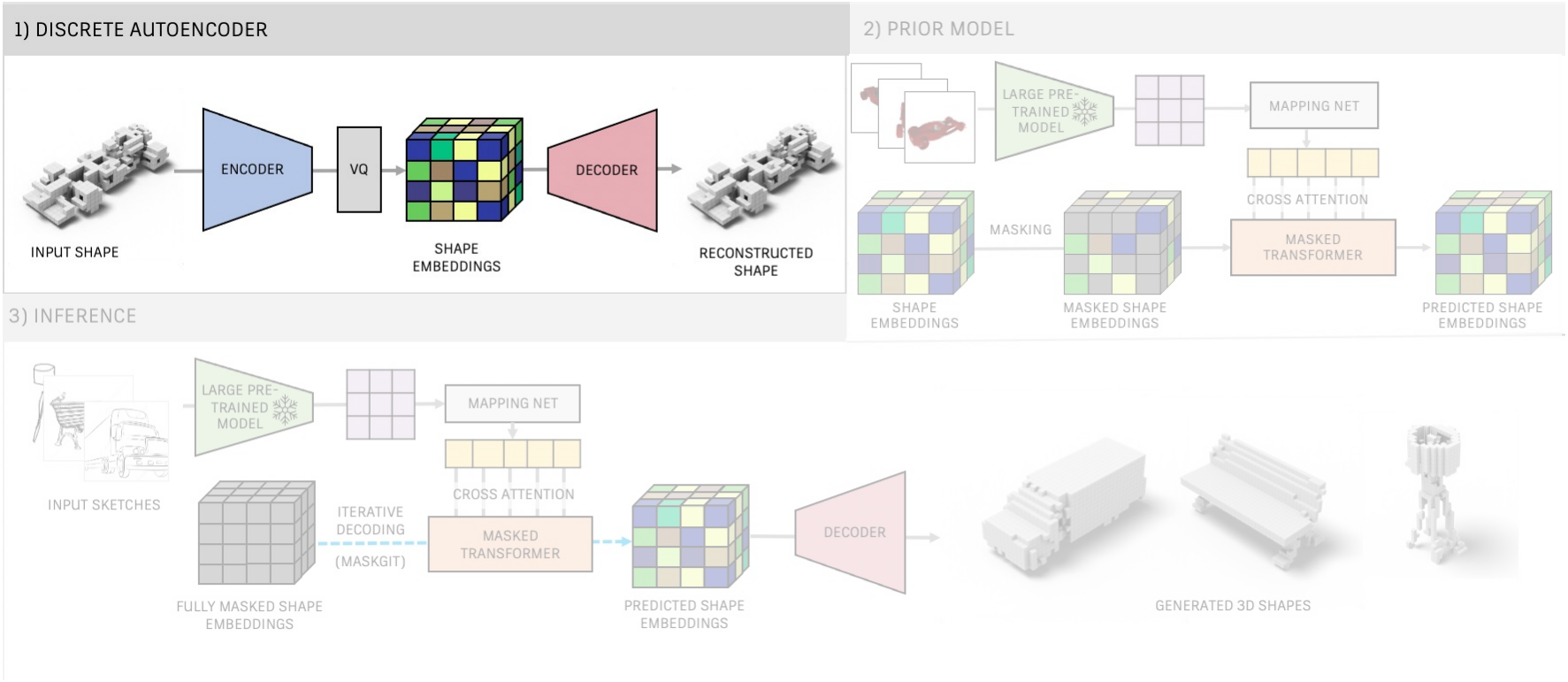
Sketch-A-Shape

- No paired data 3D-sketches
- Pre-trained large models
- Preserve stylistic details
- Several possible 3D representation

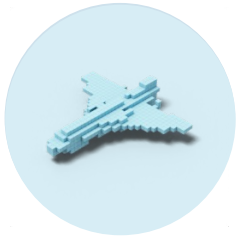


Example usage on Maya

Overview


1) DISCRETE AUTOENCODER
ENCODER
VQ
DECODER
INPUT SHAPE
SHAPE EMBEDDINGS
RECONSTRUCTED SHAPE
2) PRIOR MODEL
LARGE PRE-TRAINED MODEL
MAPPING NET
CROSS ATTENTION
MASKING
MASKED TRANSFORMER
SHAPE EMBEDDINGS
MASKED SHAPE EMBEDDINGS
PREDICTED SHAPE EMBEDDINGS
3) INFERENCE
INPUT SKETCHES
LARGE PRE-TRAINED MODEL
MAPPING NET
CROSS ATTENTION
ITERATIVE DECODING
(MASKGIT)
MASKED TRANSFORMER
DECODER
FULLY MASKED SHAPE EMBEDDINGS
PREDICTED SHAPE EMBEDDINGS
GENERATED 3D SHAPES

Discrete Autoencoder

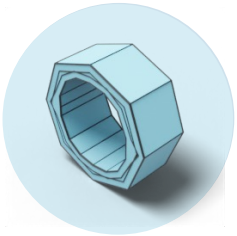


VOXELS

VQ-VAE

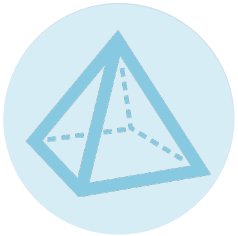
IMPLICIT

OCCUPANCY NET



CAD

SKEXGEN

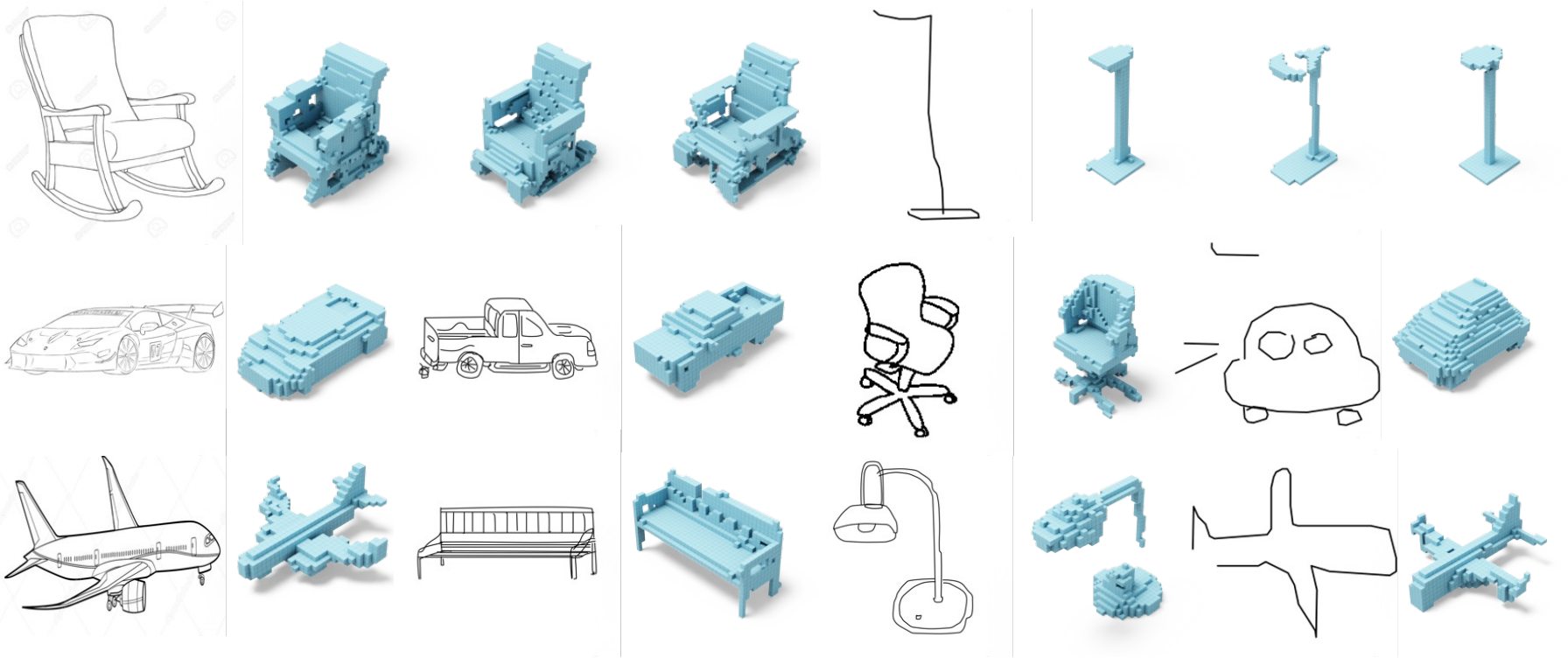


CAN BE ANYTHING

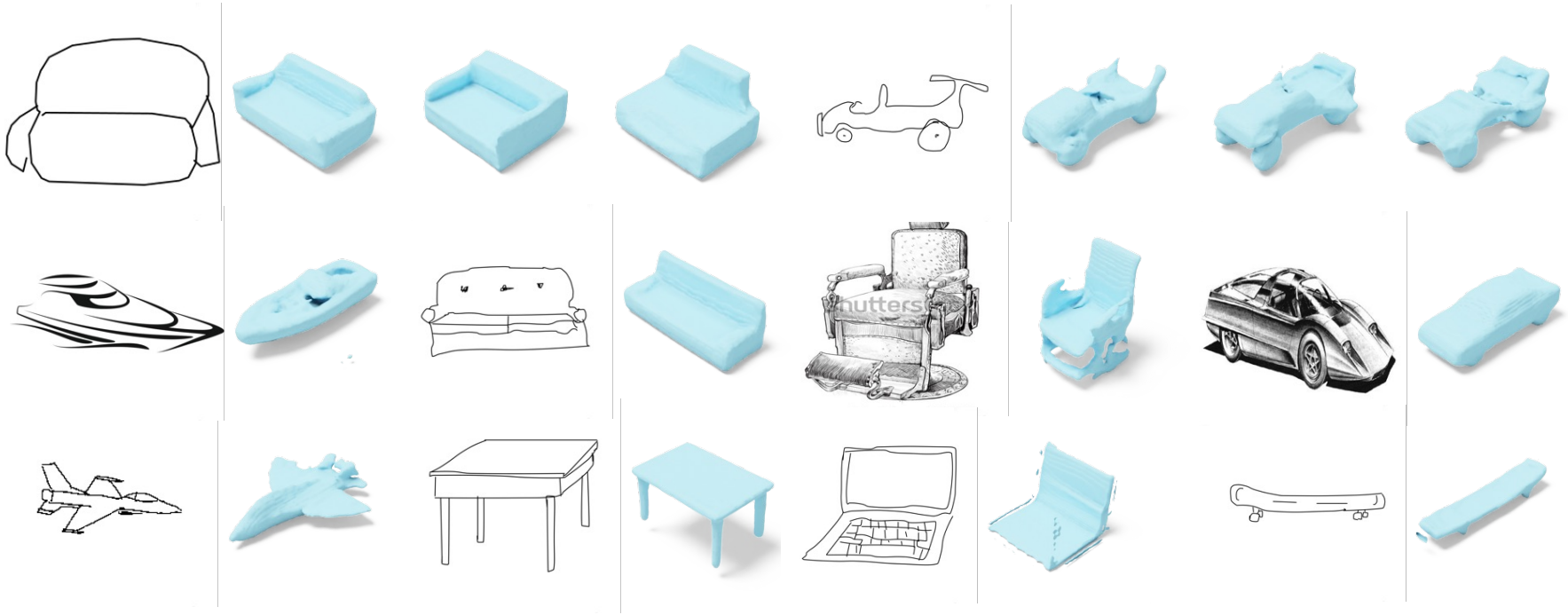
...

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. (2017)
Xu, Xiang, et al. SkexGen: Autoregressive generation of CAD construction sequences with disentangled codebooks. (2022)

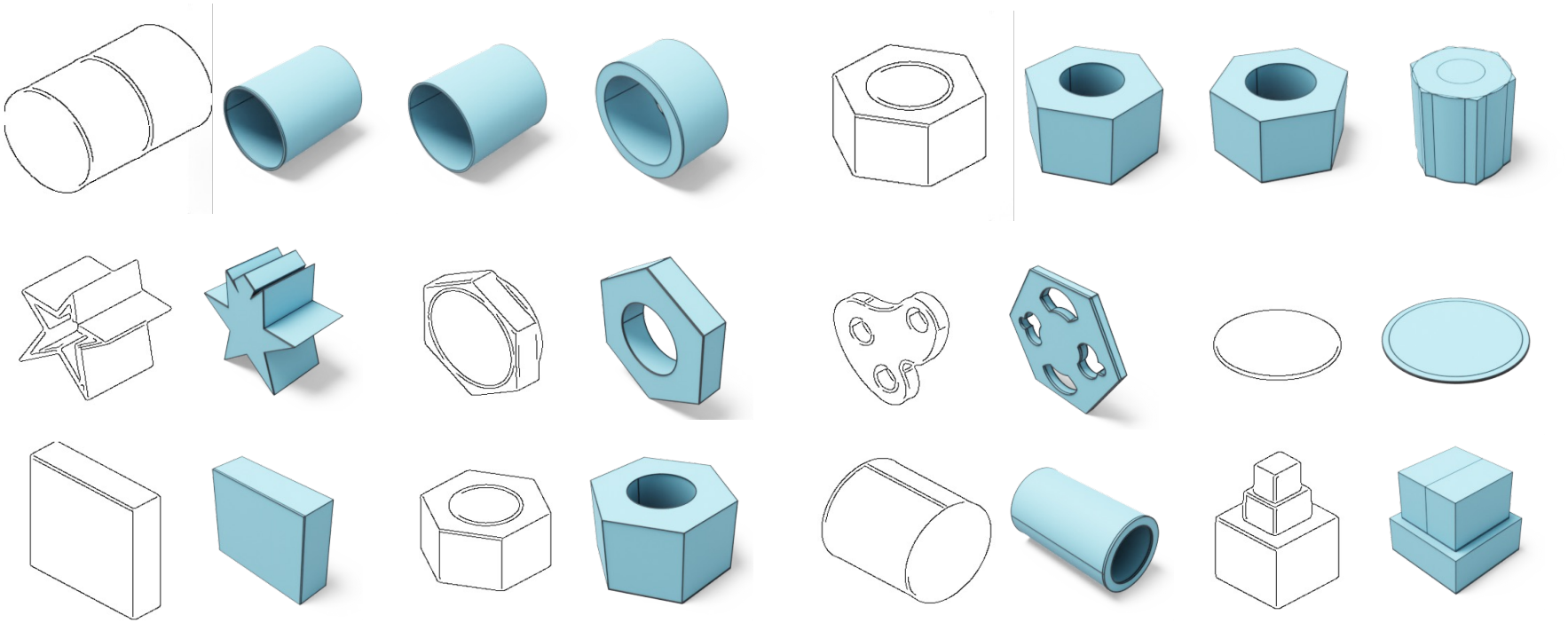
Results



Results: implicit

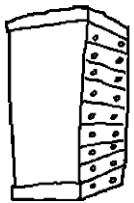


Results: CAD



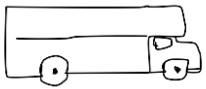
Datasets

- Training datasets: ShapeNet, DeepCAD
- Evaluation sketch datasets



ShapeNet-Sketch

ImageNet-Sketch

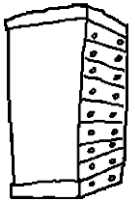


TU-Berlin

QuickDraw

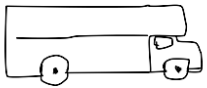
Quantitative evaluation

- Human perceptual evaluation
- Comparison with supervised methods
  - SketchSampler
  - Sketch2Model
- Metrics
  - Accuracy
  - IoU



ShapeNet-Sketch
3D ground truth

ImageNet-Sketch



TU-Berlin

QuickDraw

Human evaluation

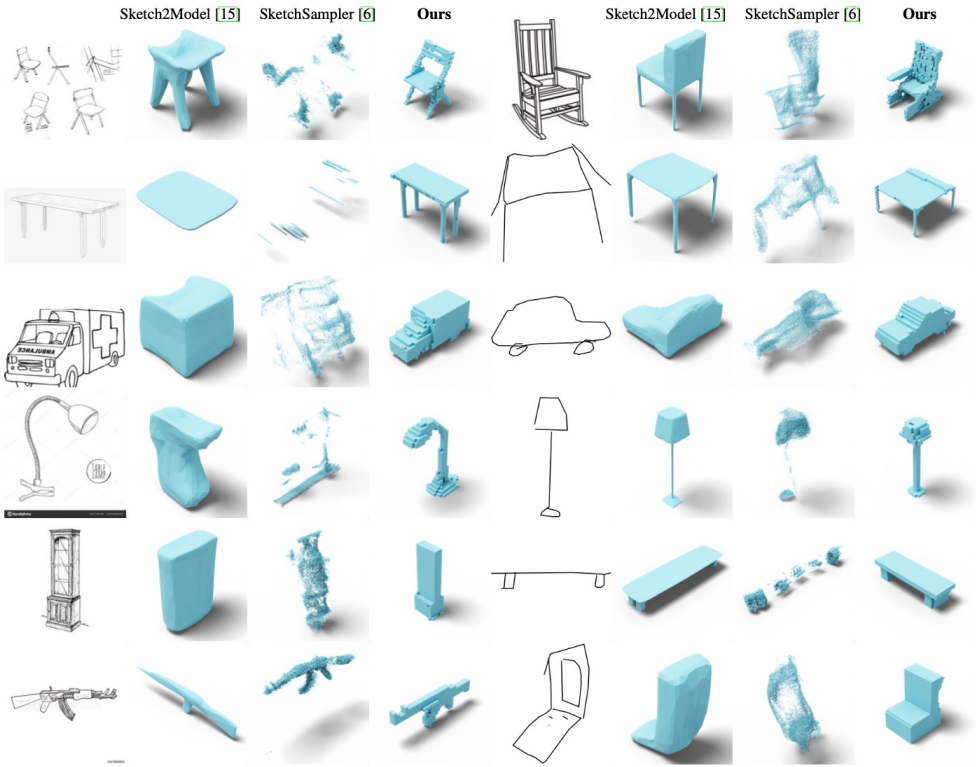
Which of the 3D models on the right hand side best matches the sketch on the left hand side?

Dataset	% correctly identified
All	71.1%
TU-Berlin	74.9%
ShapeNet-Sketch	73.1%
ImageNet-Sketch	68.1%
QuickDraw	67.9%

Comparisons

Method	Type	IOU ↑
Sketch2Mesh [7]	Supervised	0.195
Sketch2Model [15]	Supervised	0.205
Sketch2Point [13]	Supervised	0.163
SketchSampler [6]	Supervised	0.244
ours	Zero-shot	0.292

Method	QD-Acc↑	TU-Acc↑	SS-Acc↑	IS-Acc↑
Point·E	12.6	40.1	43.2	18.9
S2M	27.4	19.8	26.0	12.0
Ours	**58.8**	**81.5**	**79.7**	**74.2**



Why does this work?

- Pre-trained model semantic understanding
  - Local grid features
  - Size
  - Training dataset

Resolution	*CFG*	*Network*	*Dataset*	QD-Acc↑	TU-Acc↑	SS-Acc↑	IS-Acc↑
1 x 512	×	B-32 [57]	OpenAI [57]	36.65	61.14	62.86	55.96
50 x 768	×	B-32 [57]	OpenAI [57]	37.85	63.25	63.78	52.79
50 x 768	✓	B-32 [57]	OpenAI [57]	38.86	65.86	67.36	49.19
197 x 768	✓	B-16 [57]	OpenAI [57]	38.47	71.66	70.72	61.10
257 x 1024	✓	L-14 [57]	OpenAI [57]	**55.45**	77.15	**74.53**	**69.06**
144 x 3072	✓	RN50x16 [57]	OpenAI [57]	34.61	70.81	58.82	59.00
196 x 4096	✓	RN50x64 [57]	OpenAI [57]	46.93	73.79	59.41	64.19
257 x 1024	✓	Open-L-14 [27]	LAION-2B [64]	54.63	**77.60**	69.03	68.35
256 x 1024	✓	DINO-L-14 [53]	DINOv2 [53]	39.73	71.12	72.10	55.94
197 x 1024	✓	MAE-L [22]	ImageNet [11]	19.31	30.52	38.79	26.65
257 x 1280	✓	MAE-H [22]	ImageNet [11]	18.70	31.63	37.47	31.42

Why does this work?

- Pre-trained model semantic understanding
  - Local grid features
  - Size
  - Training dataset
- Rendering from several points of view
- Data augmentation

Conclusion & Future work

- 3D generative model conditioned on local features can do sketch to 3D
- Different abstraction
- Multiple 3D representation
- More data to be able to generate almost everything