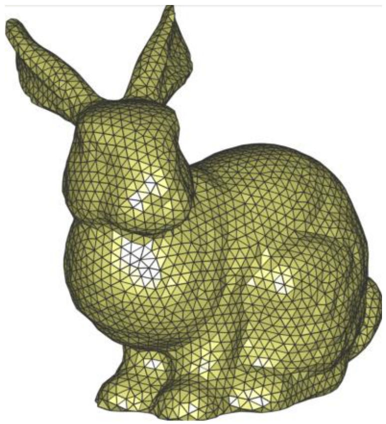
Data-Driven Geometry Processing
3D Deep Learning II

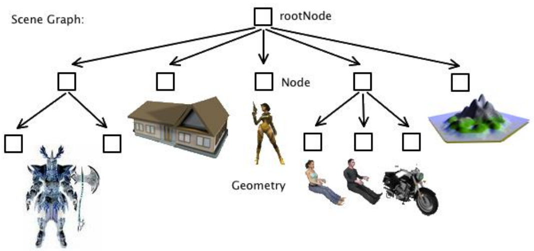
Qixing Huang

May 2th 2018

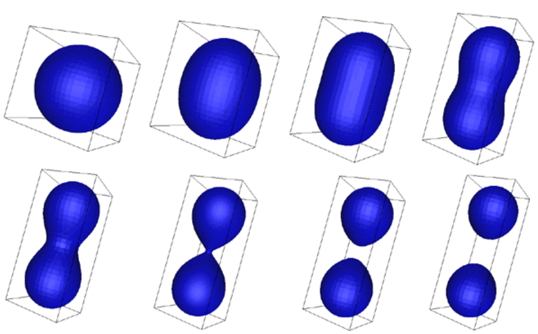
3D Surface Representations



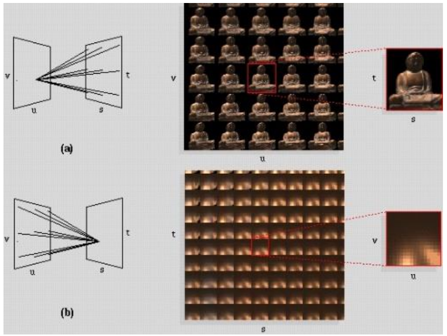
Triangular mesh



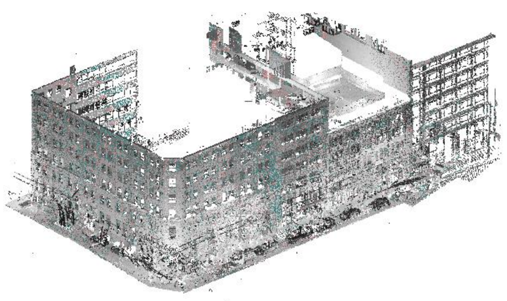
Part-based models



Implicit surface



Light Field Representation

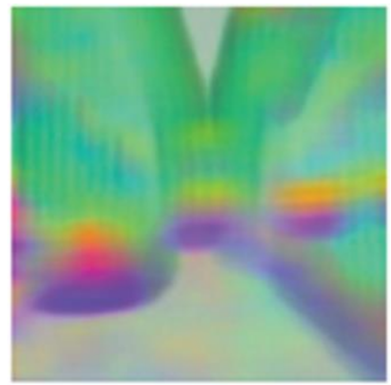


Point cloud

Matching in Embedding Spaces [CVPR’ 16]

Existing methods usually follow a two-step approach (e.g., SIFT flow)

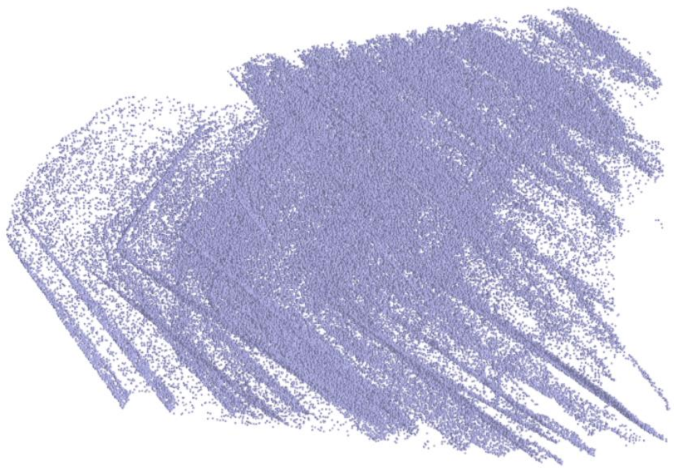
- Local descriptor computation



- Dense pixel labeling via MRF inference
  - Preserve descriptors
  - Preserve smoothness

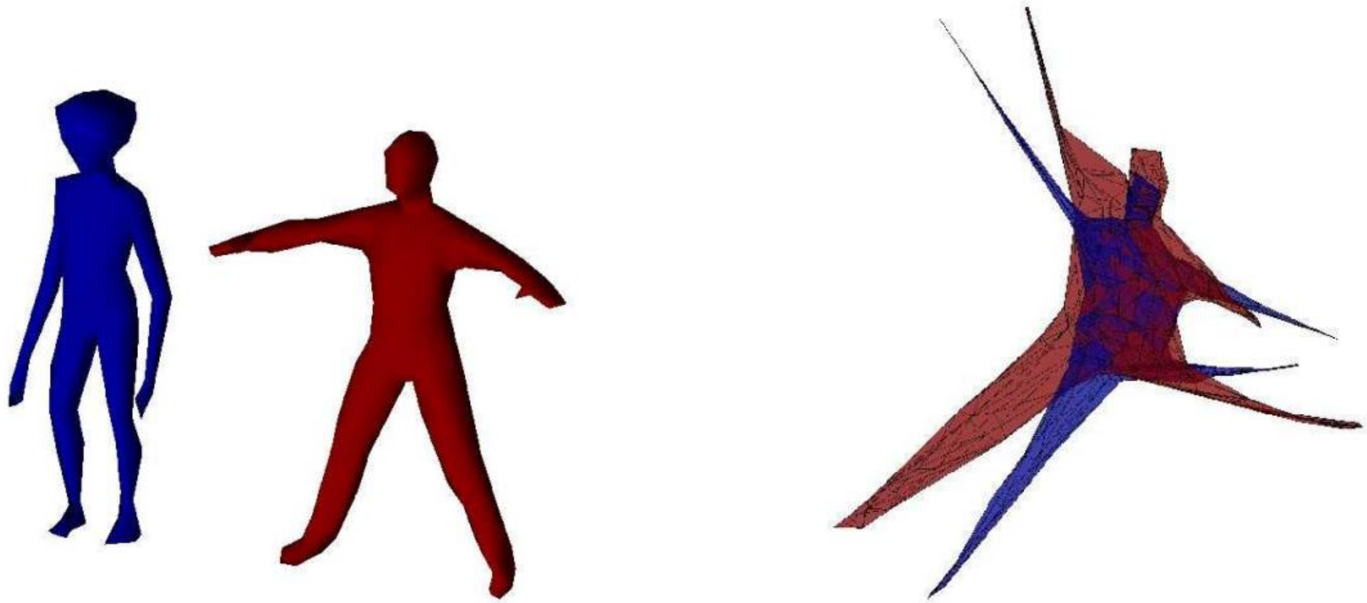
Issues of such two-step approach

Partial similarity



Inefficient when
matching multiple objects

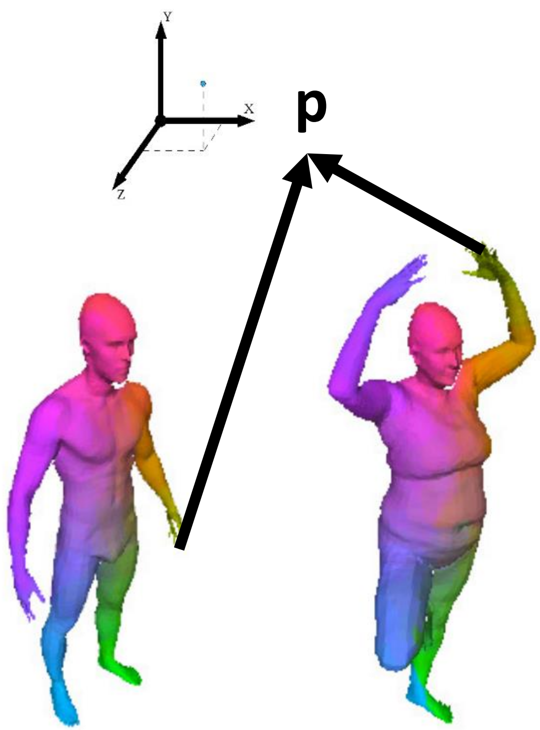
Embedding --- establishing correspondences in the embedding space



Spectral embedding [Liu et al. 06]

Sensitive to 1) partial similarity, and 2) geometric and topological changes

Properties of the desired embedding space



Corresponding points are matched in the embedding space

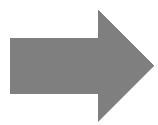
Embedding preserves continuity

The benefits of object embedding

- Correspondences become nearest neighbor query
  - Efficiency for multiple object matching
    O(n) embeddings + $O(n^2)$ queries
  - Partial similarity
  - Fuzzy correspondences

The biggest message of deep neural networks

- Approximate any function given sufficient data



Focus on depth images

- Scanning devices generate depth images

- Complete shape embedding are aggregated from depth image embeddings
  – 3D convolution is not ready yet

Architecture

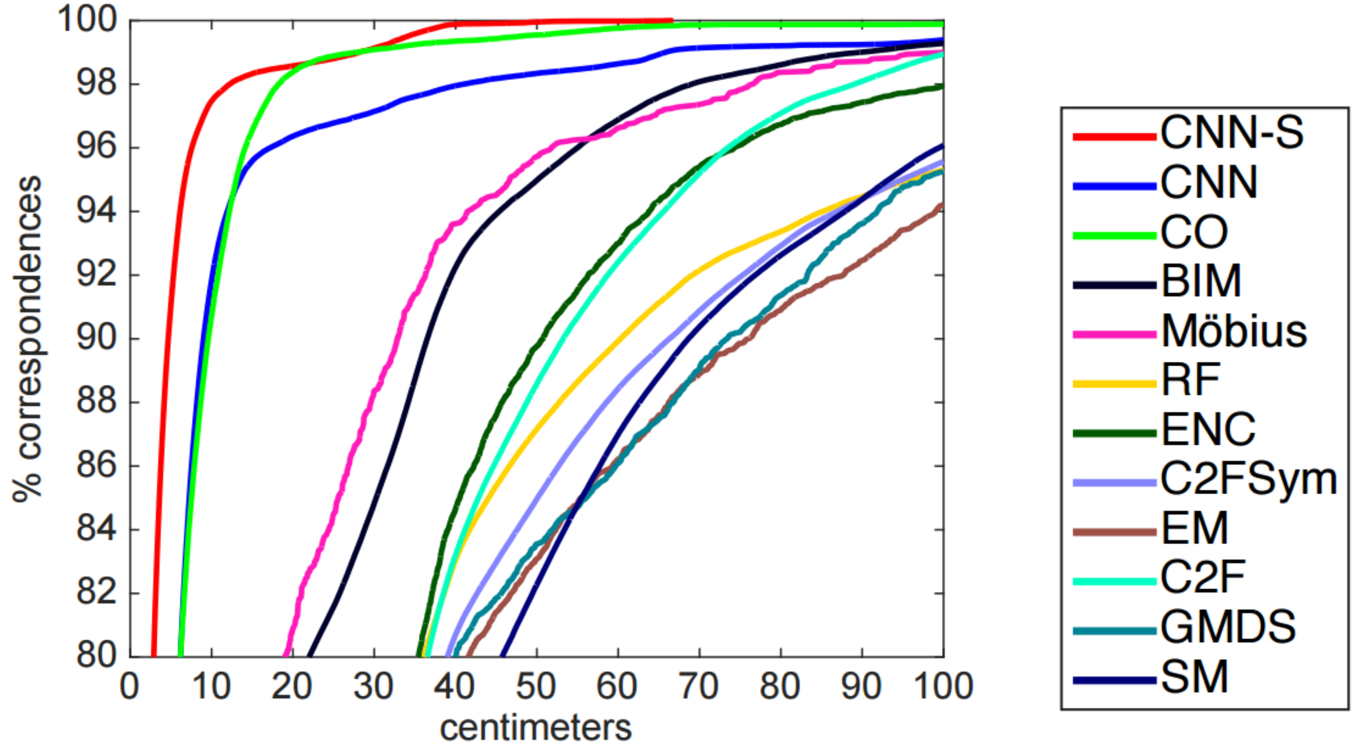
	0	1	2	3	4	5	6	7	8	9	10
layer	**image**	**conv**	**max**	**conv**	**max**	**2×conv**	**conv**	**max**	**2×conv**	**int**	**conv**
filter-stride	-	11-4	3-2	5-1	3-2	3-1	3-1	3-2	1-1	-	3-1
channel	1	96	96	256	256	384	256	256	4096	4096	16
activation	-	relu	lrn	relu	lrn	relu	relu	idn	relu	idn	relu
size	512	128	64	64	32	32	32	16	16	128	512
num	1	1	4	4	16	16	16	64	64	1	1

The input is a depth image

The output is a per-pixel descriptor (dim 16)

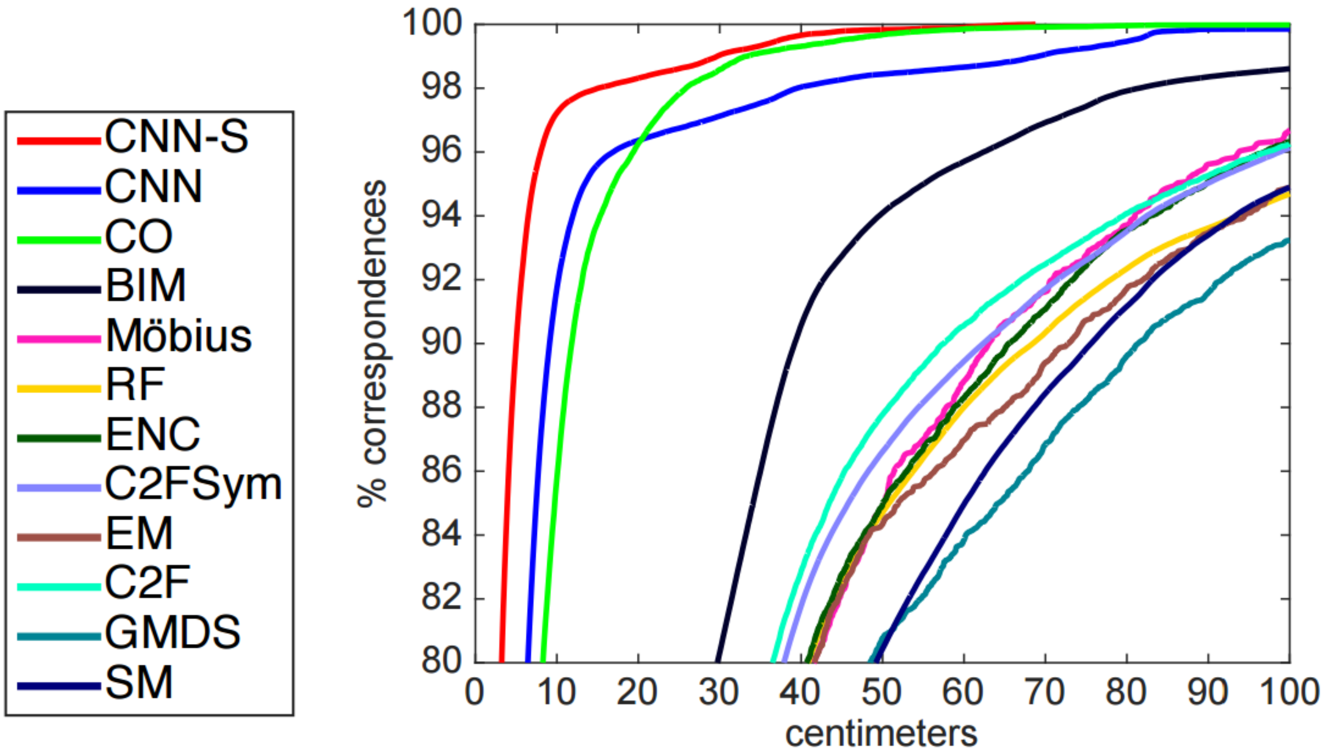
Convolution + Deconvolution

Evaluation on the FAUST dataset



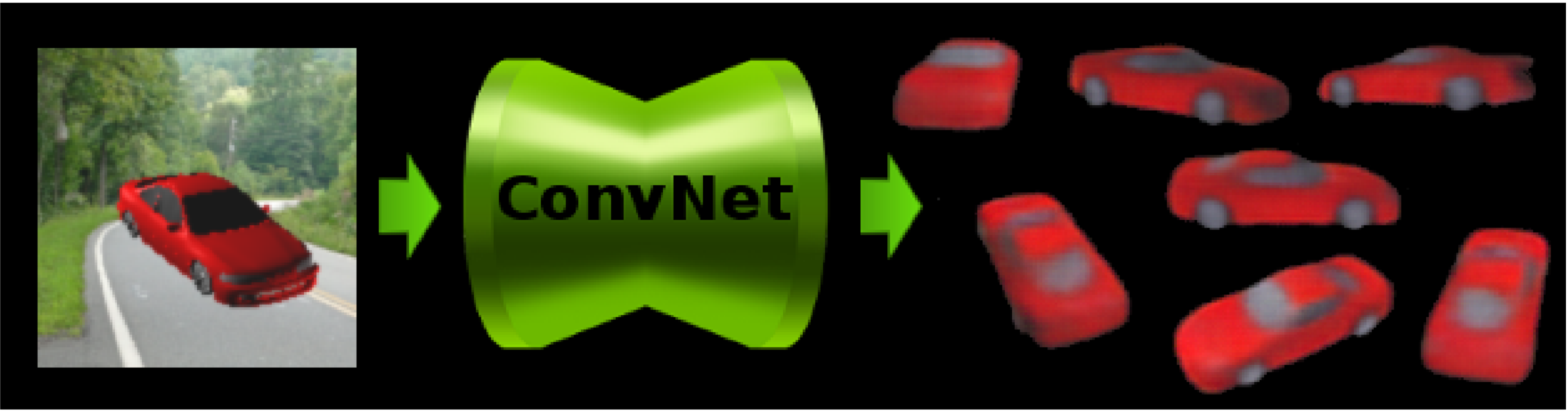
Cumulative error distribution, intra-subject

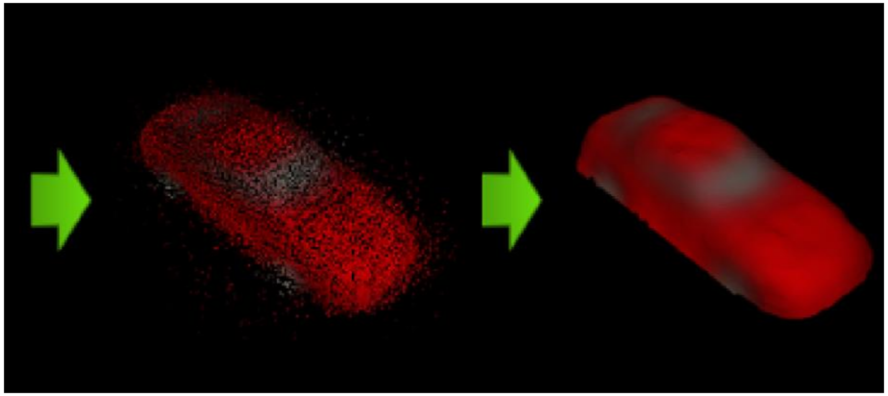
Evaluation on the FAUST dataset

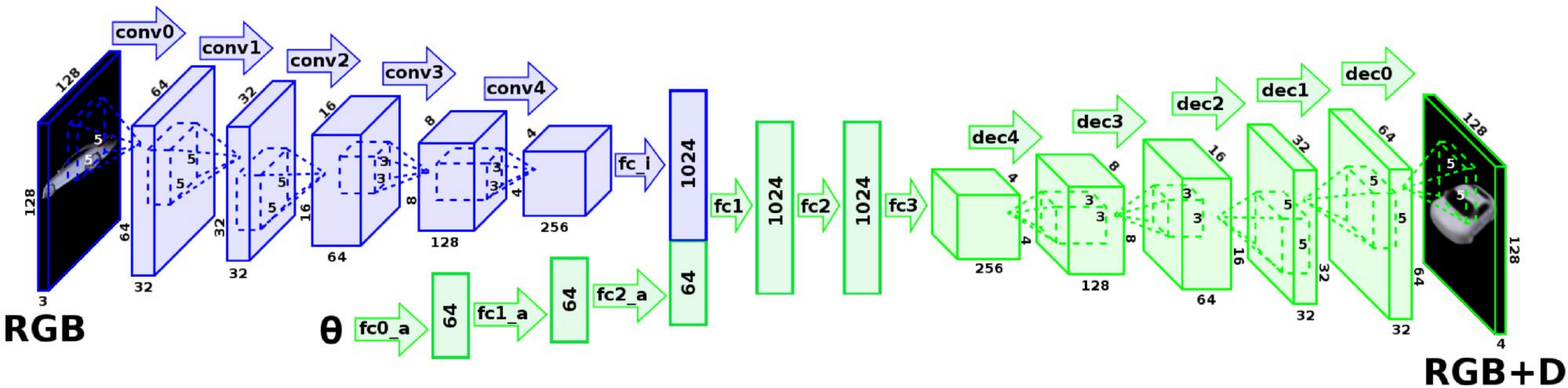


Cumulative error distribution, inter-subject

Multi-view 3D Models from Single Images With a Convolutional Network [ECCV’ 16]


ConvNet





RGB
conv0
conv1
conv2
conv3
conv4
fc_i
1024
64
θ
fc0_a
fc1_a
fc2_a
fc1
fc2
fc3
1024
1024
dec4
dec3
dec2
dec1
dec0
RGB+D

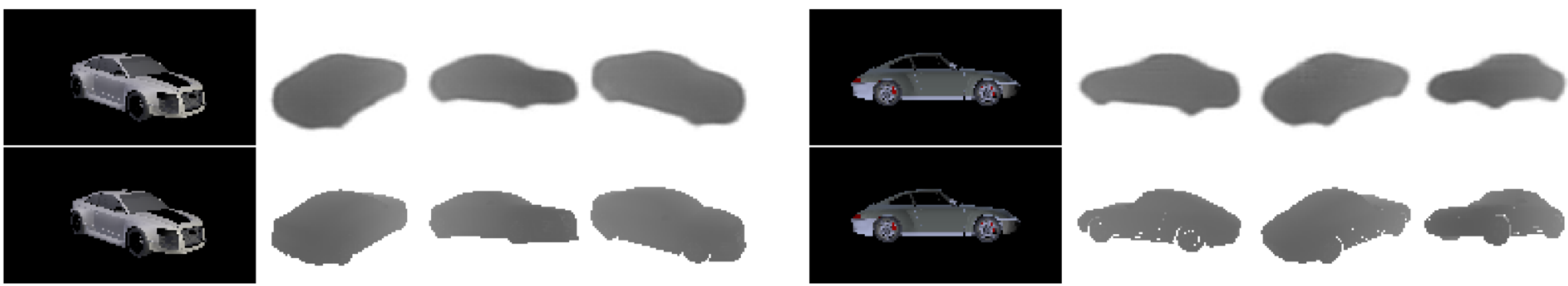


Fig. 5. Depth map predictions (**top row**) and the corresponding ground truth (**bottom row**). The network correctly estimates the shape.

Multi-view 3D Models from Single Images with a Convolutional Network

Maxim Tatarchenko, Alexey Dosovitskiy, Thomas Brox

Department of Computer Science
University of Freiburg
{tatarchm, dosovits, brox}@cs.uni-freiburg.de

ECCV 2016

Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction without 3D Supervision [Yan et al. 16]

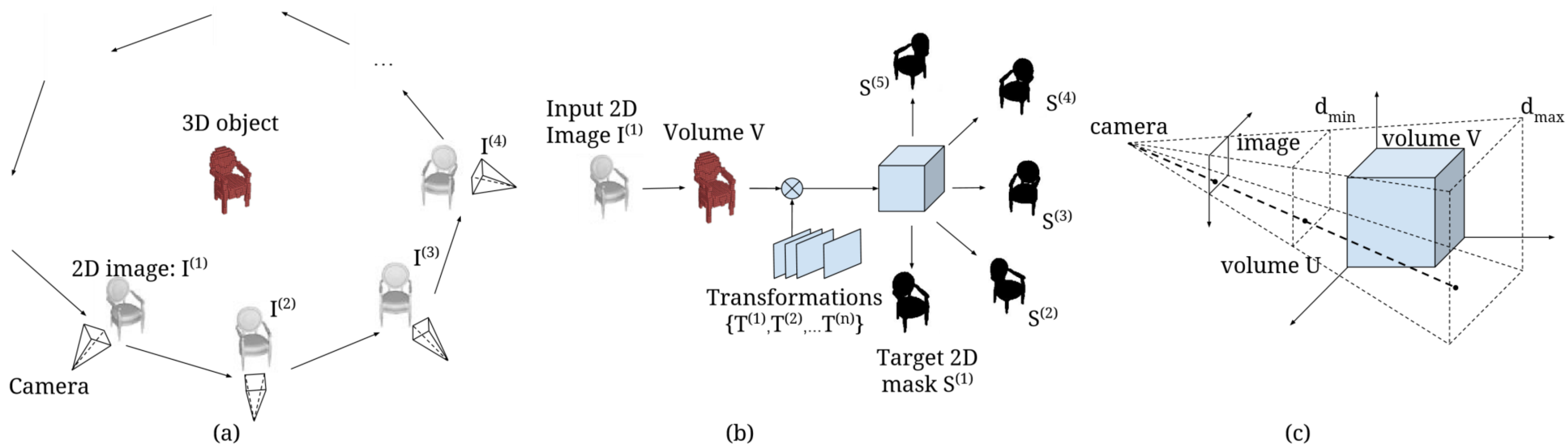
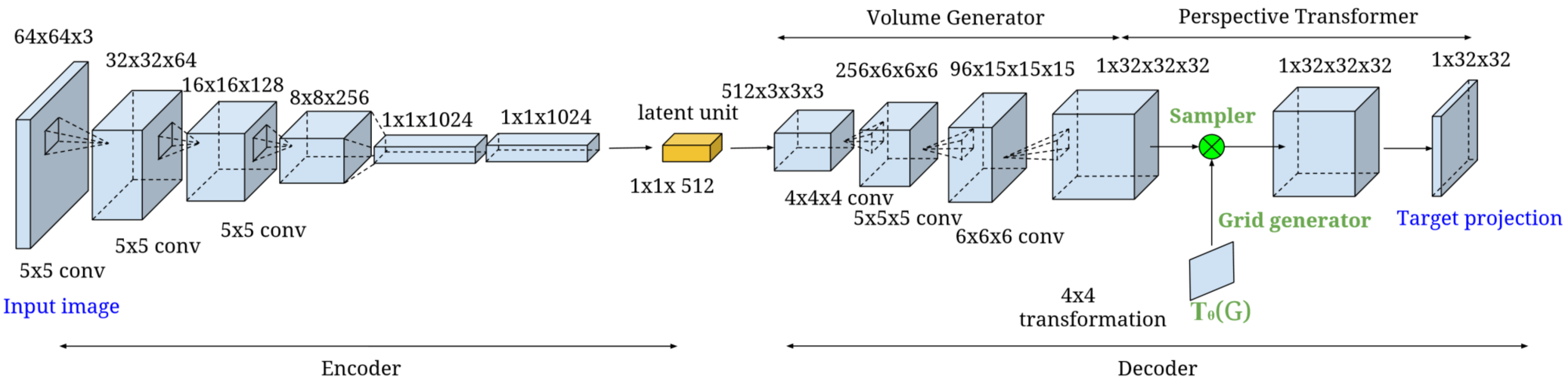


Figure 1: (a) Understanding 3D object from learning agent's perspective; (b) Single-view 3D volume reconstruction with perspective transformation. (c) Illustration of perspective projection. The minimum and maximum disparity in the screen coordinates are denoted as d_{min} and d_{max}.

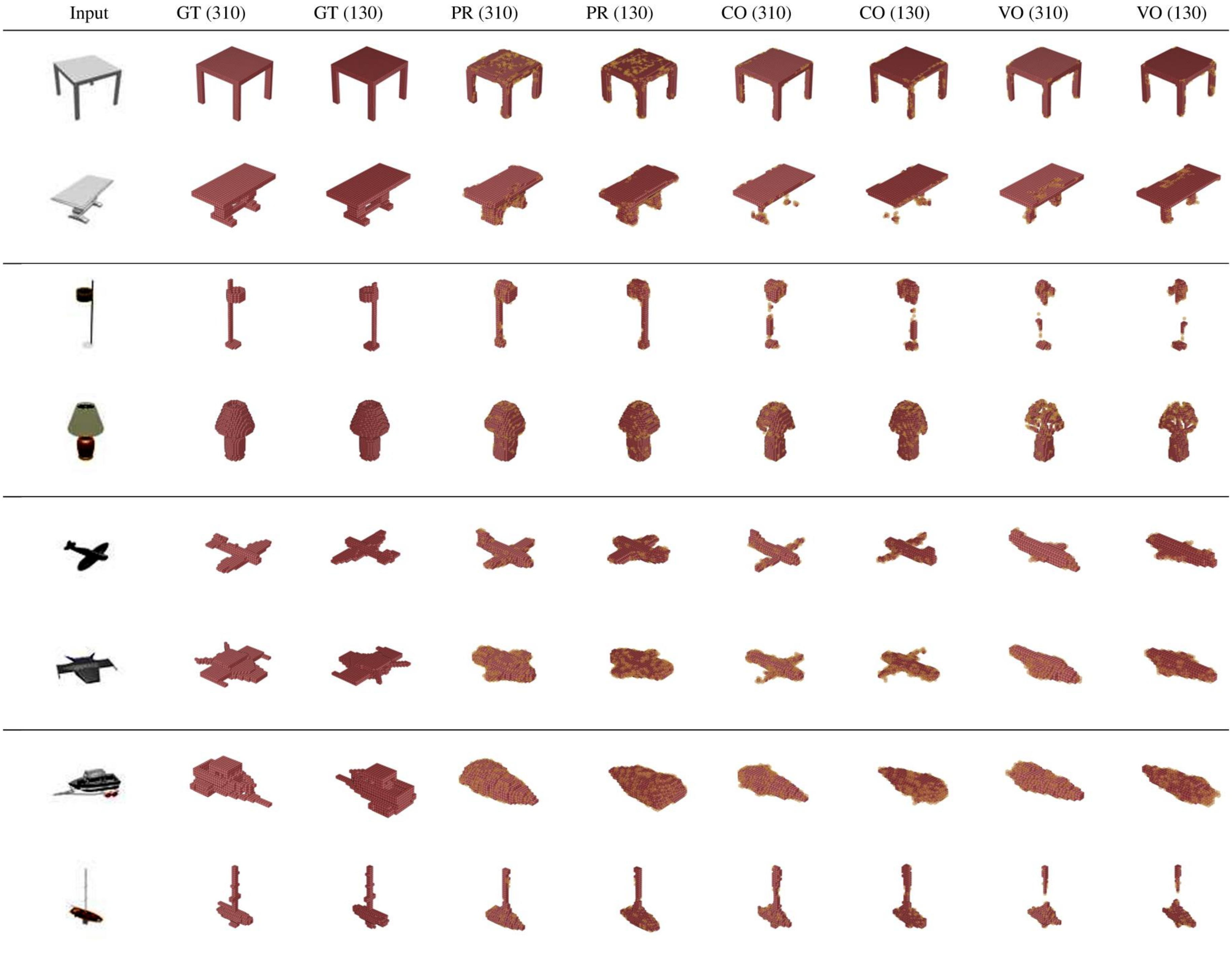
$$\mathcal{L}_{vol}(I^{(k)}) = ||f(I^{(k)}) - \mathbf{V}||_2^2$$

$$\mathcal{L}_{proj}(I^{(k)}) = \sum_{j=1}^{n} \mathcal{L}_{proj}^{(j)}(I^{(k)}; S^{(j)}, \alpha^{(j)}) = \frac{1}{n}\sum_{j=1}^{n} ||P(f(I^{(k)}); \alpha^{(j)}) - S^{(j)}||_2^2$$

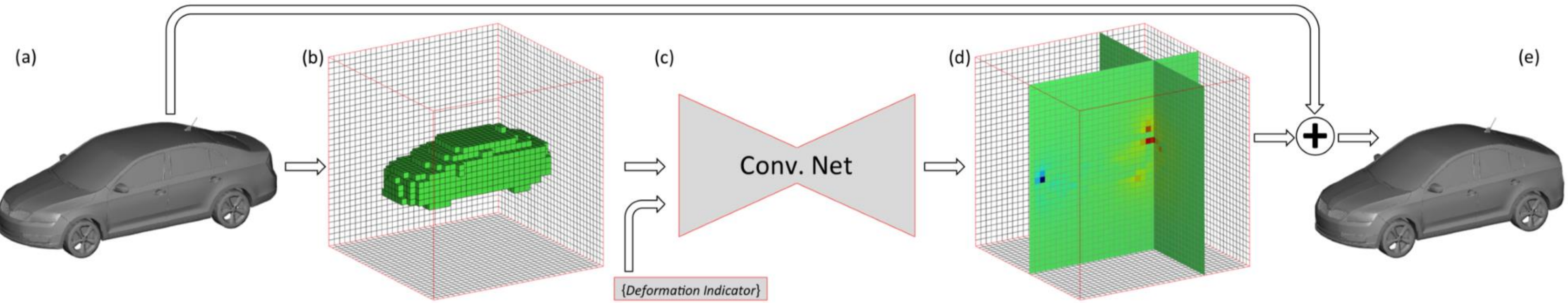
$$\mathcal{L}_{comb}(I^{(k)}) = \lambda_{proj}\mathcal{L}_{proj}(I^{(k)}) + \lambda_{vol}\mathcal{L}_{vol}(I^{(k)})$$

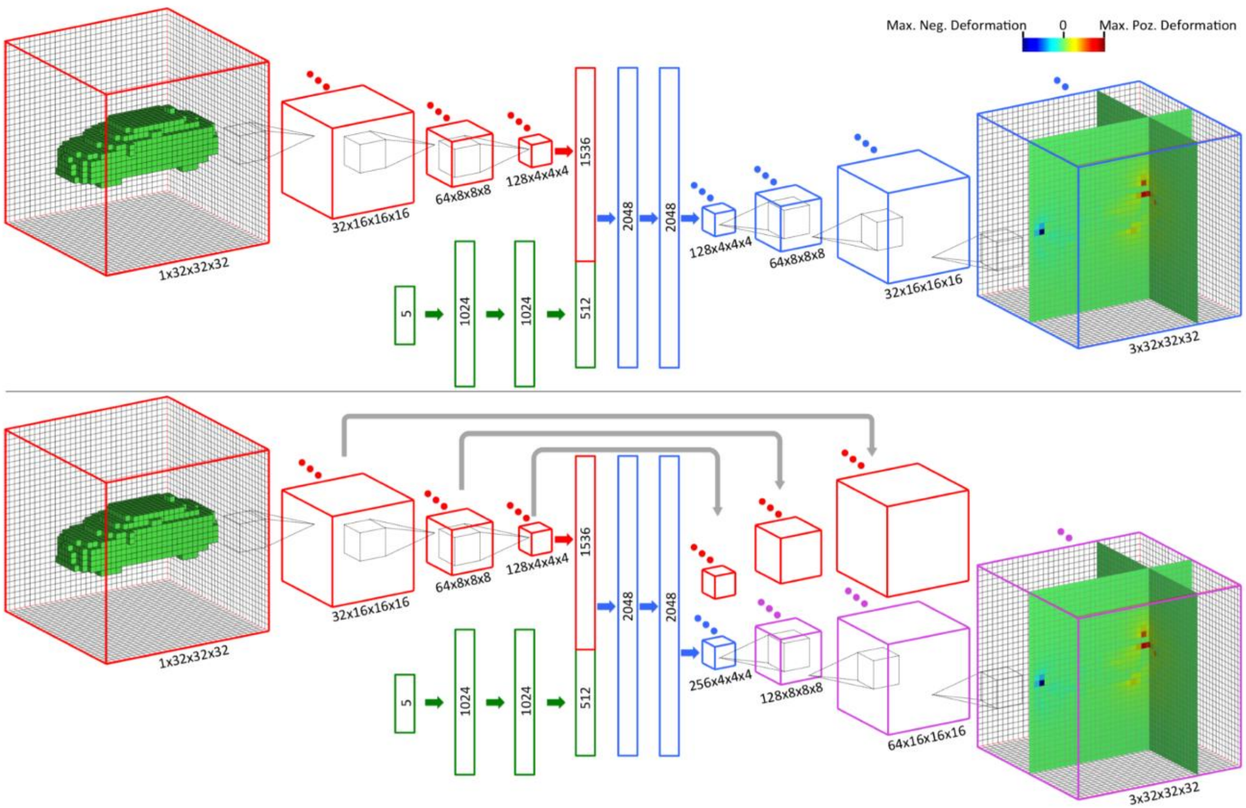


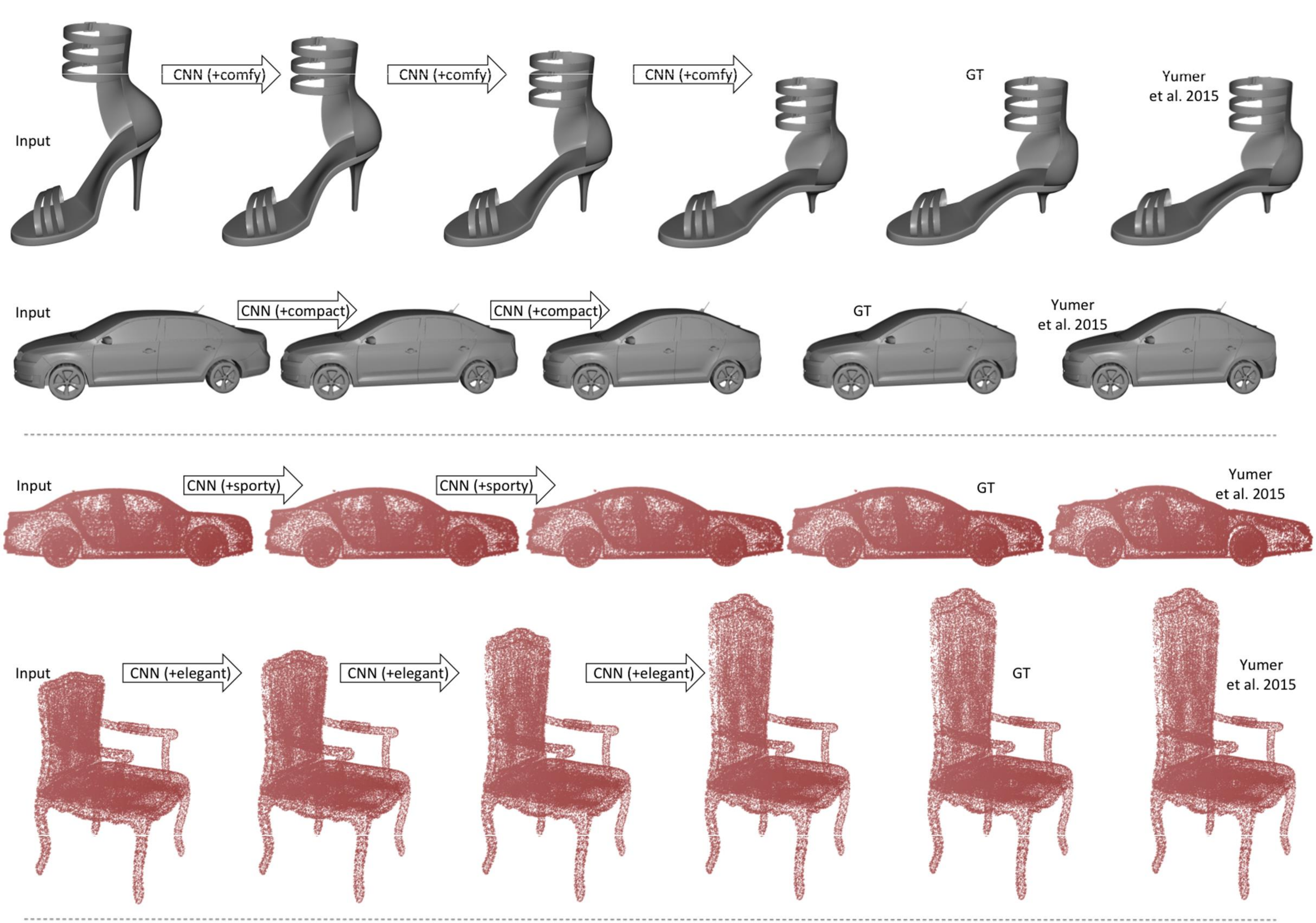
64x64x3
32x32x64
16x16x128
8x8x256
1x1x1024
1x1x1024
latent unit
1x1x 512
512x3x3x3
256x6x6x6
96x15x15x15
1x32x32x32
1x32x32x32
1x32x32
Volume Generator
Perspective Transformer
Sampler
Grid generator
Target projection
4x4x4 conv
5x5x5 conv
6x6x6 conv
4x4
transformation
Tθ(G)
5x5 conv
5x5 conv
5x5 conv
Input image
Encoder
Decoder


Input
GT (310)
GT (130)
PR (310)
PR (130)
CO (310)
CO (130)
VO (310)
VO (130)

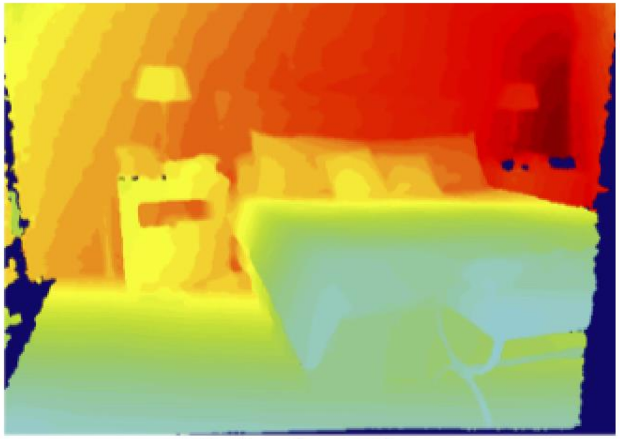
Learning Semantic Deformation Flows with 3D Convolutional Networks [Yumer and Mitra 2016]


(a)
(b)
(c)
(d)
(e)
Conv. Net
{Deformation Indicator}


Max. Neg. Deformation
0
Max. Poz. Deformation
1x32x32x32
32x16x16x16
64x8x8x8
128x4x4x4
1536
512
5
1024
1024
2048
2048
128x4x4x4
64x8x8x8
32x16x16x16
3x32x32x32
1x32x32x32
32x16x16x16
64x8x8x8
128x4x4x4
1536
512
5
1024
1024
2048
2048
256x4x4x4
128x8x8x8
64x16x16x16
3x32x32x32

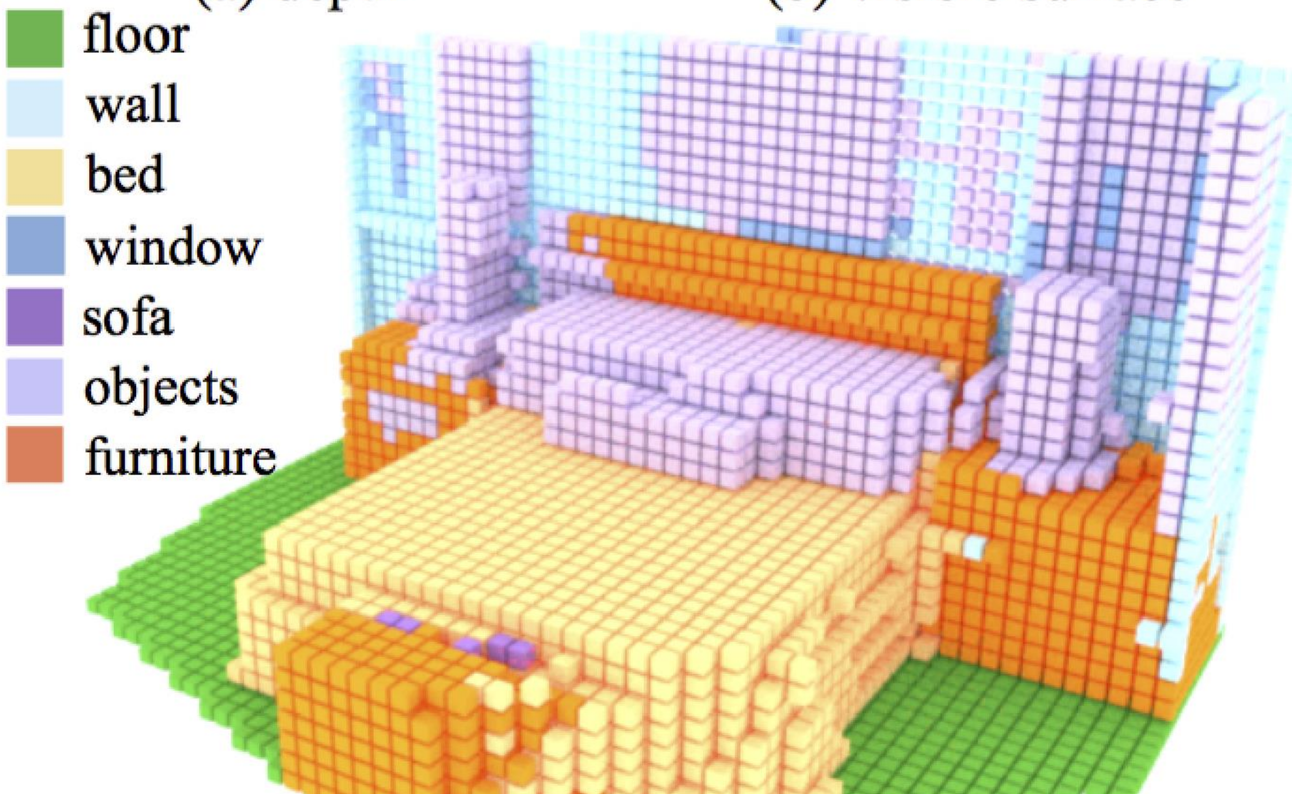

Input
CNN (+comfy)
CNN (+comfy)
CNN (+comfy)
GT
Yumer et al. 2015
Input
CNN (+compact)
CNN (+compact)
GT
Yumer et al. 2015
Input
CNN (+sporty)
CNN (+sporty)
GT
Yumer et al. 2015
Input
CNN (+elegant)
CNN (+elegant)
CNN (+elegant)
GT
Yumer et al. 2015

Semantic Scene Completion from a Single Depth Image [Song et al. 17]

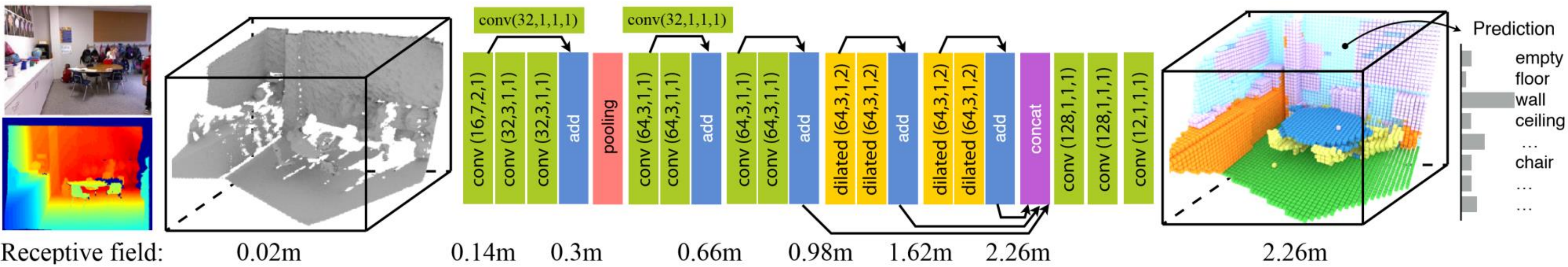


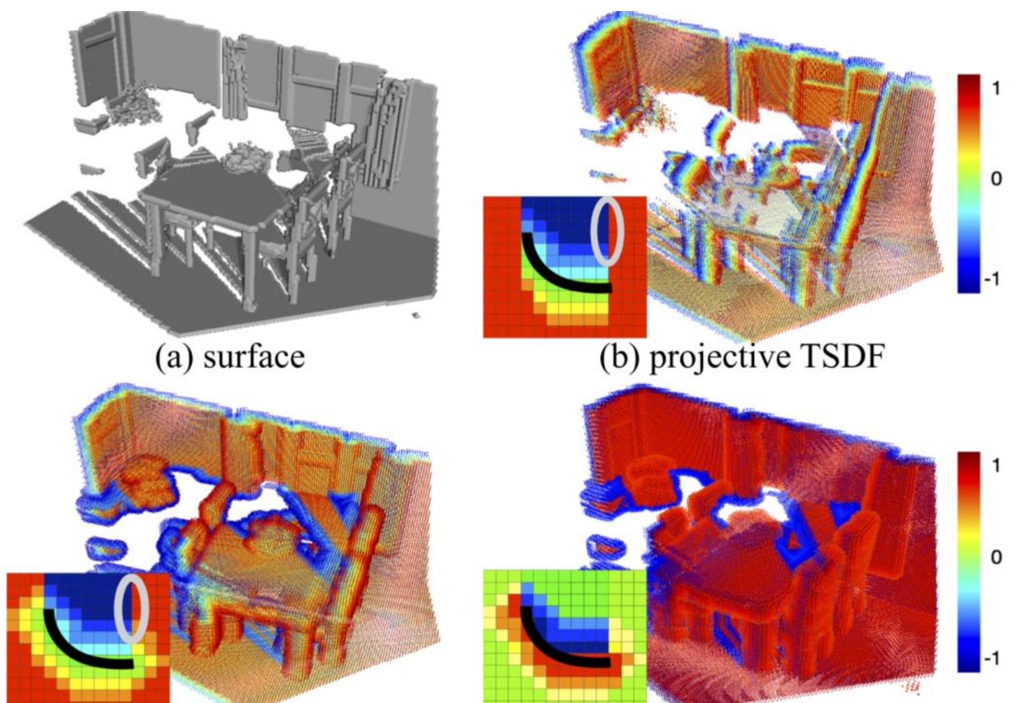
(a) depth

(b) visible surface



(c) output



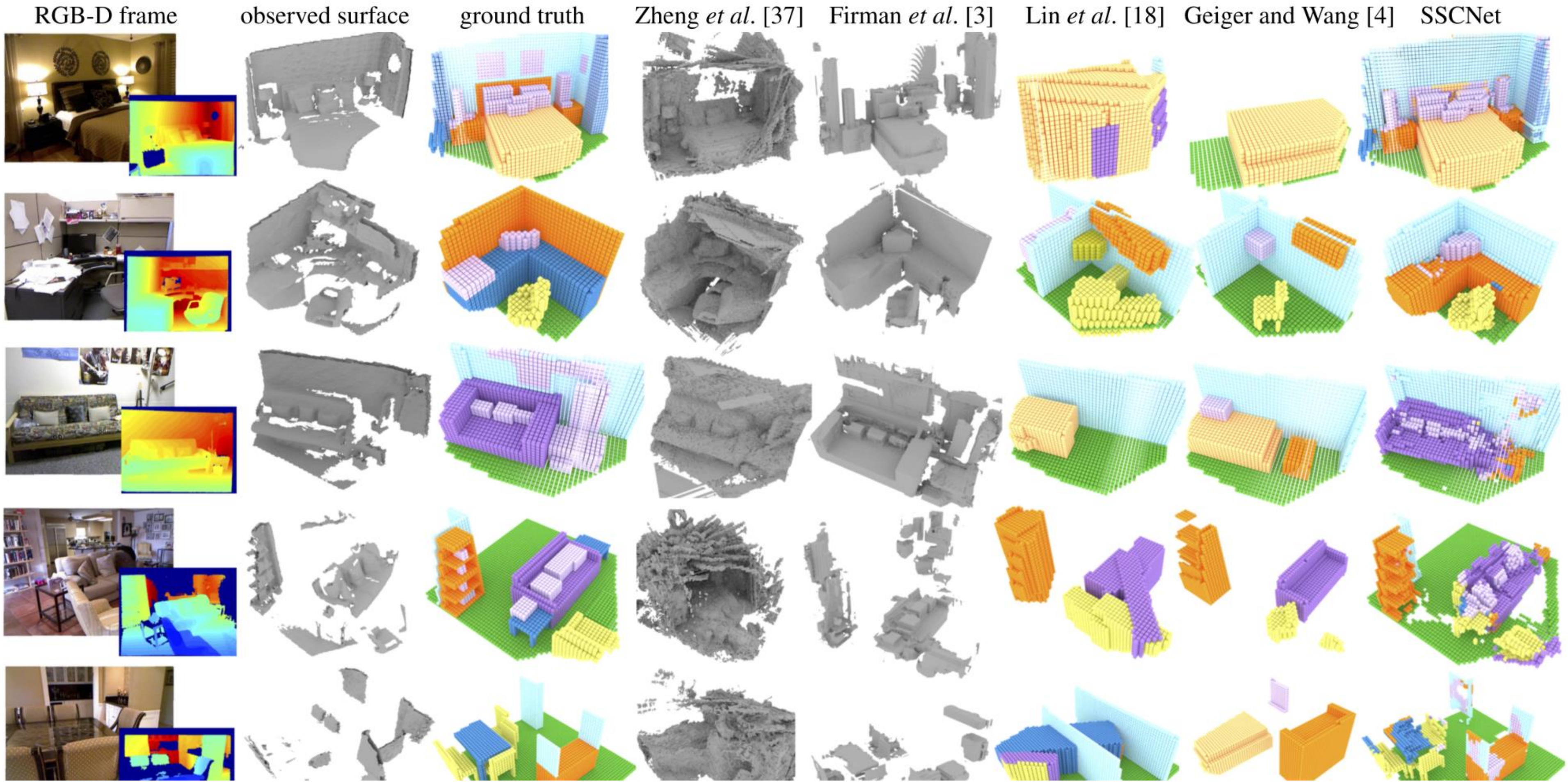


(a) surface

(b) projective TSDF

(c) TSDF

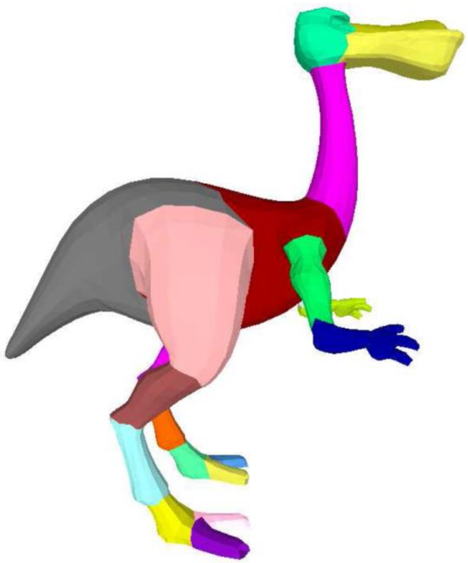
(d) flipped TSDF


RGB-D frame
observed surface
ground truth
Zheng et al. [37]
Firman et al. [3]
Lin et al. [18]
Geiger and Wang [4]
SSCNet

Other Topics (not Covered)

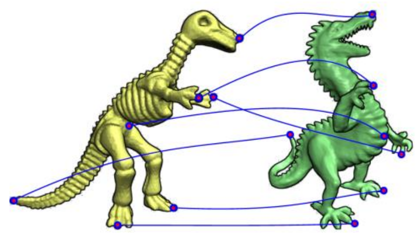
Shape Analysis

- Design algorithms to extract semantic information from one or a collection of shapes



[Karz and Tal 03]

Segmentation



[van Kaick et al. 11]

Matching

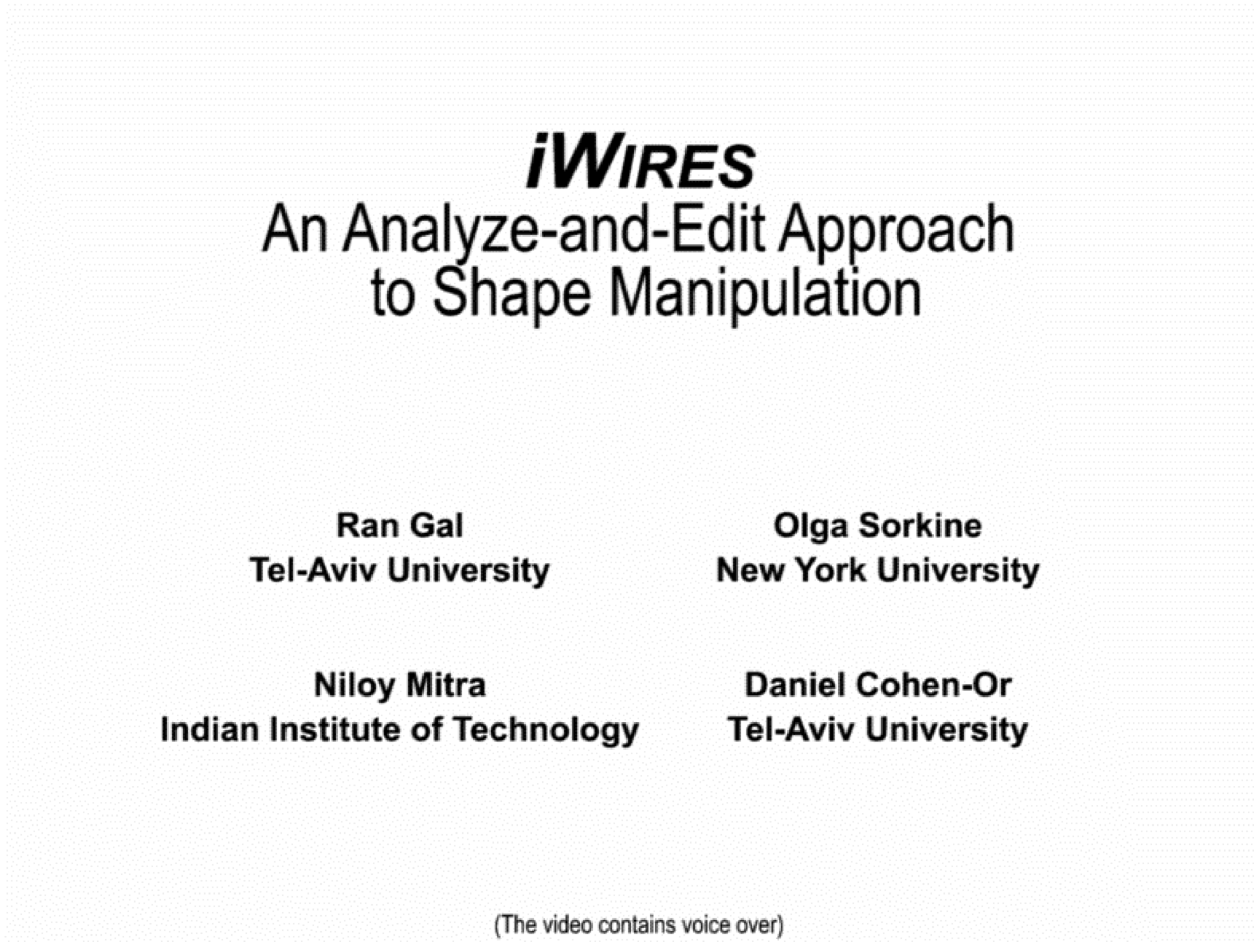
[Funkhouser et al. 05]

Retrieval

[Mitra et al. 06]

Classification & Clustering

Shape Modeling

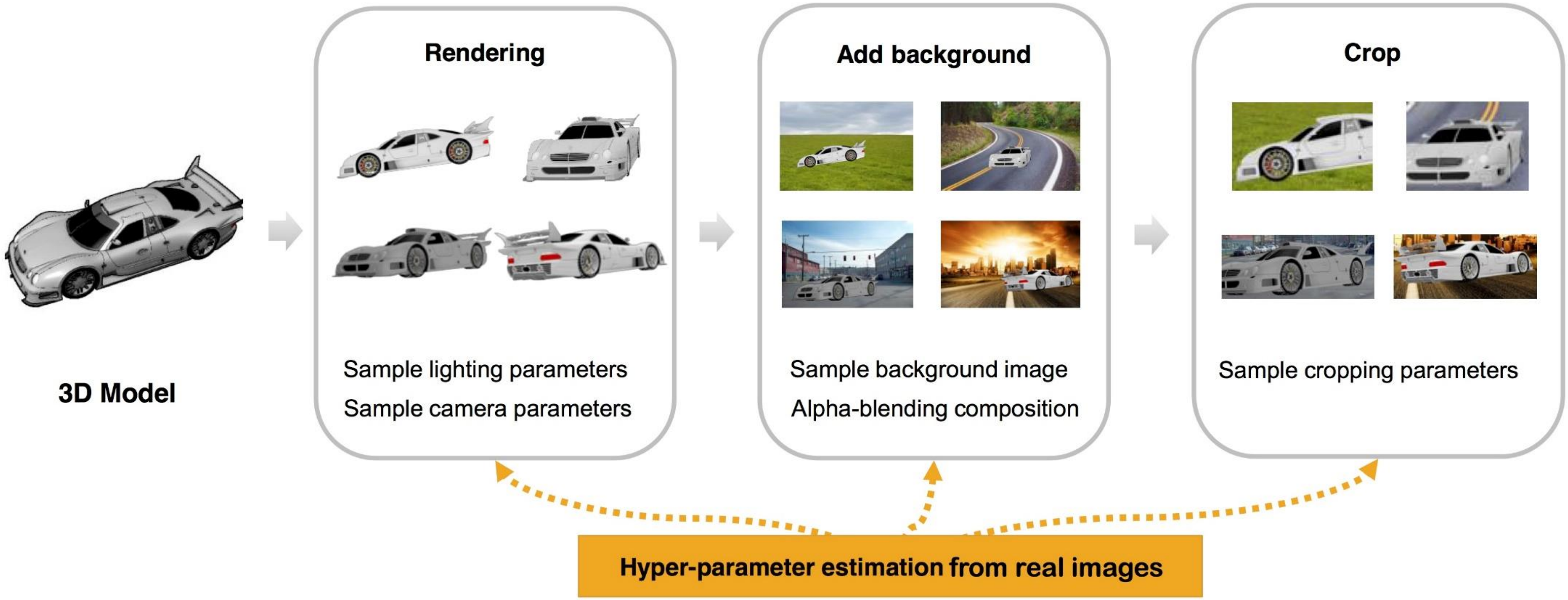


Character Animation

Animating Human Dressing, Alex Clegg, Jia Tan, Greg Turk, and C. Karen Liu, SIGGRAPH 2015

Graphics & AI

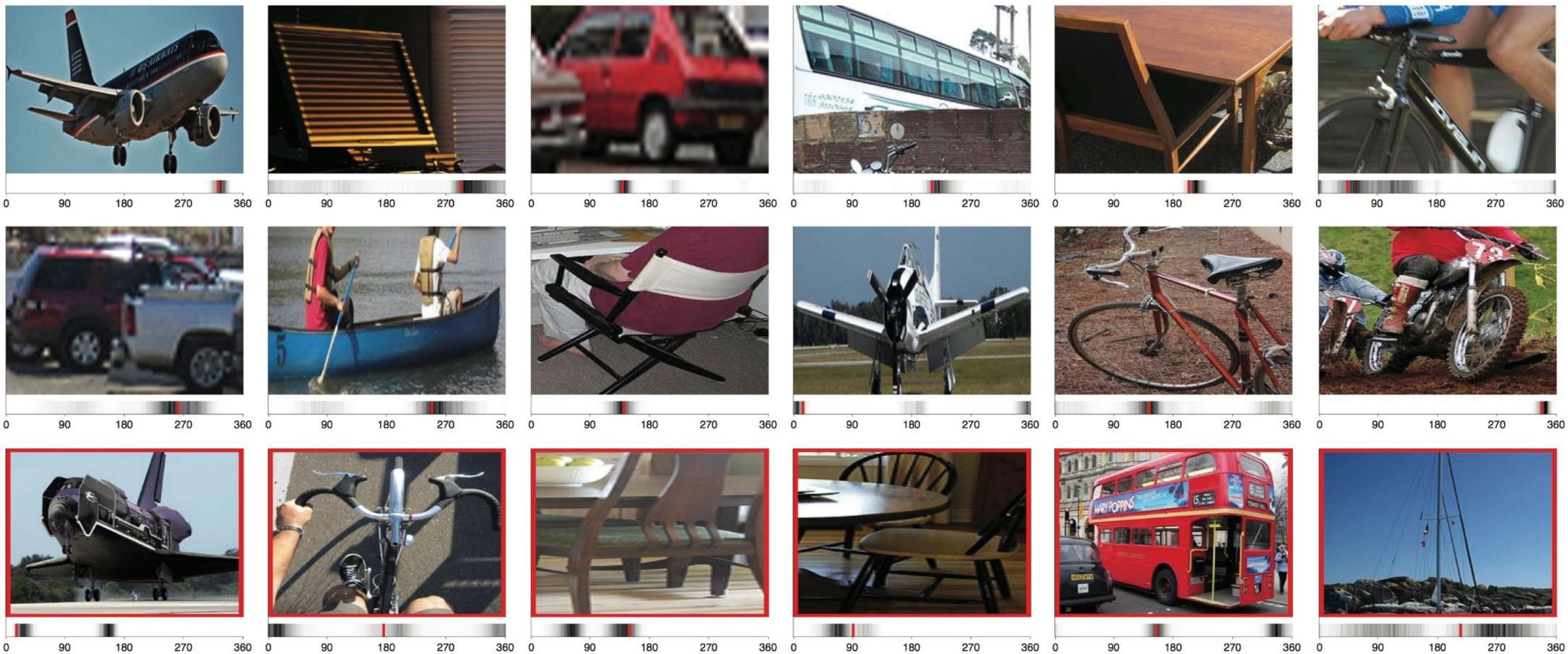
Render-for-CNN



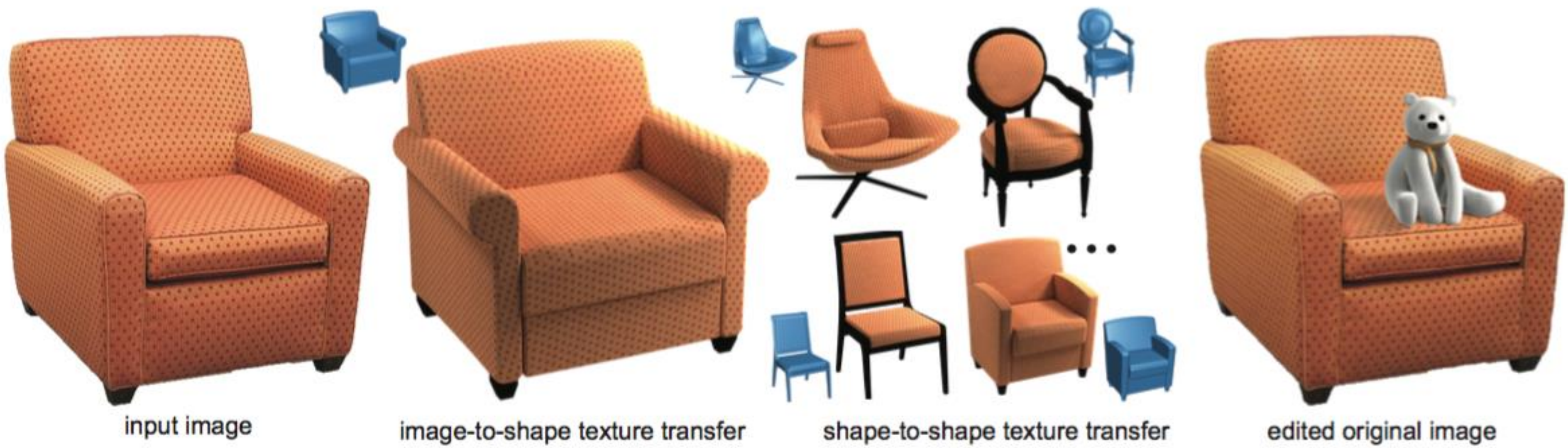
[Su et al. 15]

Synthetic Image Examples

Viewpoint Estimation Results



SIGGRAPH ASIA' 2016



Shape Captioning

- There is a bed with three pillows and a bedside table next to it.
- The room appears to be a bedroom. A blue bed and white nightstand are pushed against the furthest wall. A window is on the left side.
- A dark bedroom with a queen bed with blue comforter and three pillows. There is a night stand. One wall is decorated with a large design and another wall has three large windows.

- There is a chair and a circular table in the middle of a floral print room.
- a corner widow room with a a table and chair sitting to the east side.
- There's a dresser in the corner of the room, and a yellow table with a brown wooden chair.

Need a feature representation of 3D Scenes

Future project

Text-2-Scene Generation

[Chang et al. 15]

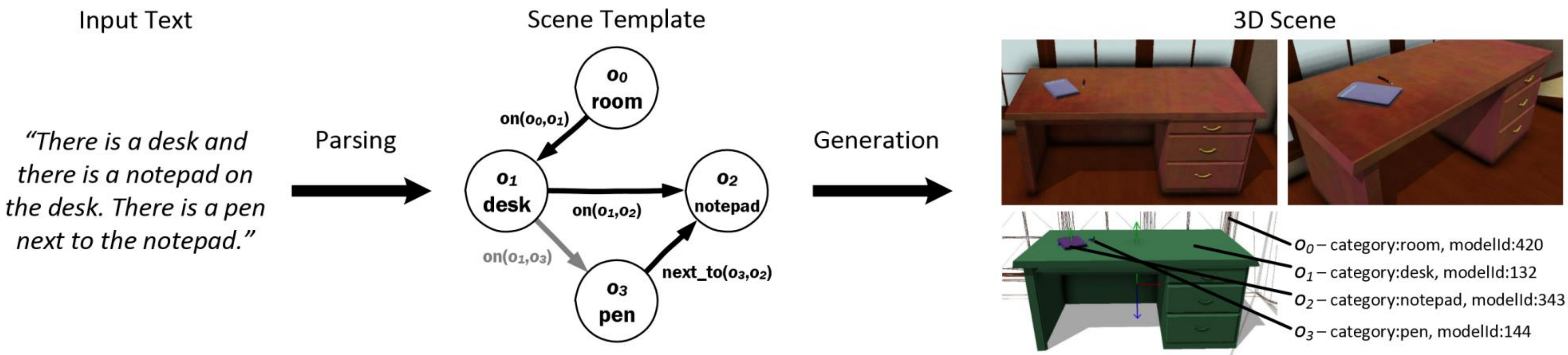
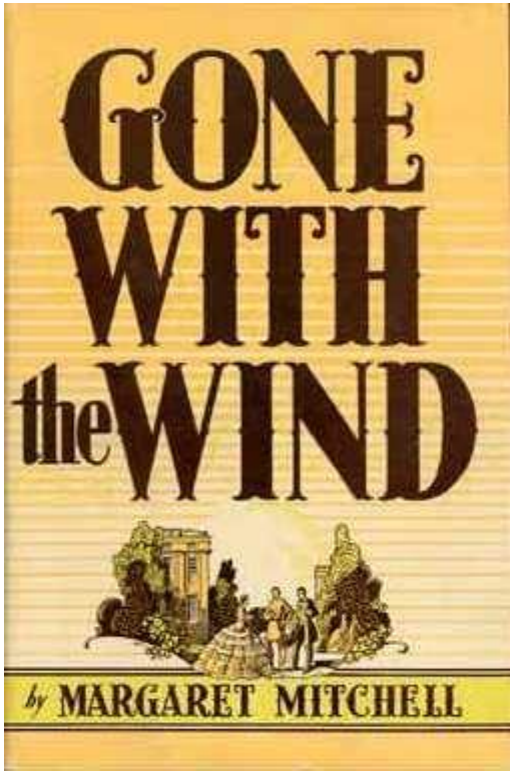


Figure 2: Illustration of the text to 3D scene generation pipeline. The input is text describing a scene (left), which we parse into an abstract scene template representation capturing objects and relations (middle). The scene template is then used to generate a concrete 3D scene visualizing the input description (right). The 3D scene is constructed by retrieving and arranging appropriate 3D models.

Text-2-Animation



With Joe Langus, Kevin Tai, and Raymood Mooney