max planck institut
informatik

UNIVERSITÄT
DES
SAARLANDES

Unsupervised Learning of 3D Object Shape and Camera Pose

Eldar Insafutdinov

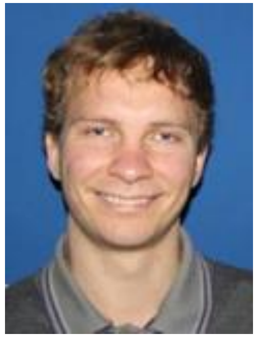
Max Planck Institute for Informatics

Skoltech Christmas Colloquium
Moscow, 26.12.2018

Background

- 3rd year PhD student at Max Planck Institute for Informatics
- I work on human pose estimation, detection and tracking

Micha
Andriluka



Leonid
Pishchulin

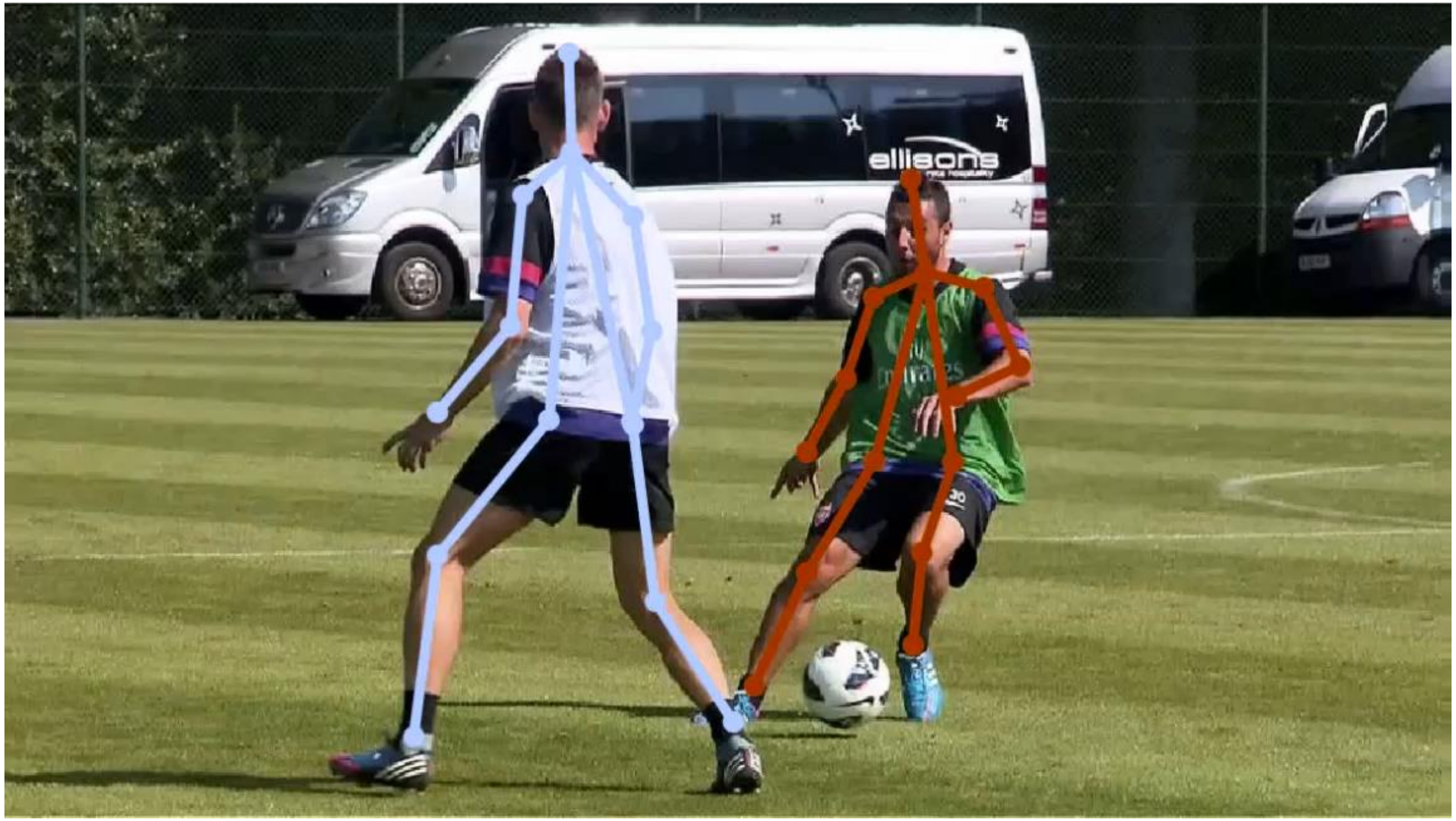
Evgeny
Levinkov

Siyu
Tang

Björn
Andres

Bernt
Schiele

Analyzing Humans in Unconstrained Scenes



[1] DeeperCut: a deeper, stronger and faster multi-person pose estimation model. *Insafutdinov et al*. ECCV 2016

[2] Arttrack: Articulated multi-person tracking in the wild. *Insafutdinov et al*. CVPR 2017

Supervised Learning

- Pose Estimation (also detection and segmentation) is inherently a supervised learning task
- Performance of the models is limited by the amount of training data
- Collecting datasets is laborious and expensive
- Unsupervised methods can theoretically achieve better performance by learning from potentially unlimited amount of data

This Talk

Unsupervised Learning of Shape and Pose with Differentiable Point Clouds
E. Insafutdinov and A. Dosovitskiy. NeurIPS 2018

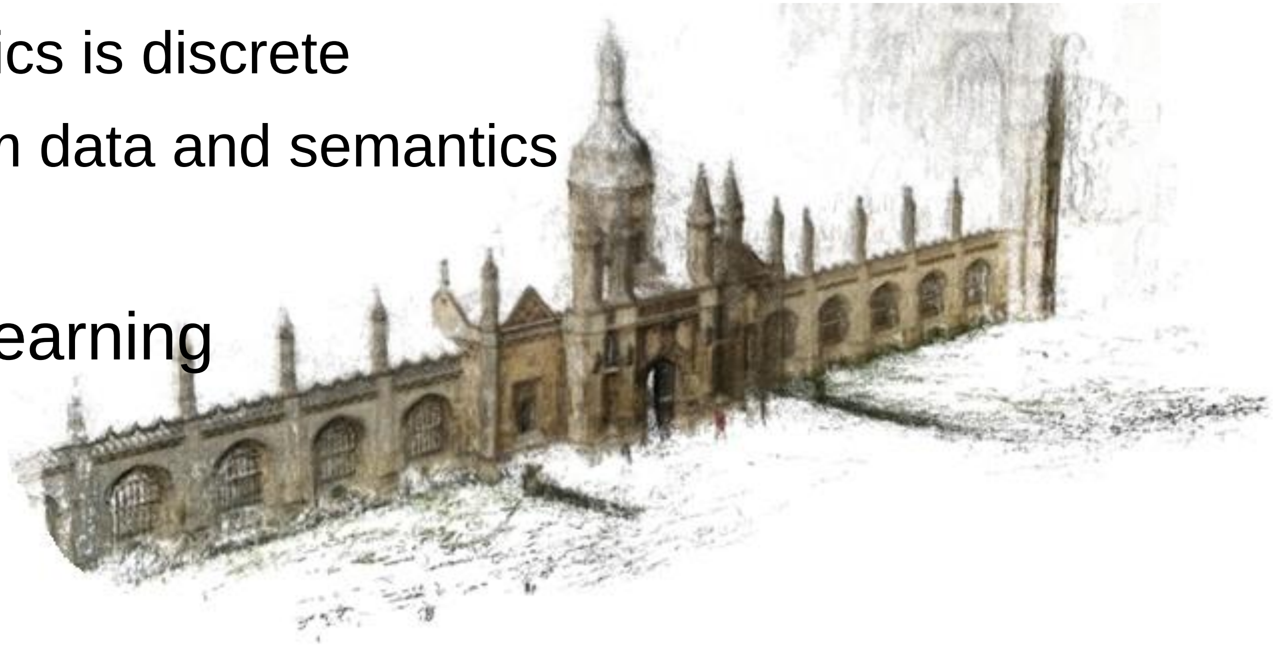
Eldar
Insafutdinov

Alexey
Dosovitskiy

Intel Intelligent Systems Lab, Munich

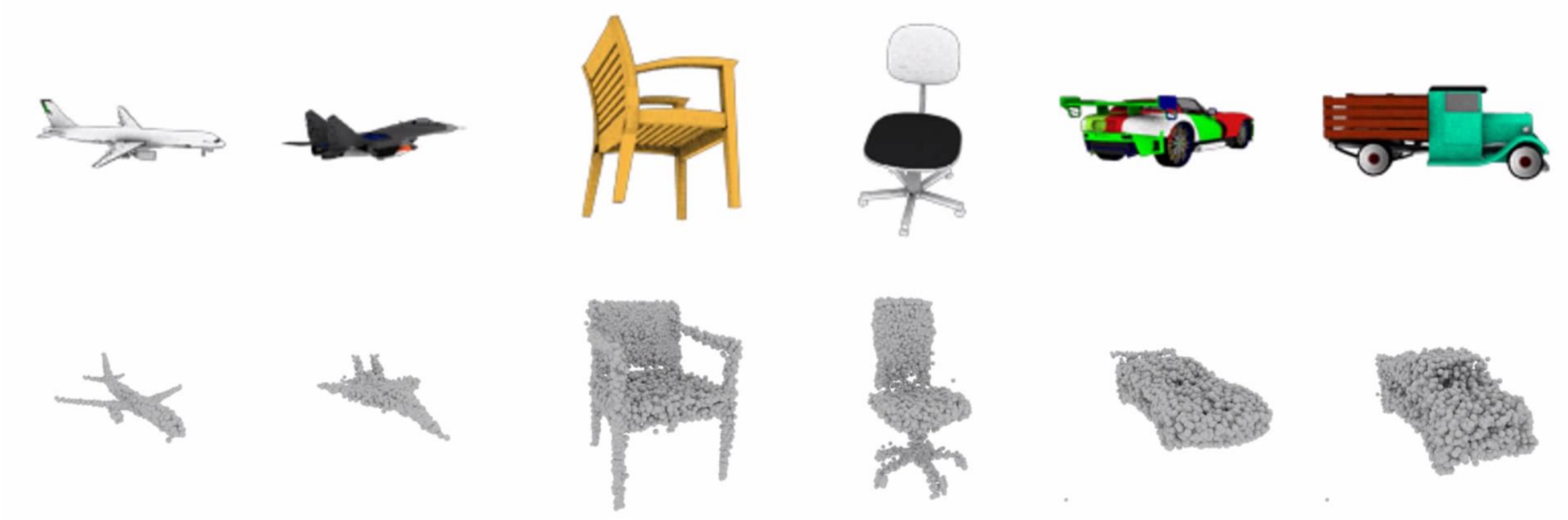
Have We Forgotten about Geometry in Computer Vision?

- A blog post by Alex Kendall
  - https://alexgkendall.com/computer_vision/have_we_forgotten_about_geometry_in_computer_vision/ 2017
- Geometry: depth, shape, pose, motion, optical flow
- Dichotomy between **Semantics** and **Geometry**
  - Geometry is **continuous** and semantics is discrete
  - Geometry is **directly observable** from data and semantics is human-defined
- Geometry facilitates unsupervised learning



3D Shape Reconstruction from Single View

- Given an image of an object from an arbitrary viewpoint:
  - Reconstruct 3D shape
  - Estimate Camera Pose (orientation and translation)
  - Critically: We do not use any 3D supervision during training



Inputs

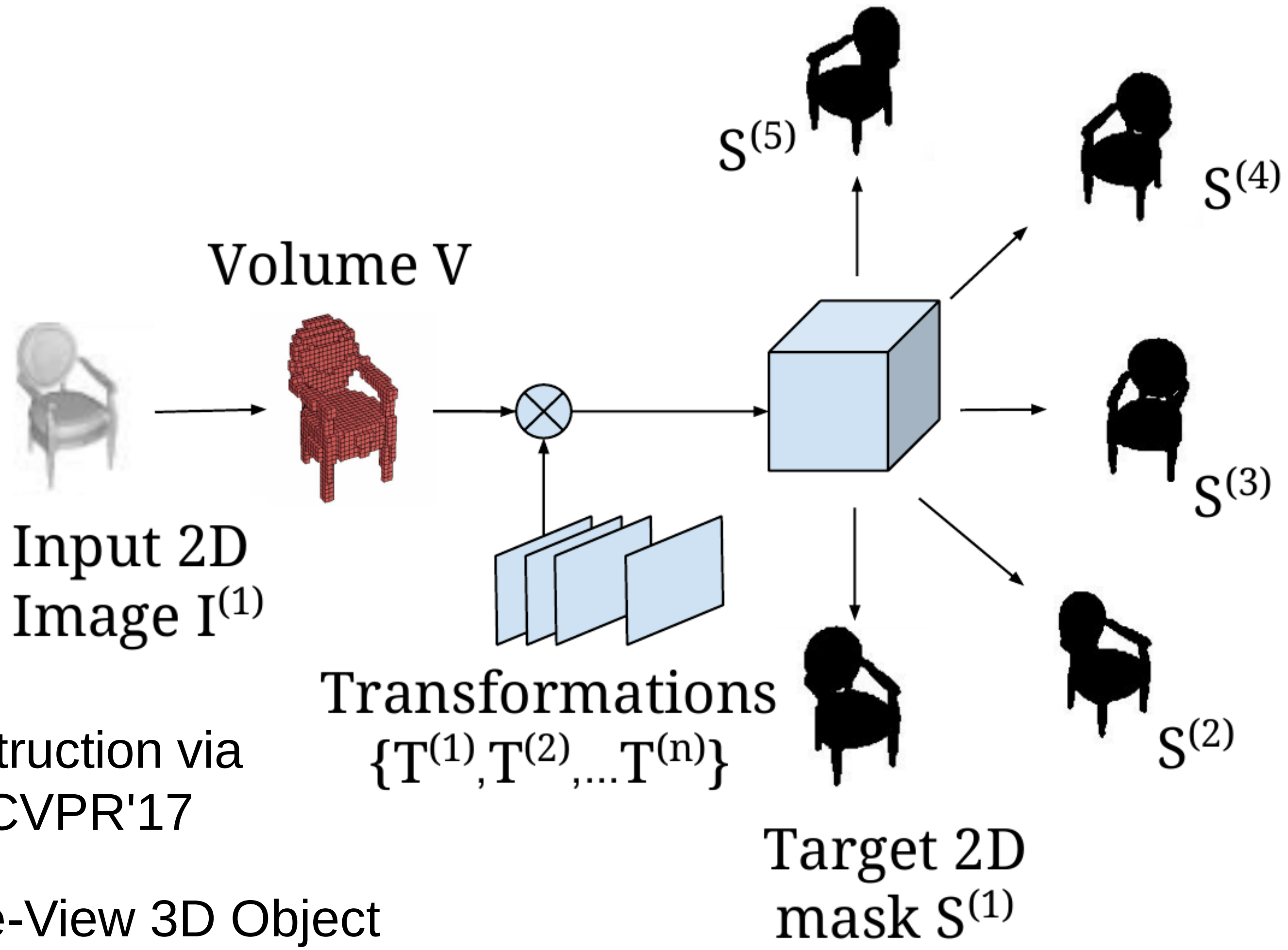
Predictions

Learning of 3D shape without 3D supervision

- Perception of the 3D world is an innate human ability
- We have a very good understanding of the 3D world
- Yet we learn to perceive the 3D without having access to the ground truth
- We mainly observe the world from **2D observations**

Prior Work: Learning of 3D Shape from 2D Supervision

- Learn by re-projection
- Project predicted 3D shape with the known camera view
- Compute loss between the projection and the provided images



Multi-view Supervision for Single-view Reconstruction via Differentiable Ray Consistency. Tulsiani et al. CVPR'17

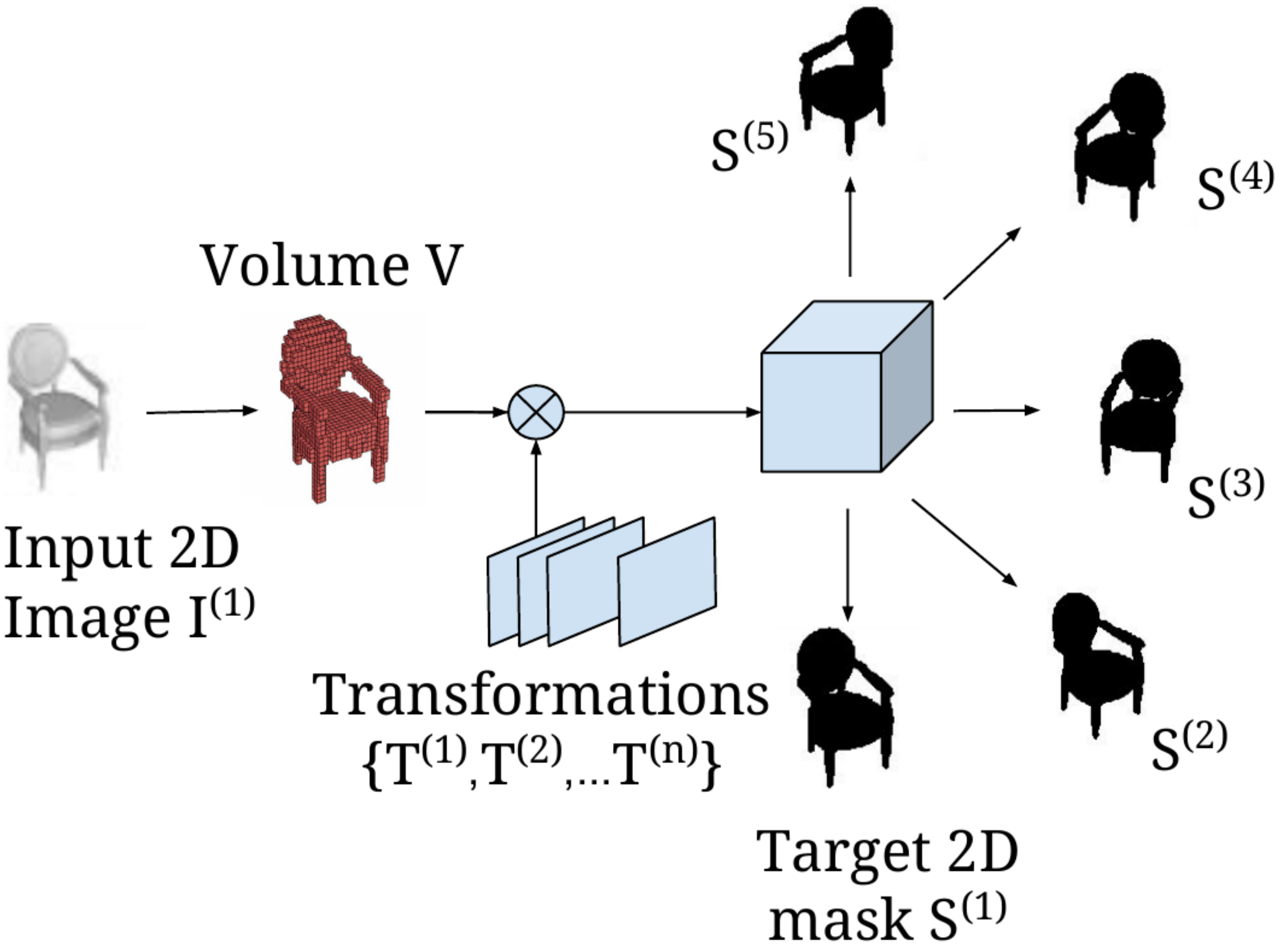
Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction without 3D Supervision. Yan et al. NIPS'16

Prior Work: Learning of 3D Shape from 2D Supervision

Limitations:

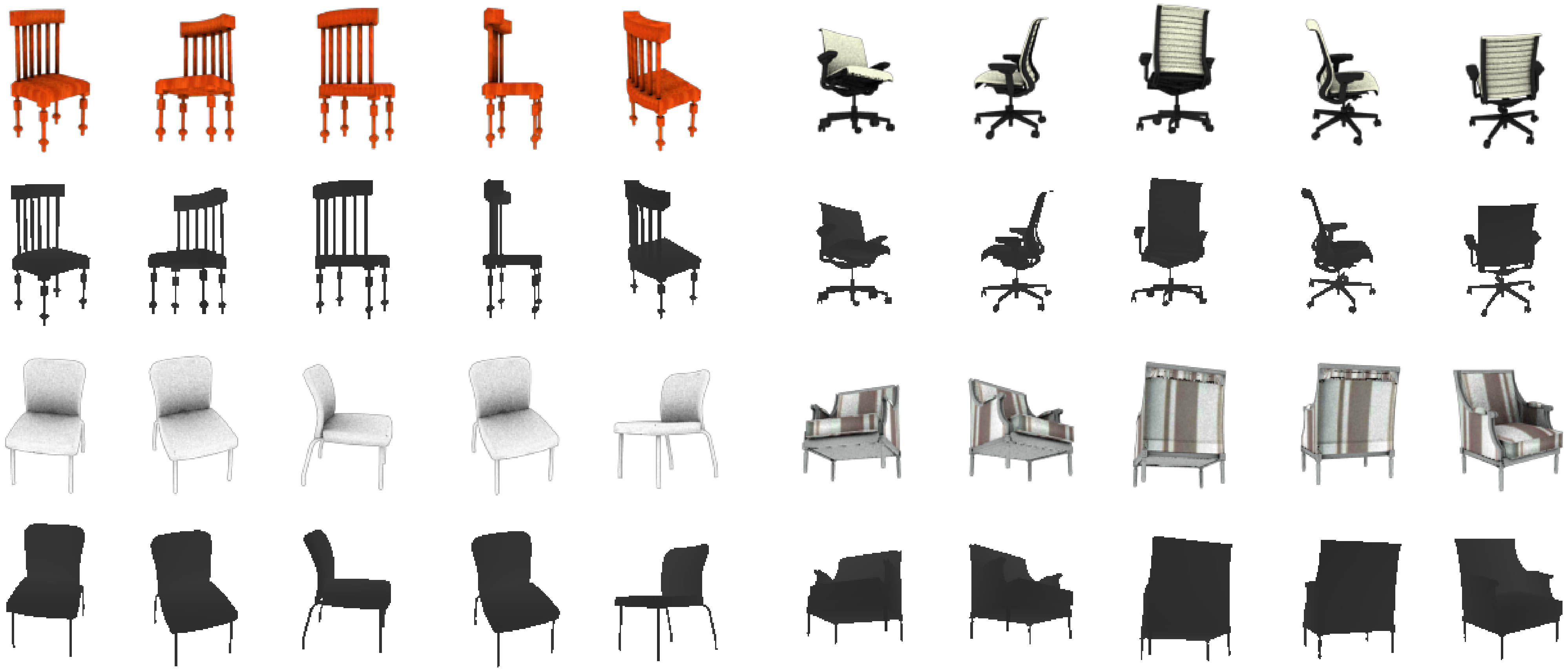
- Relies on the known camera
- Low-res voxel representation
  - Usually 32x32x32

- We address them in our work by:
  - Relaxing the requirement of camera pose supervision
  - Differentiable projection of Point Clouds



Training Data

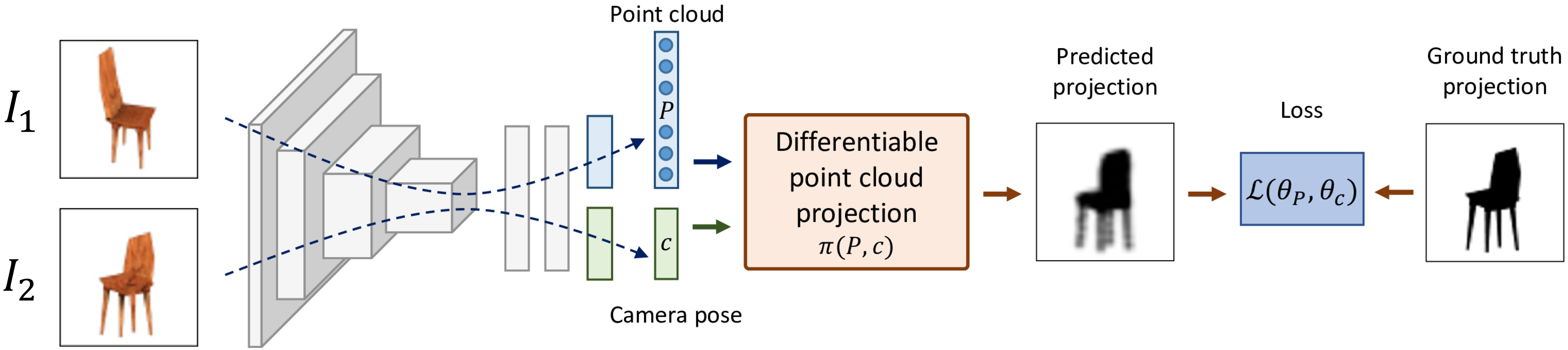
- For each training shape we sample a small number of random views
- Each view is a 2D projection (binary mask, depth or RGB)



Learning of Shape by Re-projection Error

- Project the shape predicted from I_1 using the camera pose from I_2
- Minimize reconstruction error with the GT projection:

$$\mathcal{L}(\theta_P, \theta_c) = \sum_{i=1}^{N} \sum_{j_1, j_2=1}^{m_i} \left\| \hat{\mathbf{p}}^i_{j_1, j_2} - \mathbf{p}^i_{j_2} \right\|^2$$



Choice of 3D Representation: Voxel Occupancy Grids

- 3D Voxel Occupancy Grids are a very popular representation
- They constitute a natural extension of 2D images to 3D space
- Voxels grids scale cubically with resolution (compute and storage)

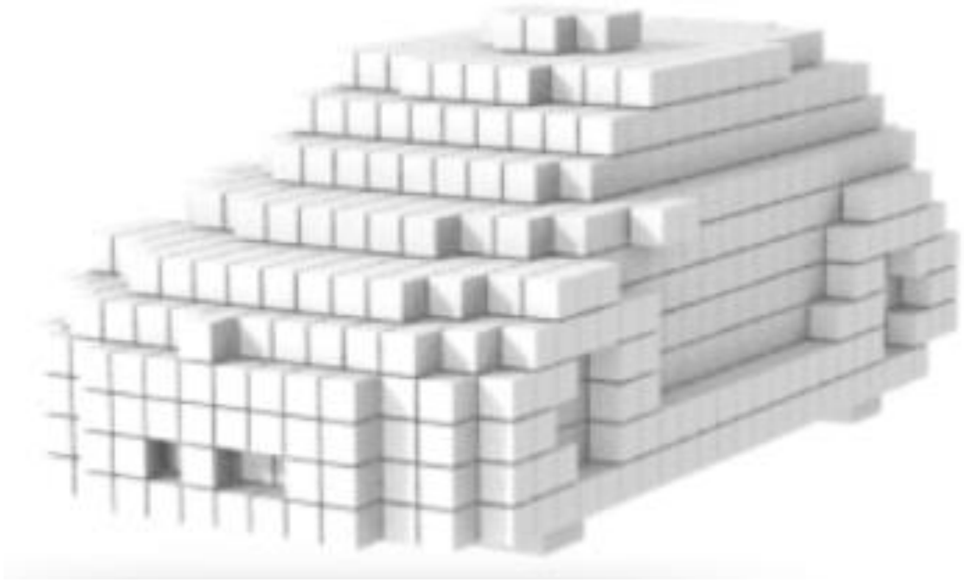
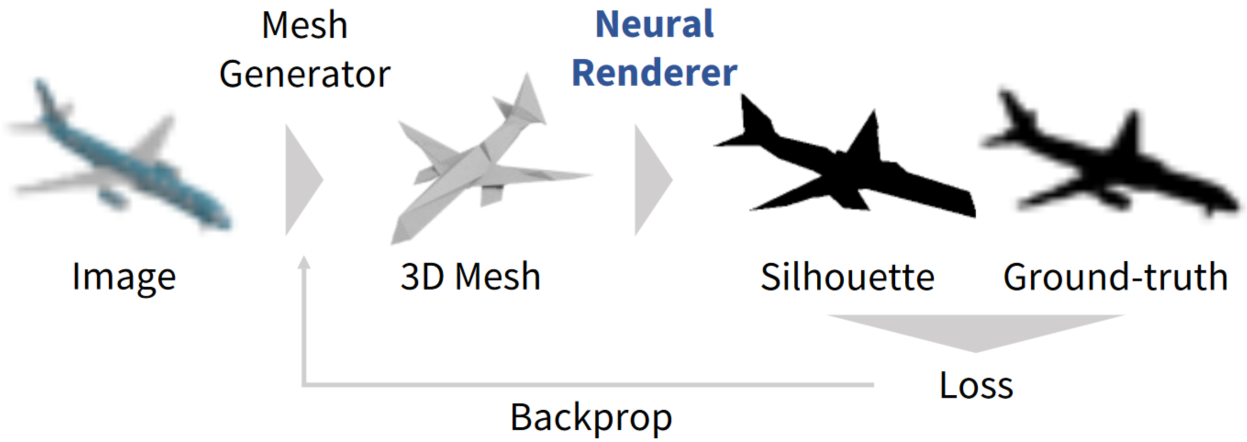


Image Credit: Learning a Multi-View Stereo Machine. A. Kar, C. Häne, J. Malik. NIPS'17.

Choice of 3D Representation: Meshes

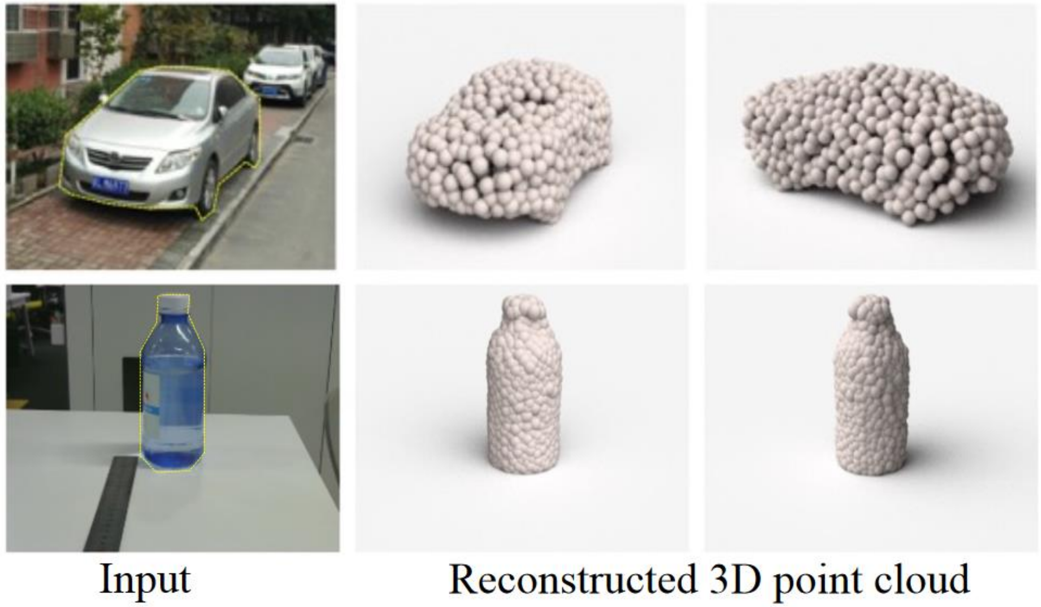
- Meshes are a popular representation with certain advantages:
  - Explicitly available surface
  - Normals are trivial to compute
- However, representing non-trivial topologies is hard



Neural 3D Mesh Renderer. Gato et. al. CVPR'18

Choice of 3D Representation: Point Clouds

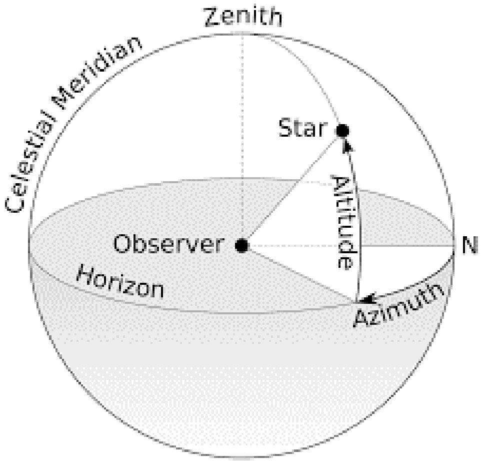
- Resolution-independent representation of shapes
- Easy to learn and apply geometric transformations to
- Unlike meshes point clouds do not encode connectivity
- We choose point clouds in our work



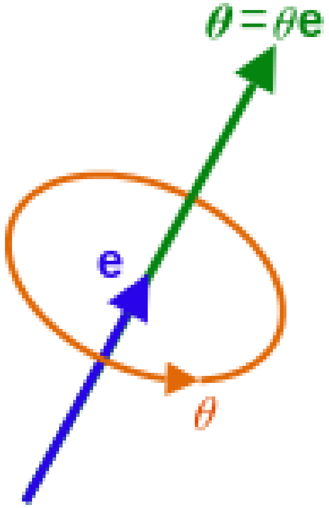
A Point Set Generation Network for 3D Object Reconstruction from a Single Image. Fan et al. CVPR'17.

Parameterisation of Orientation

- Discussed in *Kendall et al. CVPR’17*
  - Rotation Matrix
    - over-parameterised, need to enforce orthonormality
  - Euler Angles
  - Axis-Angle representation (unit vector and angle)



Euler Angles



Axis-Angle

Quaternions and Spatial Rotations

- We choose quaternions to represent rotations
- Quaternions are a number system that extends complex numbers:

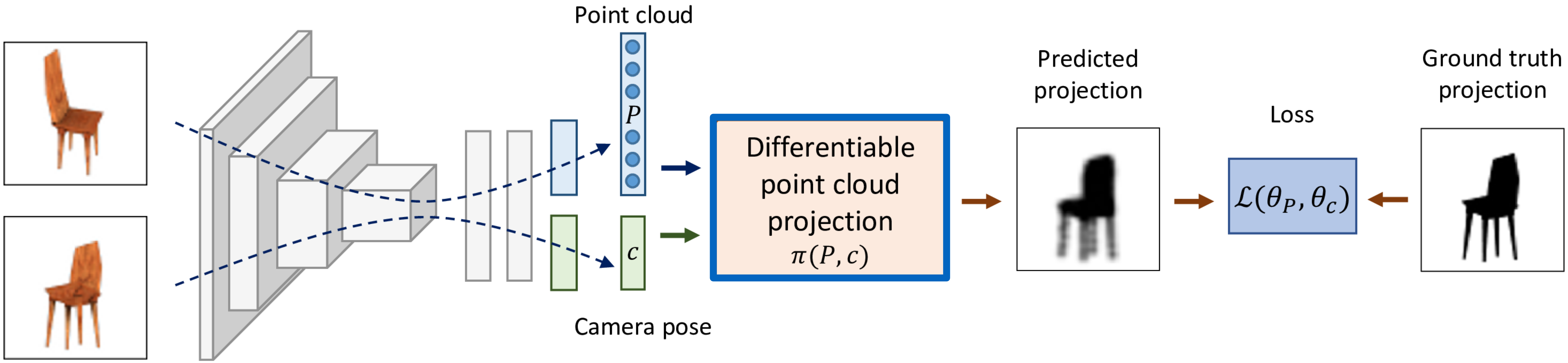
$$q = a + bi + cj + dk$$

- Quaternions can represent spatial rotations
- Rotation of 3D point **p** is defined as

$$\mathbf{p}' = \mathbf{q}\mathbf{p}\mathbf{q}^{-1}$$

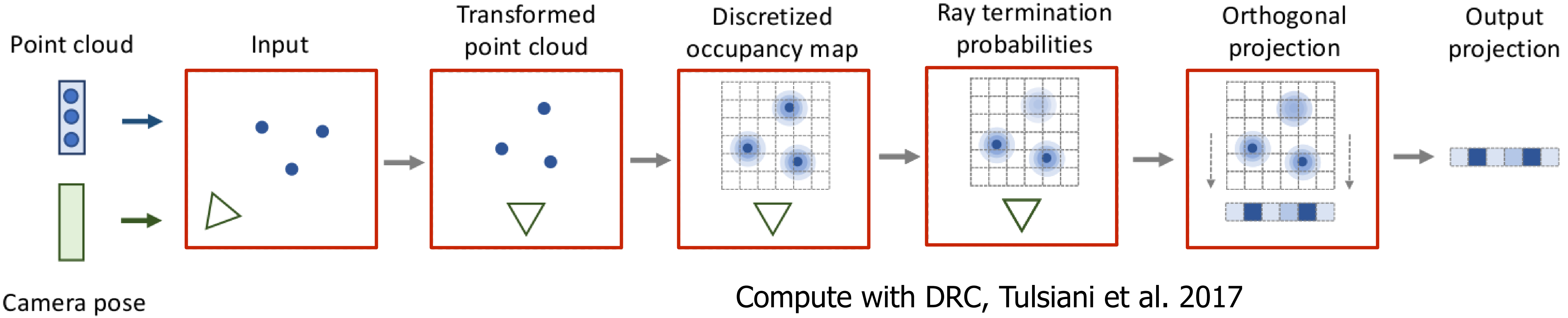
Differentiable Point Cloud Projection

- The key component of our model is the **Differentiable** Point Cloud Renderer $\pi(P, c)$


Point cloud
P
c
Camera pose
Differentiable point cloud projection $\pi(P, c)$
Predicted projection
Loss
$\mathcal{L}(\theta_P, \theta_c)$
Ground truth projection

Differentiable Point Cloud Projection

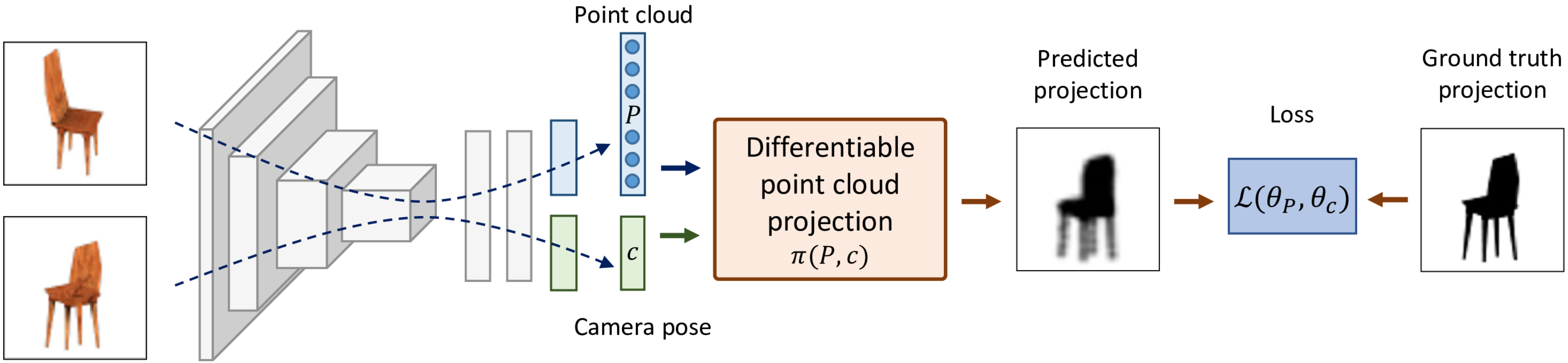
- Given a 3D point $x_i = (x_{i,1}, x_{i,2}, x_{i,3})$ and camera transform *T* (extrinsic and intrinsic parameters):
  1. Apply a perspective camera transformation $x'_i = Tx_i$
  2. Represent x_i as a Gaussian density $f_i(x) = c_i \exp\left(-\frac{1}{2}(x - x'_i)^T \Sigma_i^{-1} (x - x'_i)\right)$
  3. Occupancy function of point cloud: $o(x) = \mathrm{clip}(\sum_{i=1}^{N} f_i(x), [0, 1])$
  4. Discretize $o(x)$ to a grid of resolution $D_1 \times D_2 \times D_3$



Compute with DRC, Tulsiani et al. 2017

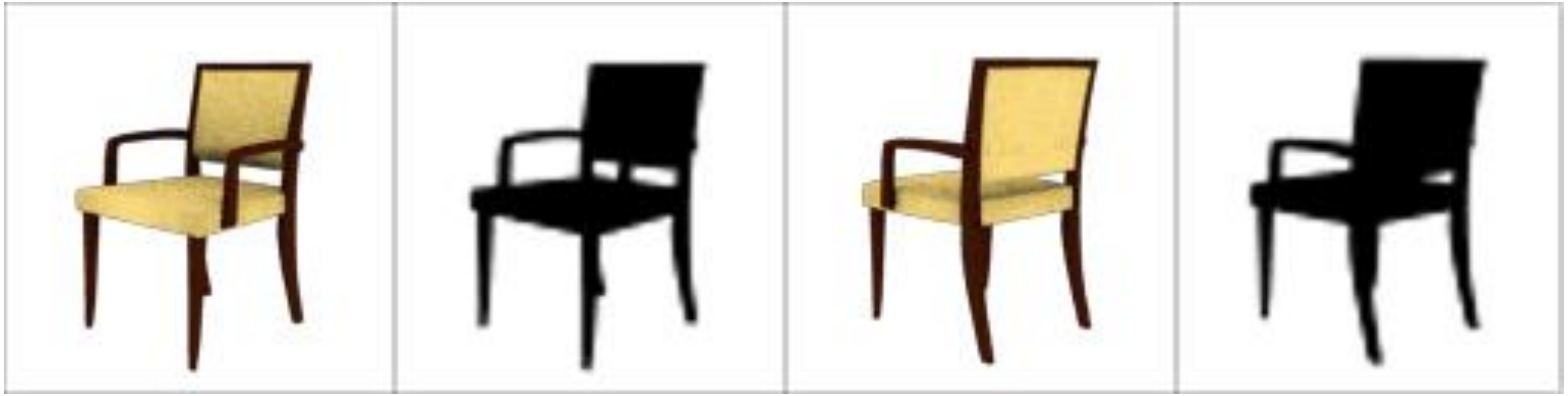
Architecture Overview

- The naive system described so far does not quite work



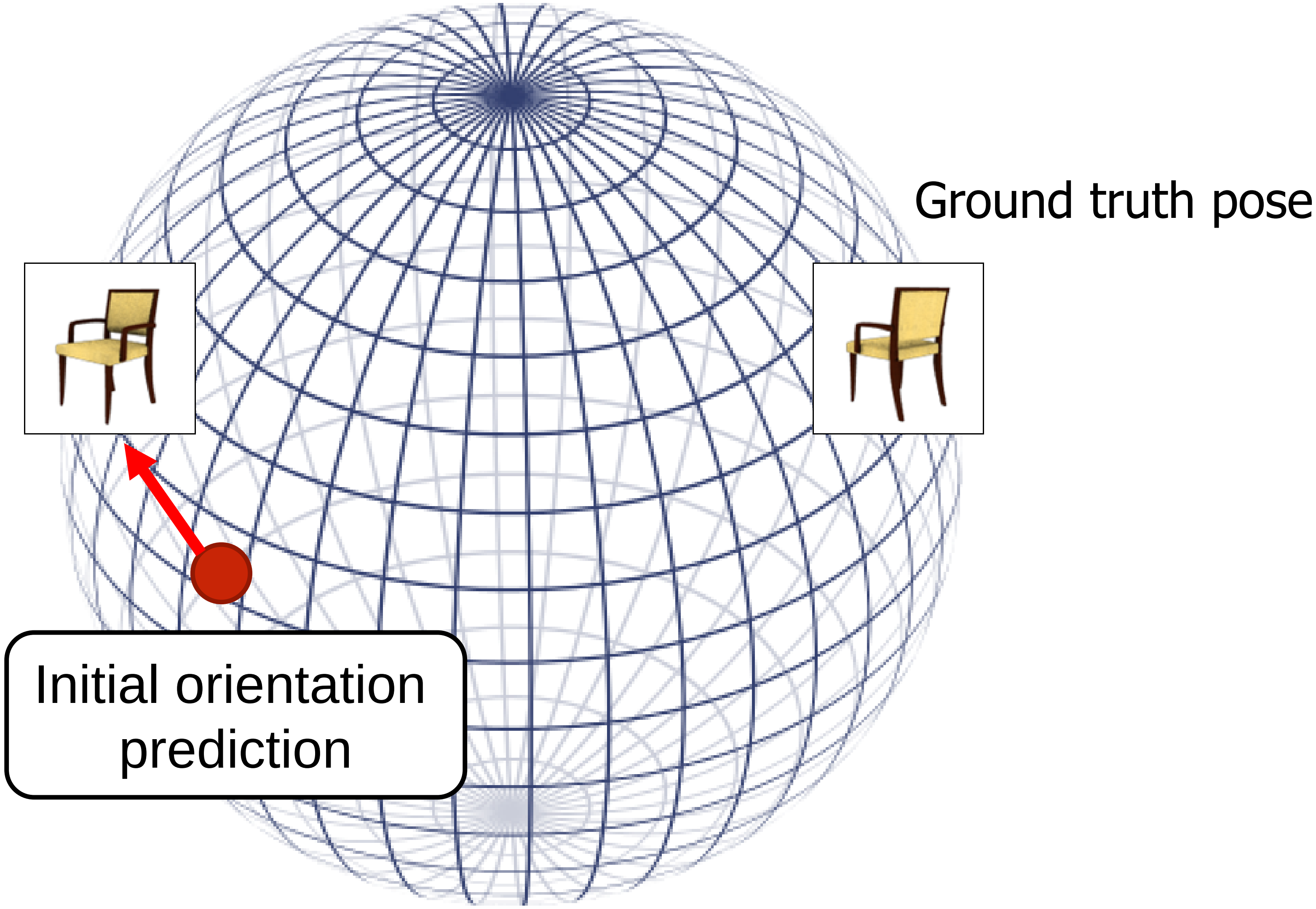
Camera pose ambiguity

- Different views have very similar looking 2D projections



Landscape of the Projection Loss w.r.t. the Camera Orientation

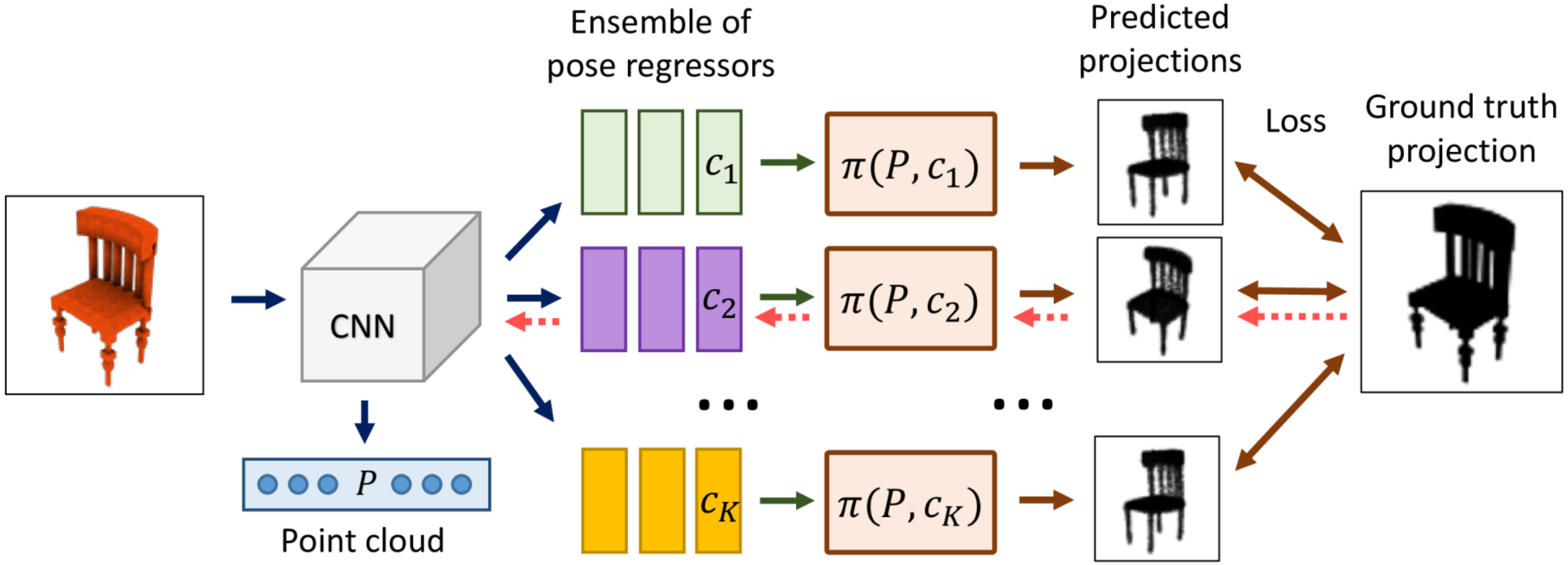
- Red arrow is the loss gradient wrt the camera orientation
- An initial pose estimate can be far from the ground truth
- This will lead training towards the wrong local minima



Ensemble of Pose Predictors

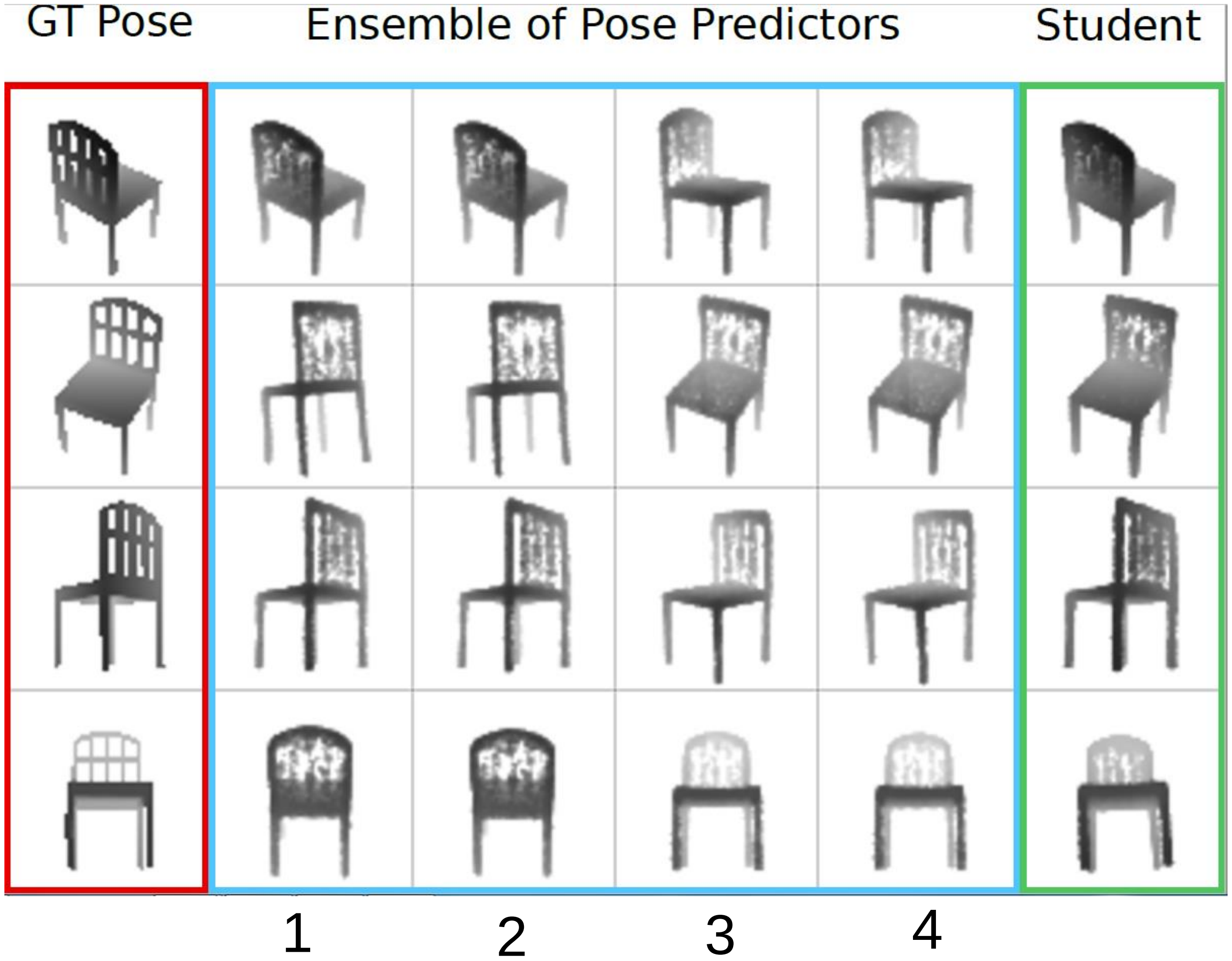
- We need to produce diverse predictions of camera orientation:
  - Train multiple pose regressors instead of one
  - Use the “hindsight” loss to train the ensemble:

$$\mathcal{L}_h(\theta_P, \theta_c^1, \ldots, \theta_c^K) = \min_{k \in [1,K]} \mathcal{L}(\theta_P, \theta_c^k).$$



Ensemble of Pose Predictors

- Each regressor learns to specialize on the subset of poses
- In parallel train a student model
- Supervise the student with best predictions from the ensemble
- Drop the ensemble at test time



Quantitative Evaluation of Point Cloud representation

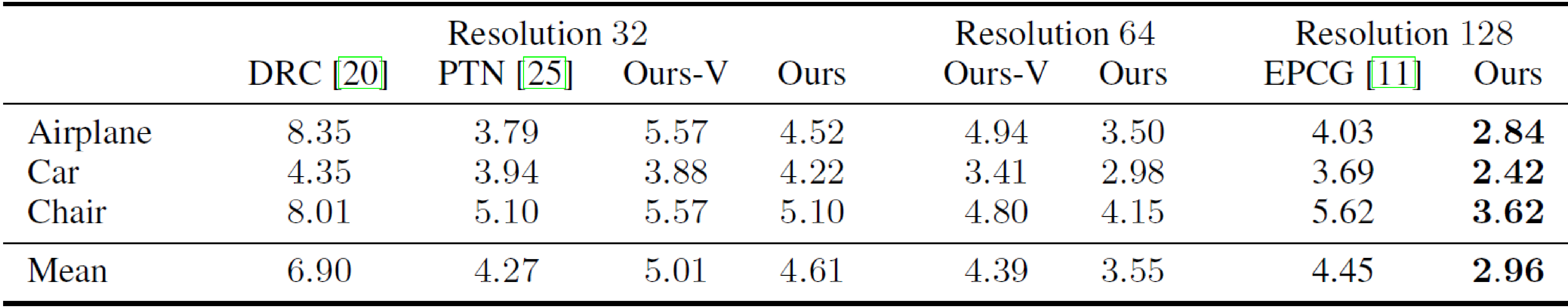
- Evaluate on the ShapeNet dataset:
  - *ShapeNet: An information-rich 3D model repository. Chang et al. 2015*
  - Category specific models on *cars*, *chairs* and *airplans*
- Evaluate of shape reconstruction accuracy with Chamfer Distance metric:

$$d_{CD}(S_1, S_2) = \underbrace{\sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2}_{\text{precision}} + \underbrace{\sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2^2}_{\text{coverage}}$$

A Point Set Generation Network for 3D Object Reconstruction from a Single Image. Fan et al. CVPR'17.

Experiments with Known Camera Pose

- We evaluate the effectiveness of our Differentiable Point Clouds



	Resolution 32				Resolution 64		Resolution 128	
	DRC [20]	PTN [25]	Ours-V	Ours	Ours-V	Ours	EPCG [11]	Ours
Airplane	8.35	3.79	5.57	4.52	4.94	3.50	4.03	**2.84**
Car	4.35	3.94	3.88	4.22	3.41	2.98	3.69	**2.42**
Chair	8.01	5.10	5.57	5.10	4.80	4.15	5.62	**3.62**
Mean	6.90	4.27	5.01	4.61	4.39	3.55	4.45	**2.96**

DRC: **D**ifferentiable **R**ay **C**onsistency. Tulsiani et al. CVPR’17

PTN: **P**erspective **T**ransformer **N**etworks. Yan et al. NIPS’16

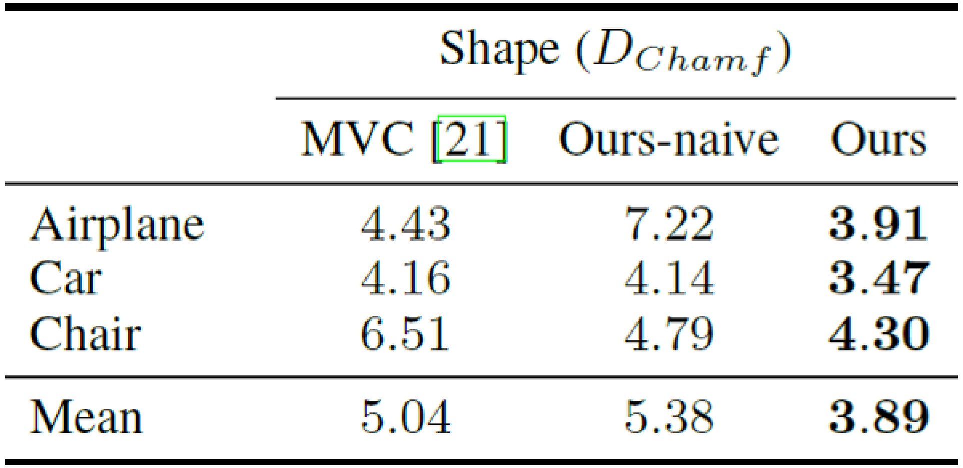
EPCG: Learning **E**fficient **P**oint **C**loud **G**eneration for Dense 3D Object Reconstruction. AAAI 2018.

Experiments with Unknown Camera Pose

Comparison with the concurrent method **MVC**:

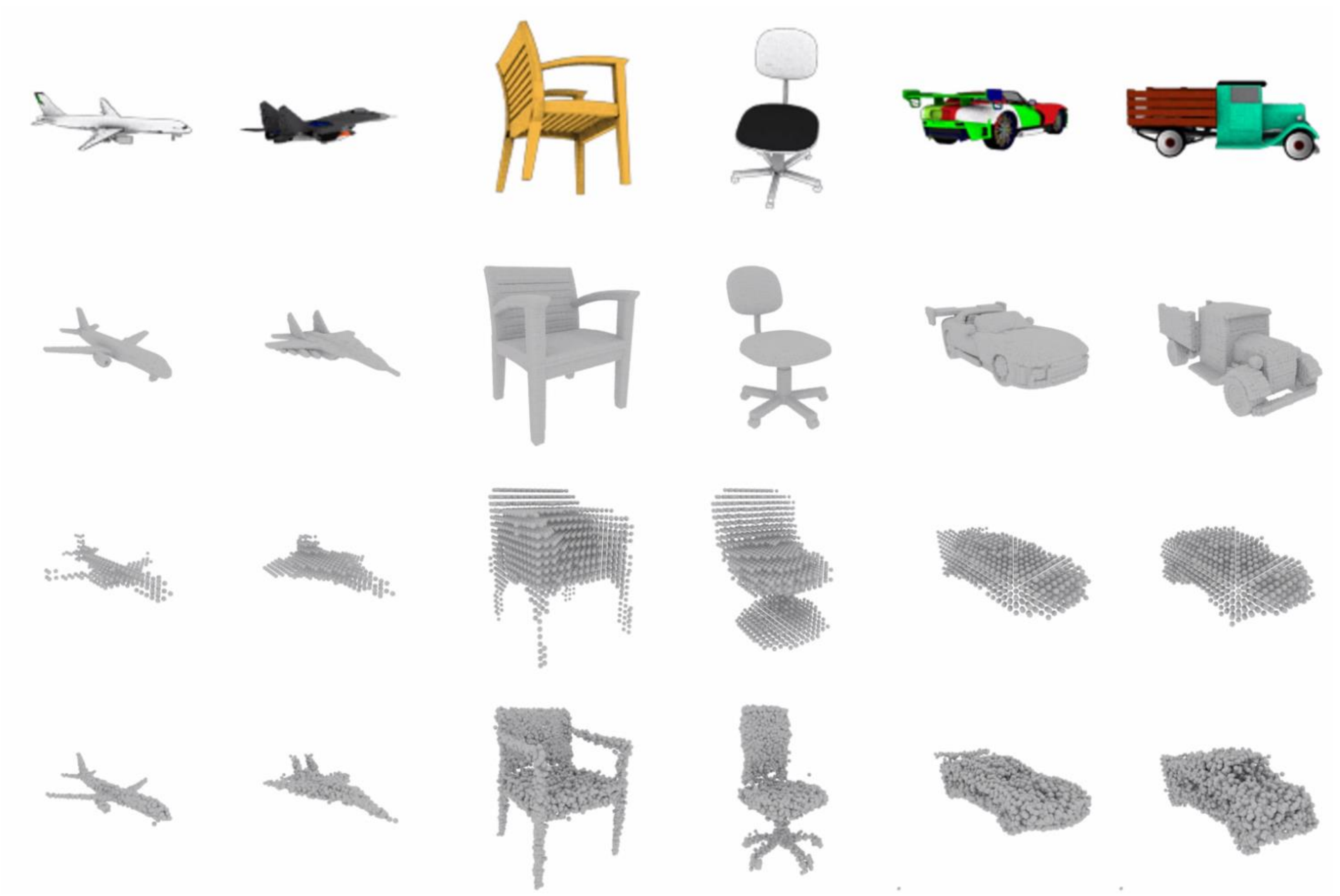
Multi-**V**iew **C**onsistency as Supervisory Signal for Learning Shape and Pose Prediction.

Tulsiani et al. CVPR 2018



	Shape (D_{Chamf})		
	MVC [21]	Ours-naive	Ours
Airplane	4.43	7.22	**3.91**
Car	4.16	4.14	**3.47**
Chair	6.51	4.79	**4.30**
Mean	5.04	5.38	**3.89**

Qualitative results



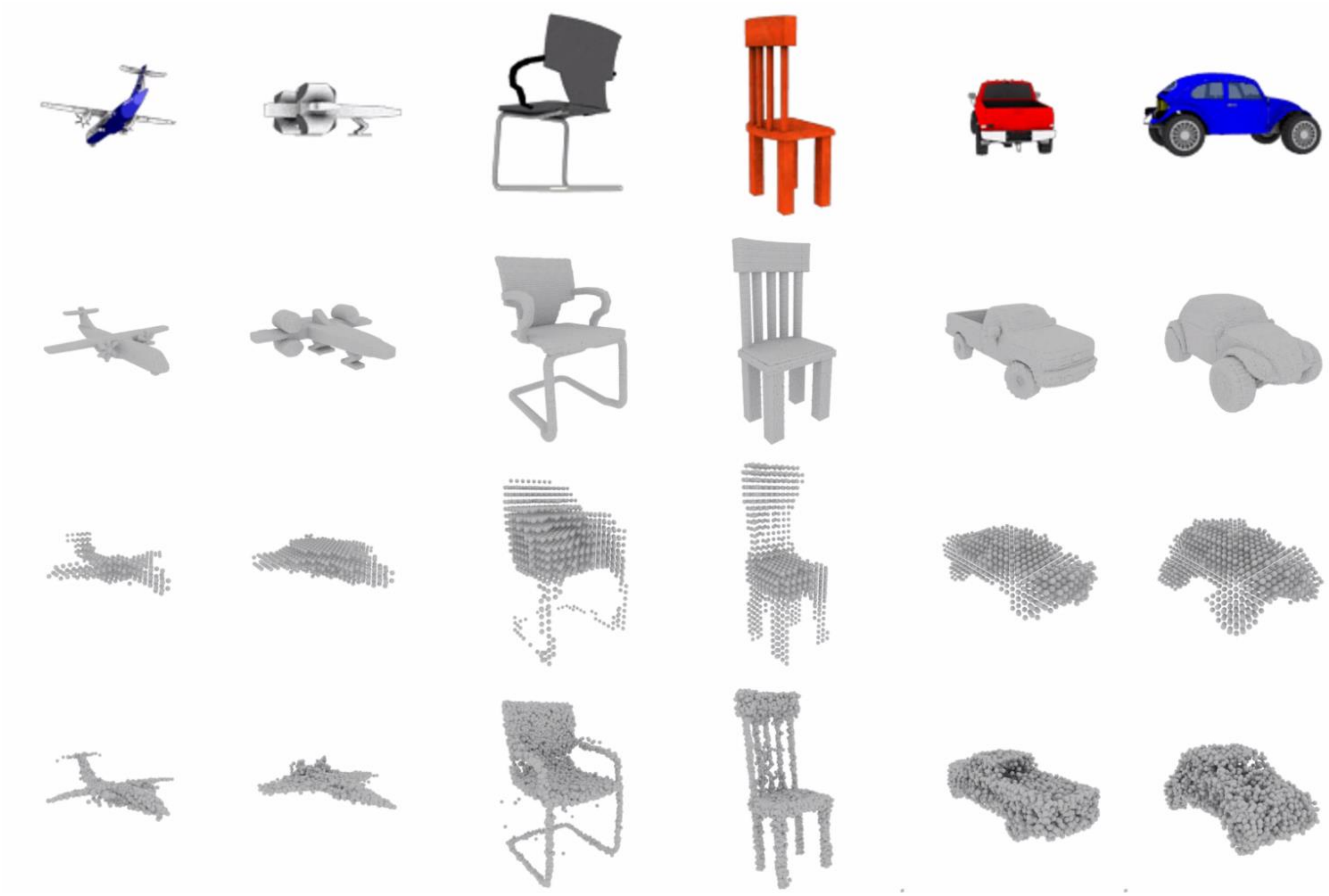
Inputs

Ground Truth

MVC, Tulsiani et al. 2018

Ours

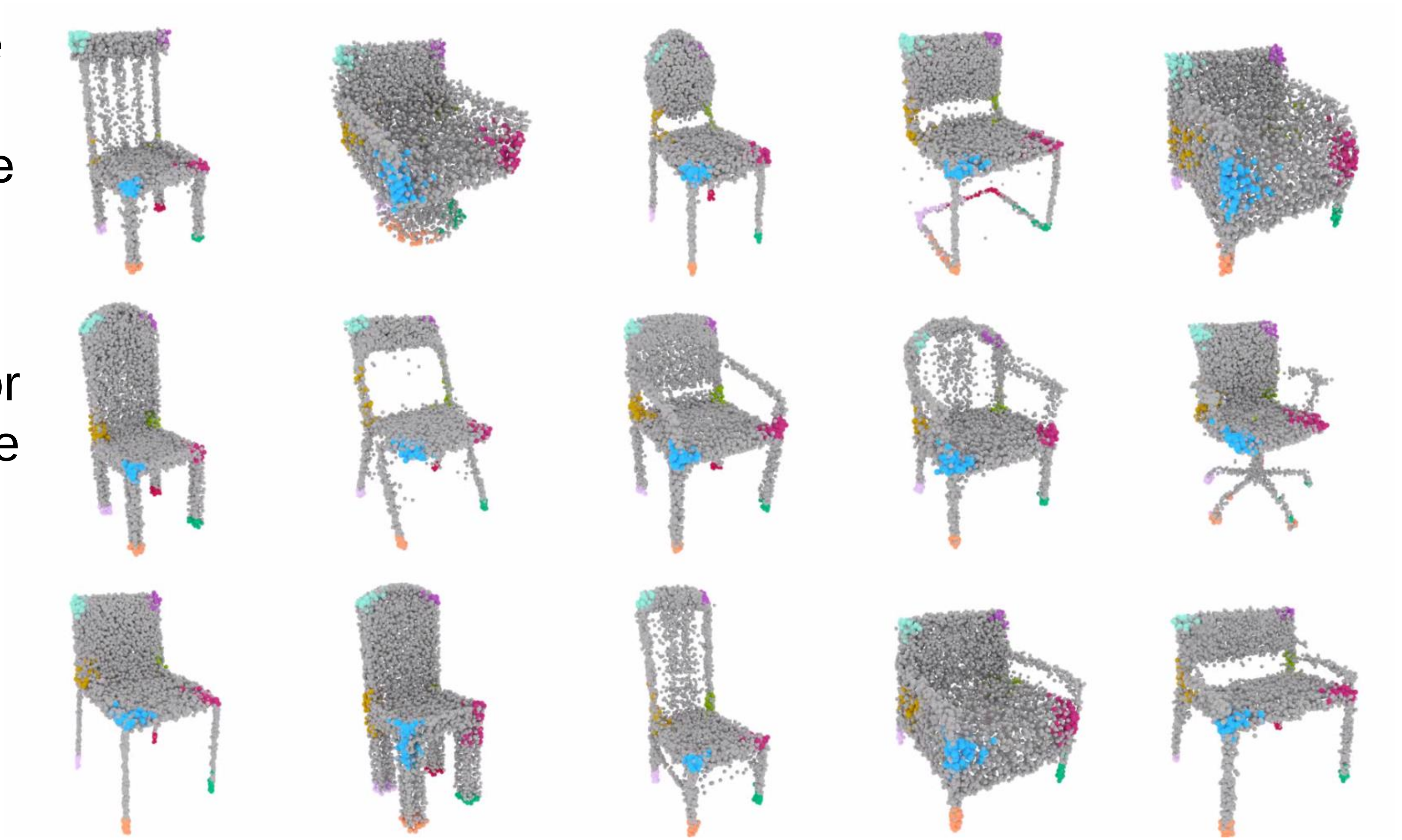
Qualitative results



Unsupervised Discovery of Semantic Correspondences

The network learns to associate the same predicted points with the same “semantic” parts of the objects.

Here the same colour is used for the points predicted by the same output units in the fully connected layer.



Summary

- Proposed differentiable point cloud renderer
  - Allows to learn high detailed shapes
  - Outperforms voxel-based methods
- We can learn object shape even without camera pose supervision
  - Thanks to the ensemble of diverse pose predictors
- Code available:
  - *https://github.com/eldar/differentiable-point-clouds*

Спасибо!