SLAM in the Era of Deep Learning

Ian Reid
School of Computer Science, University of Adelaide
Australian Institute for Machine Learning
Australian Centre for Robotic Vision

Acknowledgements

- Saroj Weerasekera, Huangying Zhan, Kejie Li, Mehdi Hosseinzadeh, Ming Cai
- Ravi Garg, Vijay Kumar, Yasir Latif, Trung Pham, Thanh-Toan Do
- Gustavo Carneiro, Niko Suenderhauf

- **Australian Research Council**
  - Laureate Fellowship
  - Centre of Excellence in Robotic Vision

Combining geometry and semantics

- SLAM solutions rooted in geometry are deeply impressive – and getting better – but:
  - Say nothing about *what* is present
  - Usually do not adequately to leverage prior knowledge
- What do we want from “ideal” SLAM?
  - Real-time scene understanding
  - Dense, accurate, large-scale
  - Semantically rich, object-based, enforces inter-object constraints, understands and uses physical constraints (and even physics)
  - etc

Deep learning for SLAM

This talk:

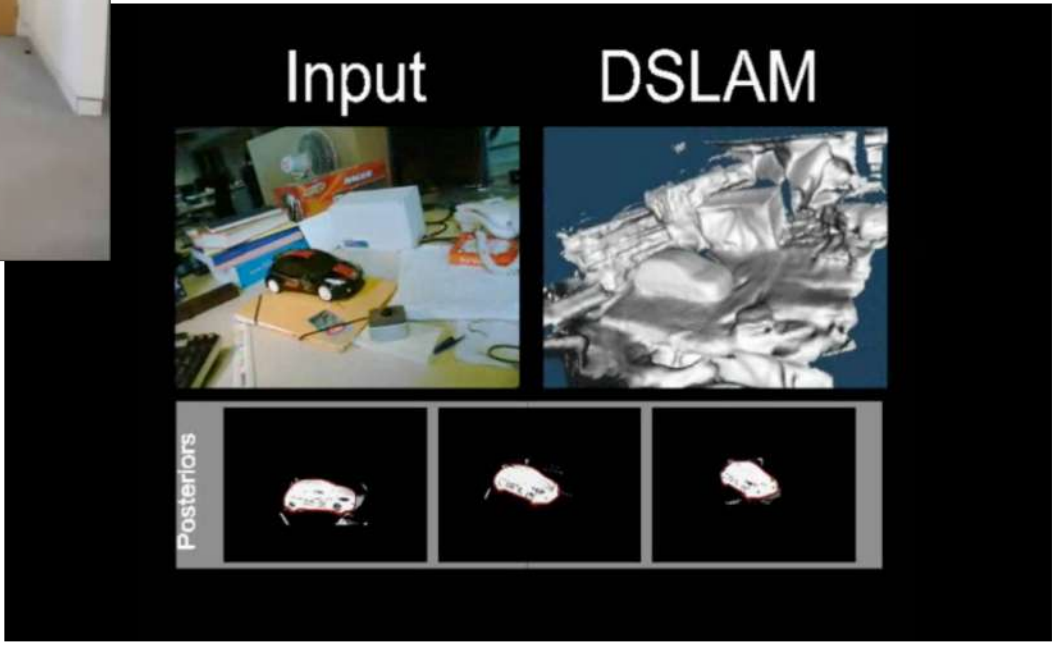
1. Using deep networks for better dense SLAM

1. Reconstructing objects as well as scenes for object-based SLAM

Before deep learning…

Flint, Mei and Reid, CVPR 2010, ICCV 2011

Dame, Prisacariu, Kahler, Segal and Reid, CVPR 2013



1. Improving dense SLAM with deep learning

Dense per pixel tasks

- *Nekrasov et al, Lightweight RefineNet, BMVC 2018*
- *Nekrasov et al, Real-Time Joint Semantic Segmentation and Depth Estimation, arXiv 2018*

Dense SLAM with deep learning

- Usual formulation of dense SLAM
  - Photometric (pixel intensity) cost

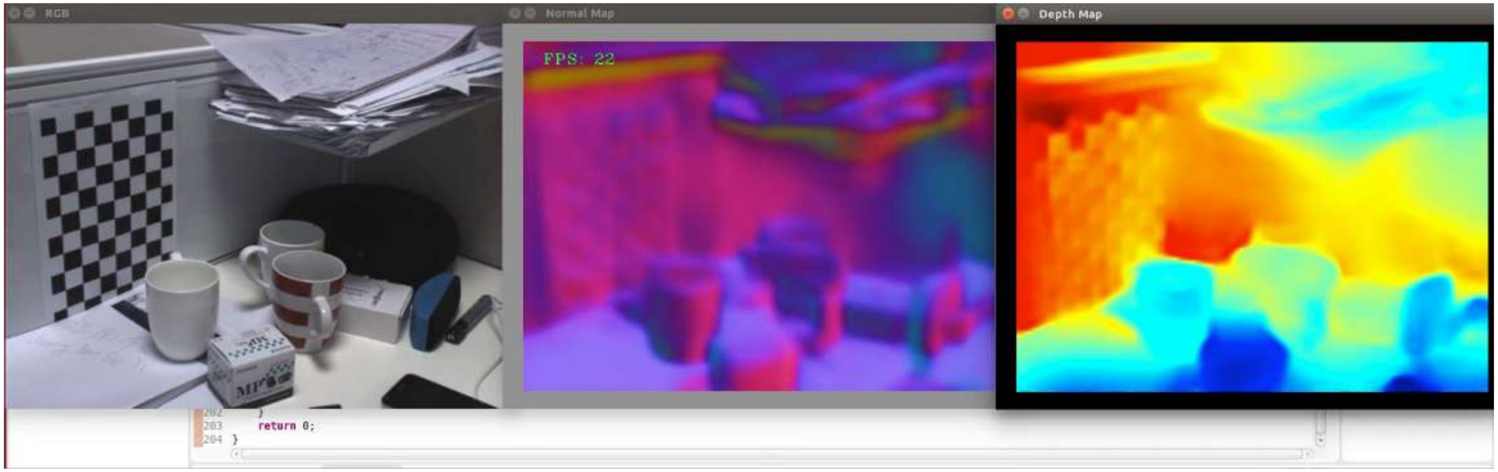
$$\mathbf{E}_{\mathbf{pixel}}(\mathbf{u}_p, \frac{1}{\mathbf{d}}) = \frac{1}{|\mathcal{I}(r)|} \sum_{m \in \mathcal{I}(r)} \left\| \rho(\mathbf{I}_m, \mathbf{u}_p, \frac{1}{d_p}) \right\|_1$$

$$\rho(\mathbf{I}_m, \mathbf{u}_p, \frac{1}{d_p}) = \mathbf{I}_r(\mathbf{u}_p) - \mathbf{I}_m(\pi(\mathrm{KT}_{mr}, \pi^{-1}(\mathbf{u}_p, d_p)))$$

  - Photometric cost no good in regions of uniform brightness
  - Global prior term to regularise (e.g. smoothness)

Dense SLAM with deep learning

1. Can learn scene structure from lots of examples
   - Use this as a prior instead of global smoothness



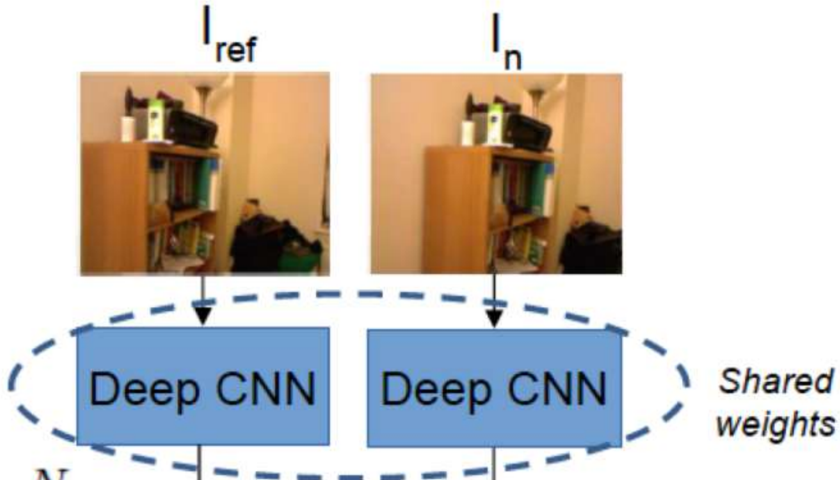
   - New prior

$$\mathbf{E}_{\mathbf{normal}}(\mathbf{d}) = \sum_{(p,q)\in\mathcal{P}} g_p \left\|\hat{\mathbf{n}}_p \cdot \ (d_q\tilde{\mathbf{x}}_q - d_p\tilde{\mathbf{x}}_p)\right\|_\epsilon$$

Weerasekera *et al*, ICRA 2017

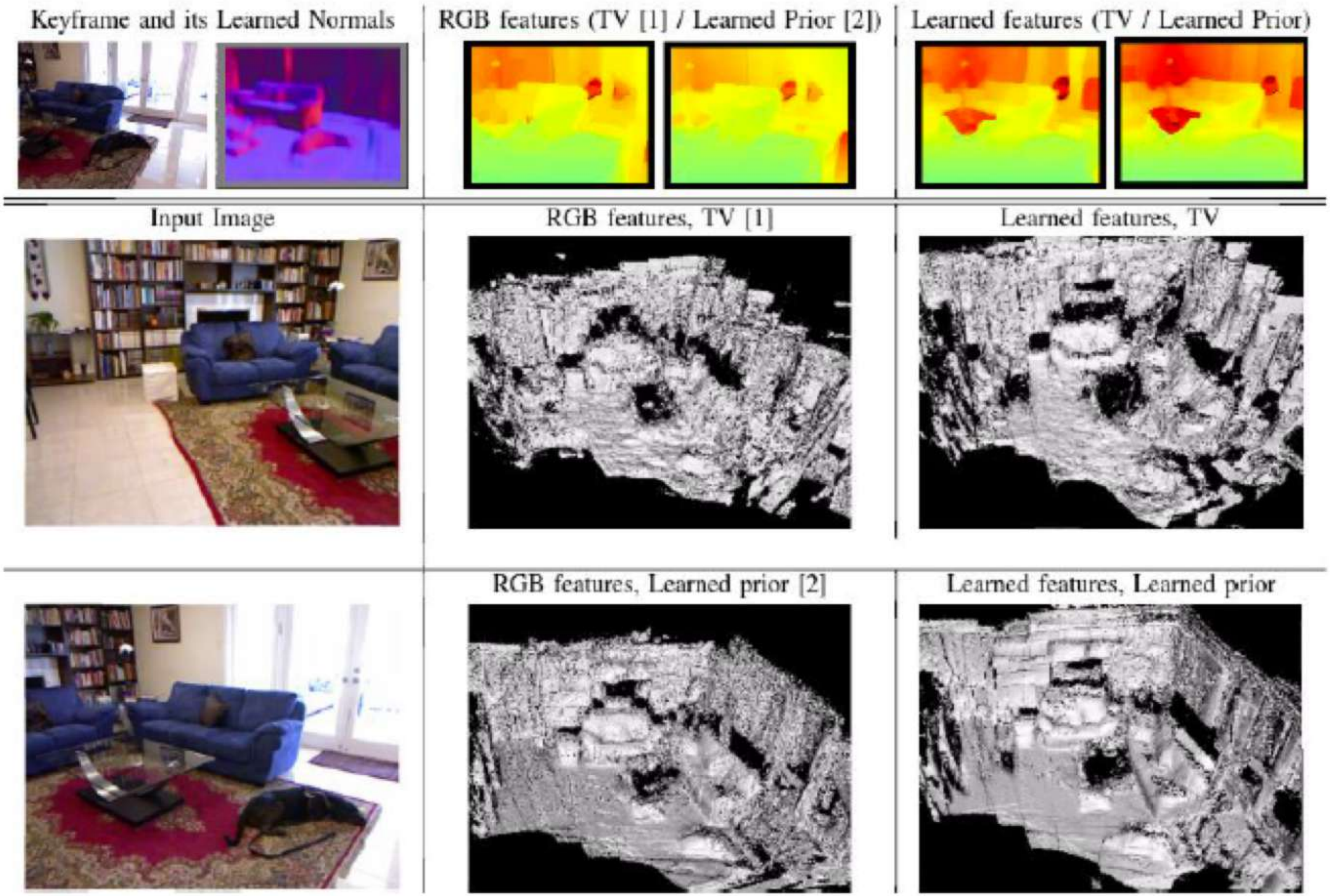
Dense SLAM with deep learning

2. Individual pixel brightnesses are not informative
   - Use deep feature representation *per pixel*



$$E_\phi(\rho_p) = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathbf{f}_{ref}(\mathbf{u}_p) - \mathbf{f}_n(\pi(T_{nr}\pi^{-1}(\mathbf{u}_p, \rho_p)))\right\|_1$$

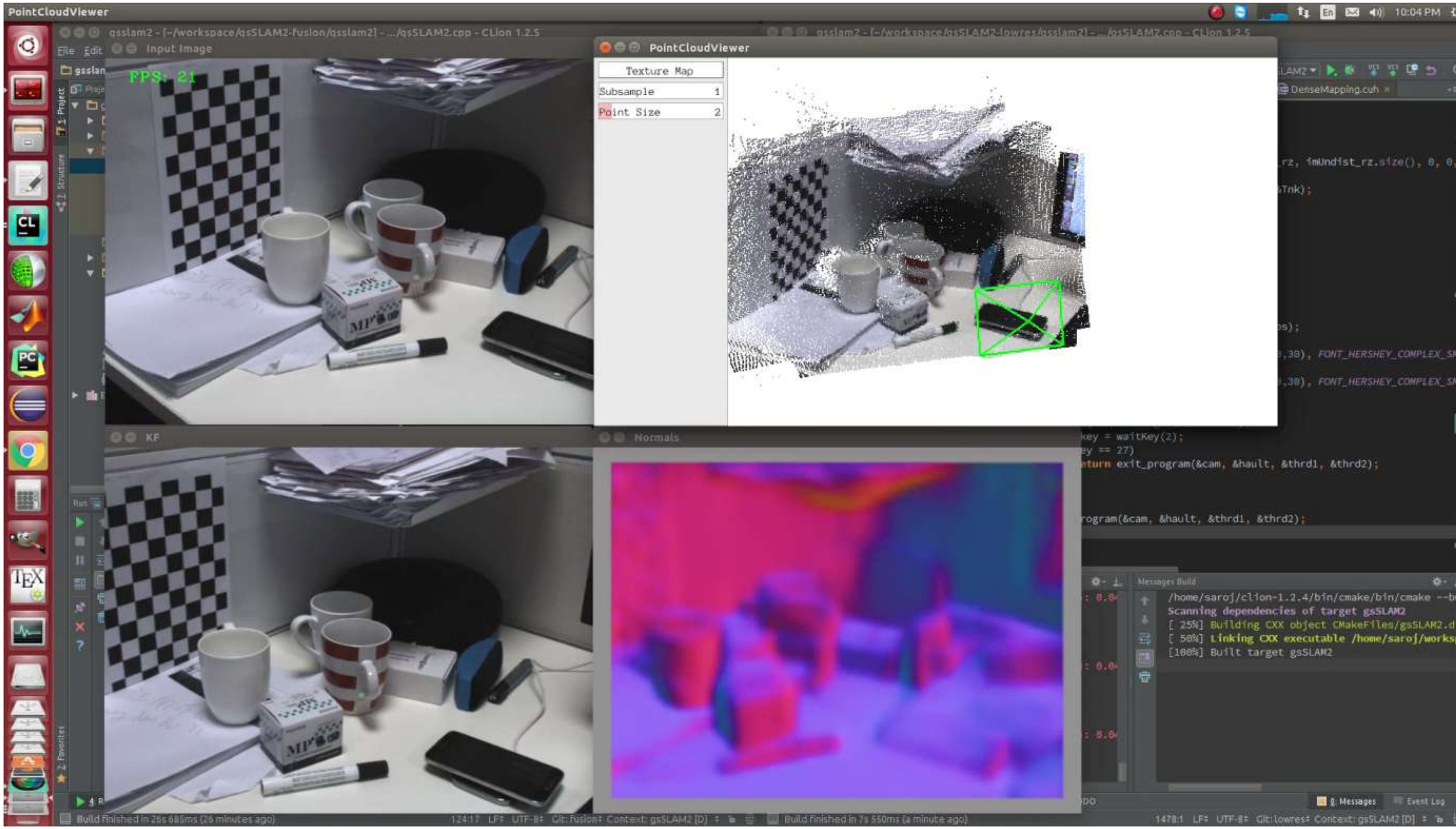
Weerasekera *et al*, ACCV 2018



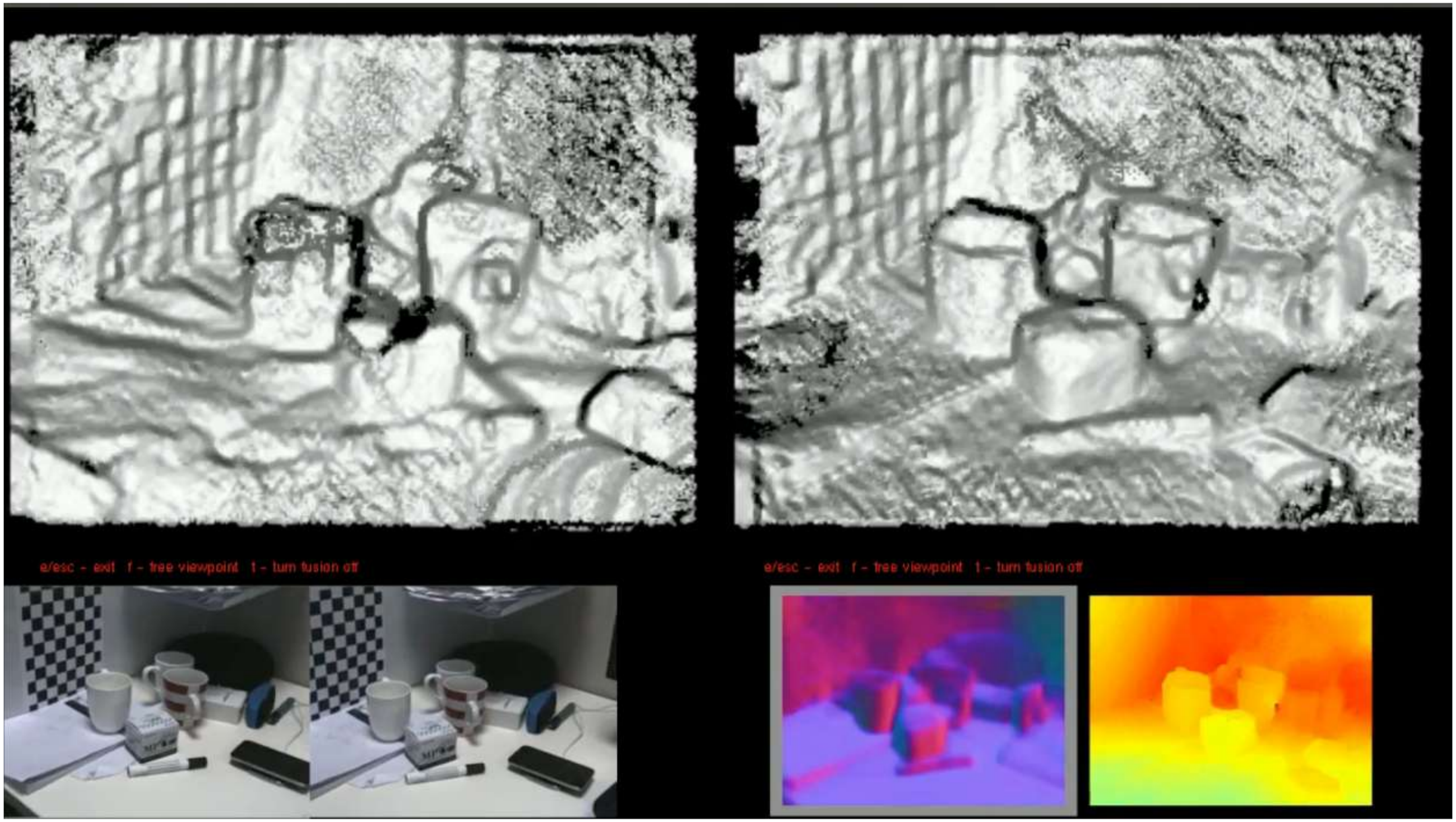
[1] Newcombe et al., ICCV 2011

[2] Weerasekera et al., ICRA 2017

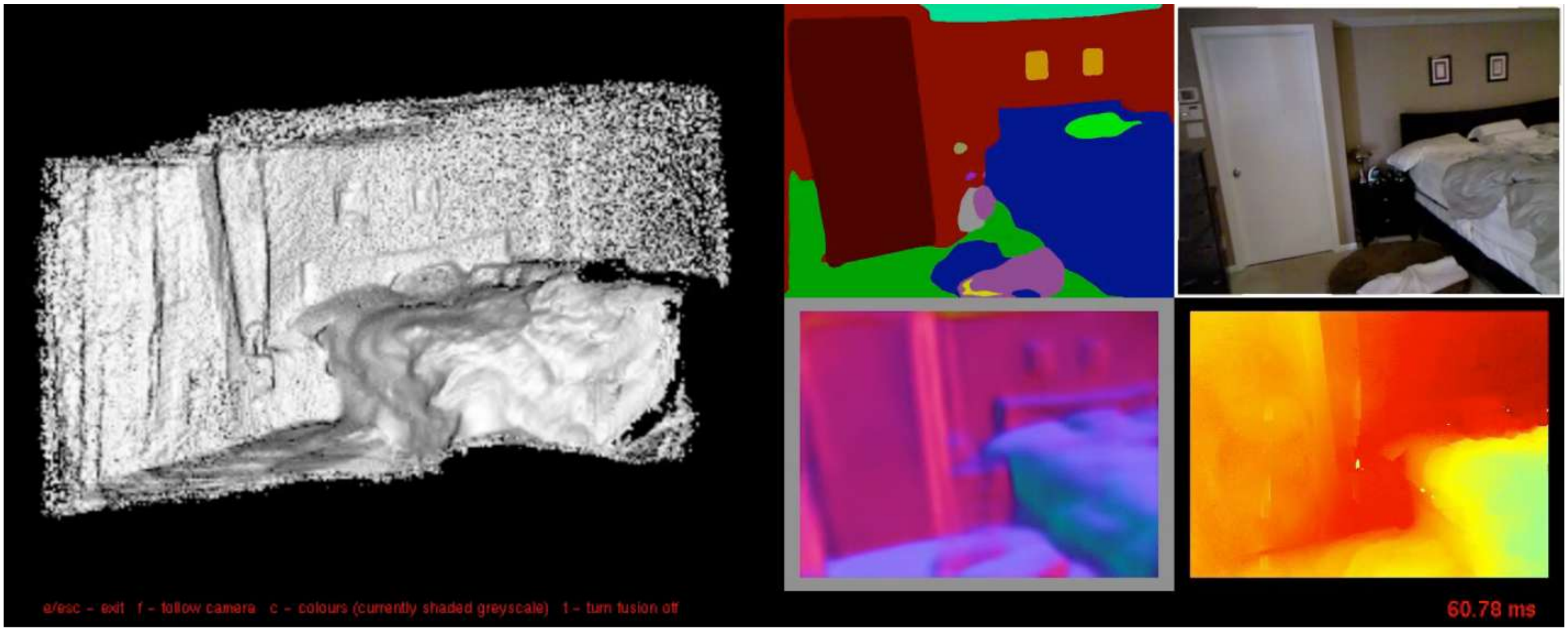
Example


PointCloudViewer
Input Image
FPS: 21
Texture Map
Subsample 1
Point Size 2
KF
Normals

Comparison with smoothness



Adding semantic segmentation

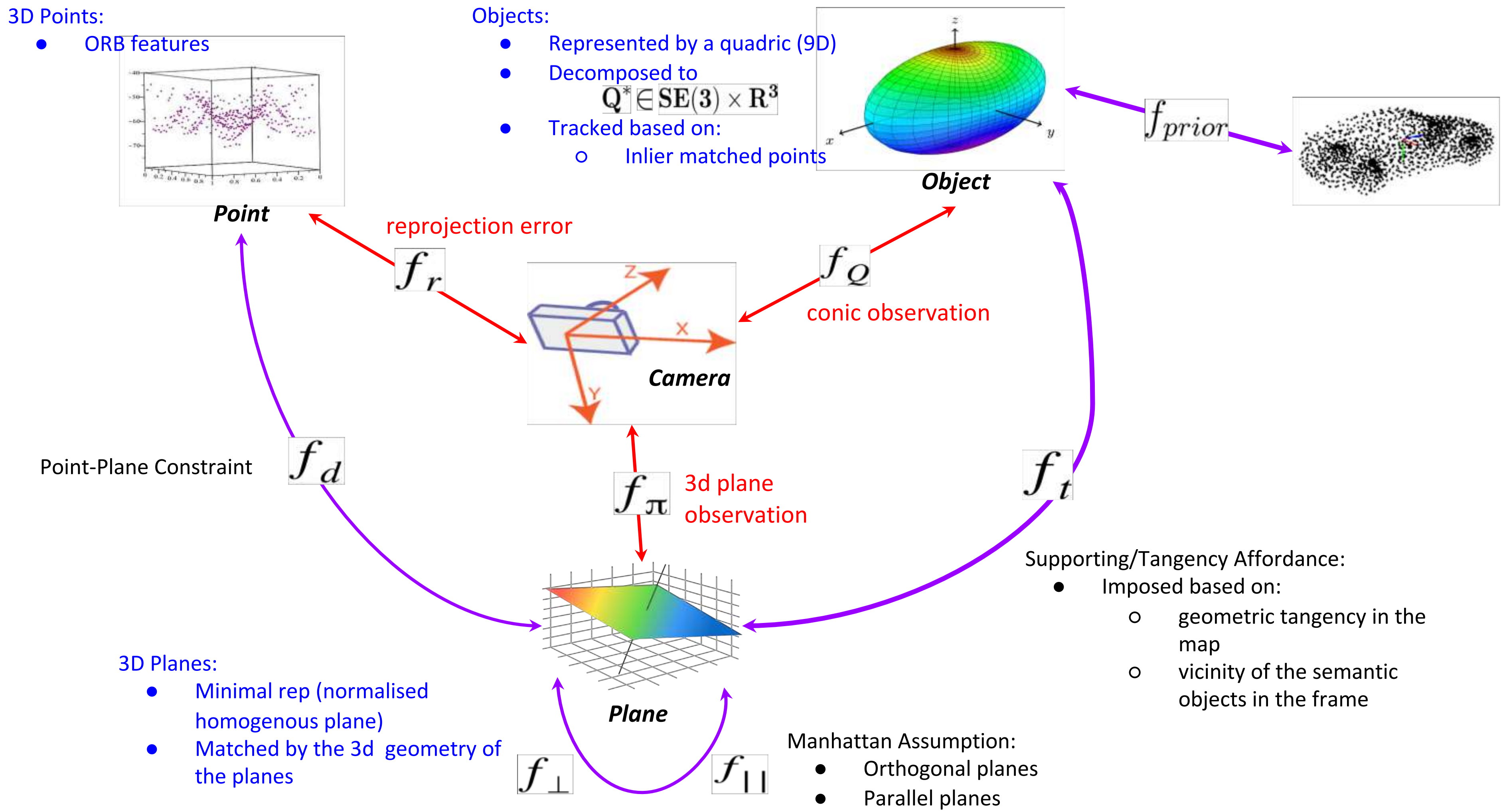


2. Towards Semantic Object SLAM

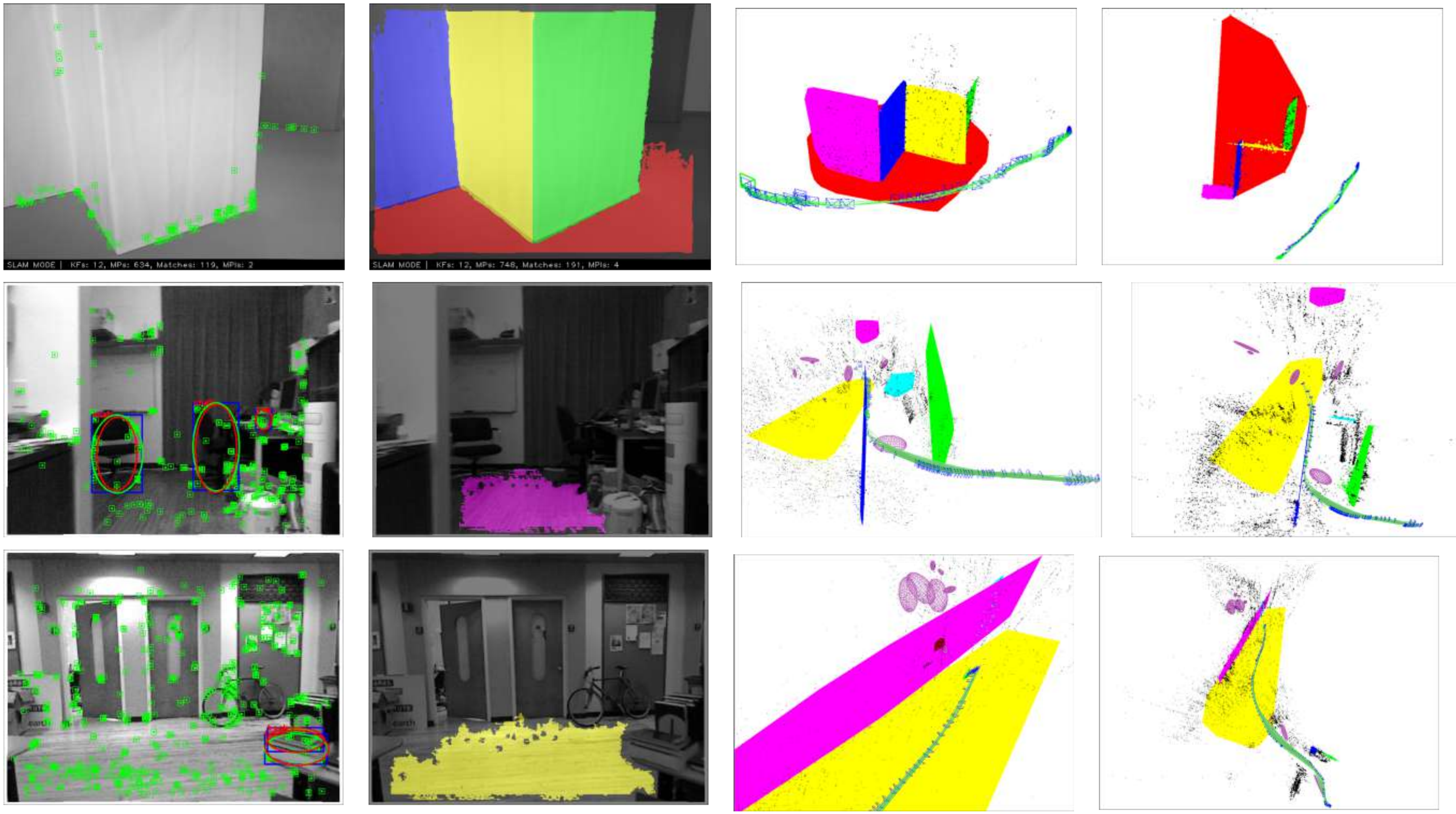
Towards Semantic Object SLAM

- Building meaningful map representation while localising the camera
  - Points
    - sparse, easier to detect, and robust for tracking
  - Planes
    - capture the large-scale structure of a general scene (indoors)
    - appropriate representation for feature-deprived regions
    - more difficult to match than points
  - Objects
    - general unseen objects
    - the most difficult to represent and track
    - start with generic (quadrics) then move to single-view learned 3D

Landmarks and Constraints



Results



fr3/str notex far

nyu office_1

nyu office 1b

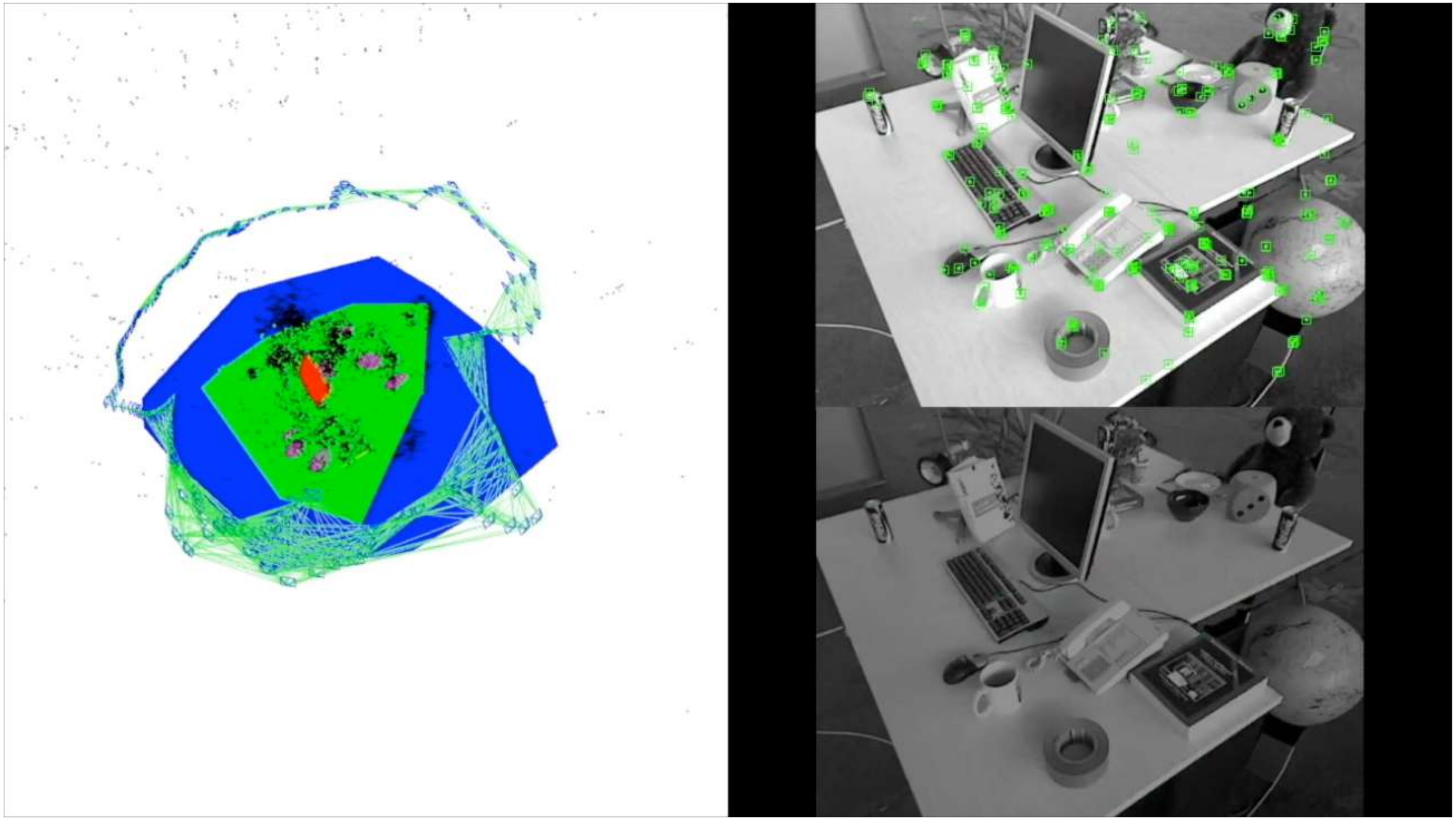
ORB-Features Detected Objects

Segmented Planes

Reconstructed Map (Side)

Reconstructed Map (Top)

Results



Quantitative Results

- Ablation study against point-based ORB-SLAM2

Table 1: Comparison against RGB-D ORB-SLAM2. PP, PP+M, PQ, and PPQ+MS mean points-planes only, points-planes+Manhattan constraint, points-quadrics only, and all of the landmarks with Manhattan and supporting constraints, receptively. RMSE is reported for ATE in cm for 10 sequences in TUM RGBD datasets. Numbers in bold in each row represent the best performance for each sequence. Numbers in [ ] show the percentage of improvement over ORB-SLAM2

Dataset	ORB-SLAM2	PP	PP+M	PQ	PPQ+MS
fr1/floor	1.4399	1.3798	**1.3246** [8.01%]	—	—
fr3/cabinet	7.9602	7.3724	**2.1675** [72.77%]	—	—
fr3/str_notex_near	1.6882	1.0883	**1.0648** [36.93%]	—	—
fr3/str_notex_far	2.0007	1.9092	**1.3722** [31.41%]	—	—
fr1/xyz	1.0457	0.9647	0.9231	0.9544	**0.9038** [13.57%]
fr1/desk	2.2668	1.5267	1.4831	1.9821	**1.4029** [38.11%]
fr2/xyz	0.3634	0.3301	0.3174	0.3453	**0.3097** [14.78%]
fr2/rpy	0.3207	0.3126	0.3011	0.3195	**0.2870** [10.51%]
fr2/desk	1.2962	1.2031	1.0186	1.1132	**0.8655** [33.23%]
fr3/long_office	1.5129	1.0601	0.9902	1.3644	**0.7403** [51.07%]

Single view object reconstruction

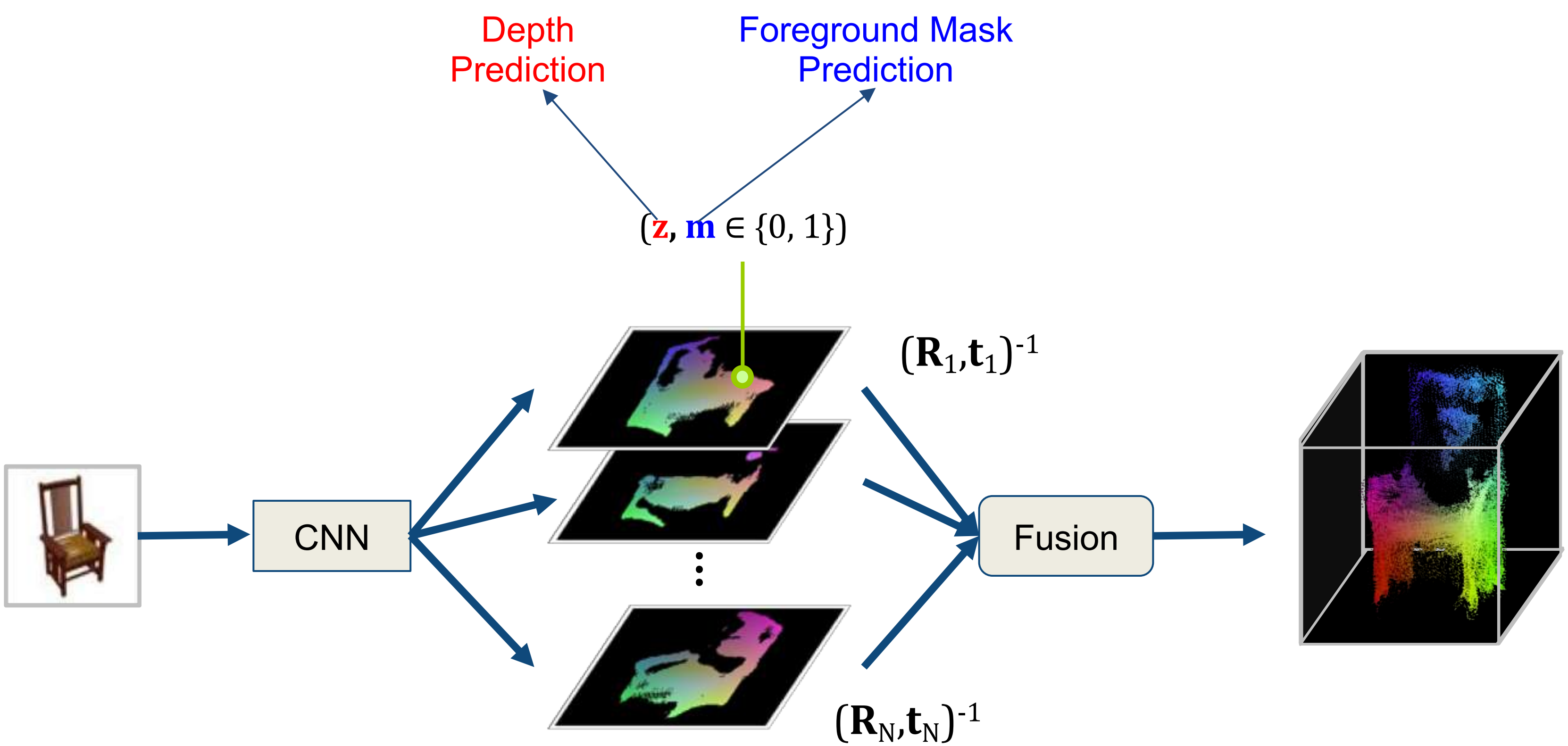
Use deep network to predict 3D shape of an object from a single view

We address two issues:

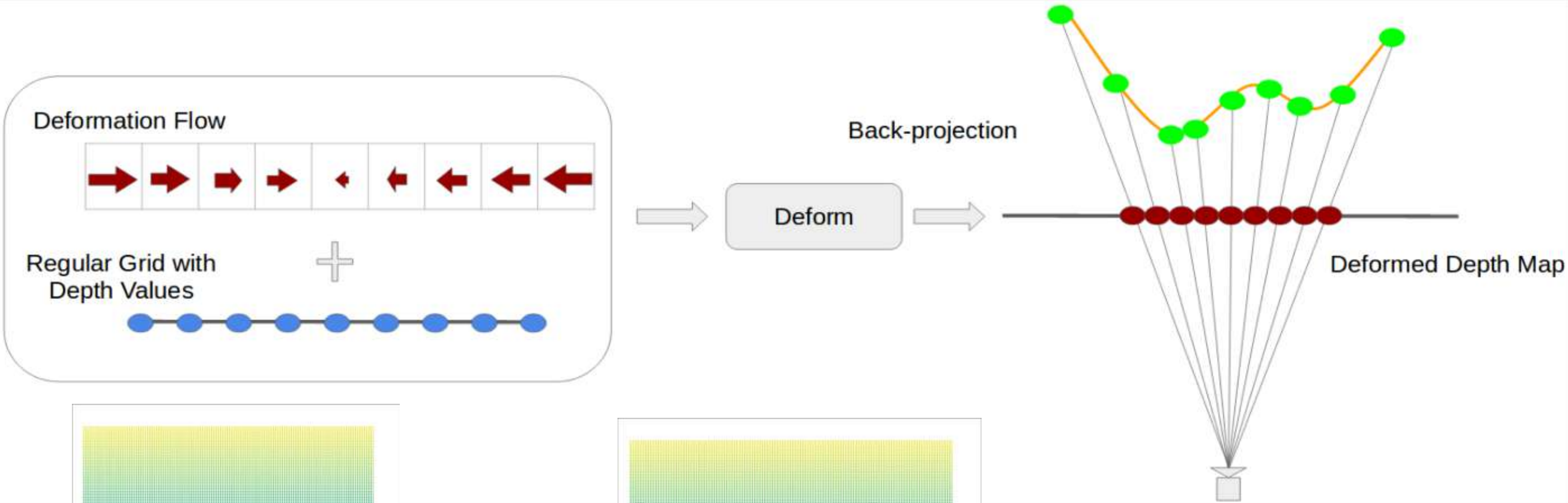
1. How to efficiently reconstruction object 3D shapes with dense point cloud in a deep learning framework.
2. Alleviate noisy point-clouds fused from multiple depth maps using multi-view consistency based on 2D distance fields.

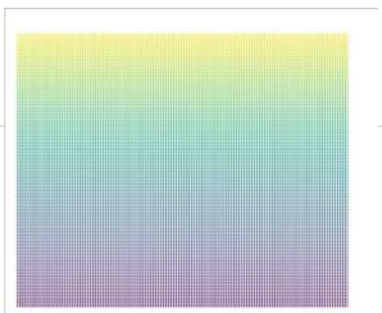
Li, Pham, Zhan, Reid, ECCV 2018

Single view object reconstruction

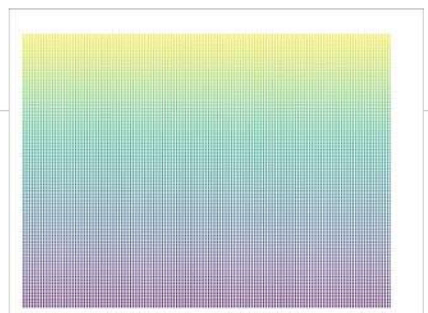


Multi-view Deform-depth Pairs

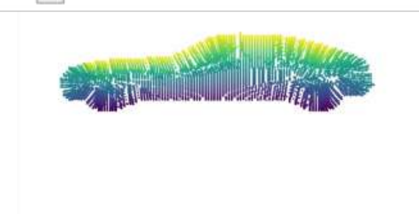




Regular Image Grid before Deformation



Deformation Animation

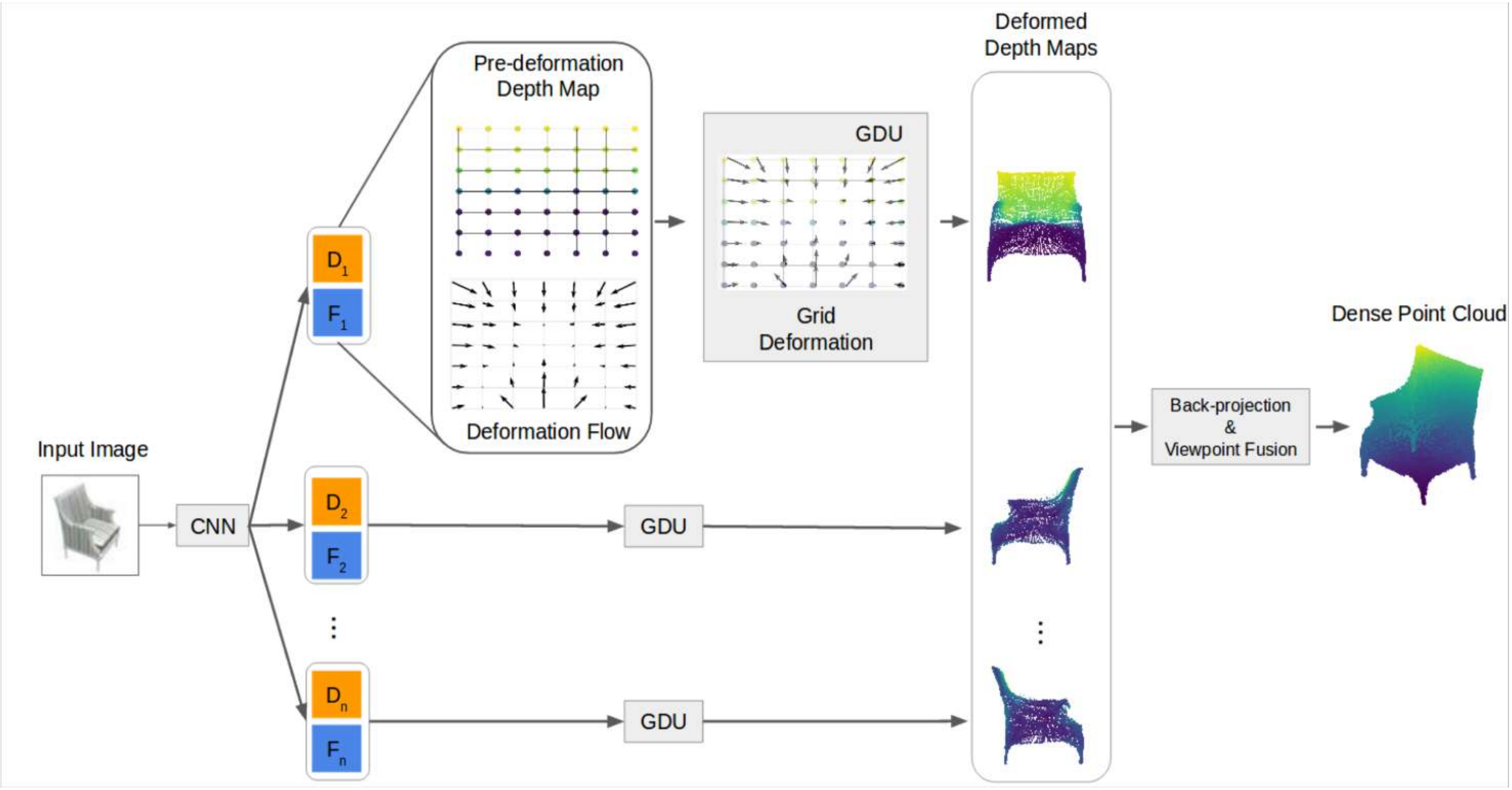


Deformed Image Grid

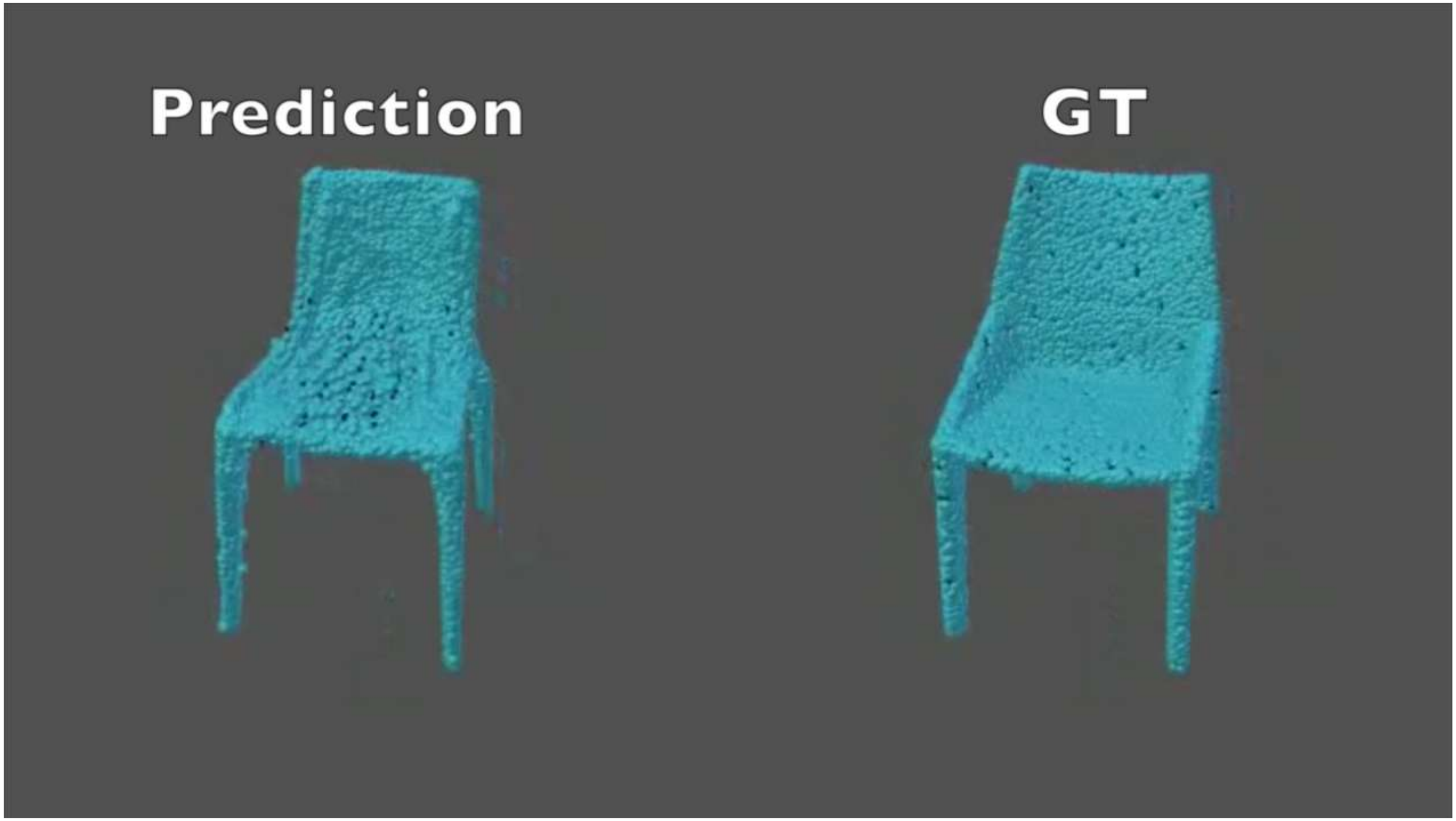
Advantages over depth-mask pairs:

- Efficient memory
- Better surface coverage
- Bypass the need of foreground/background thresholding

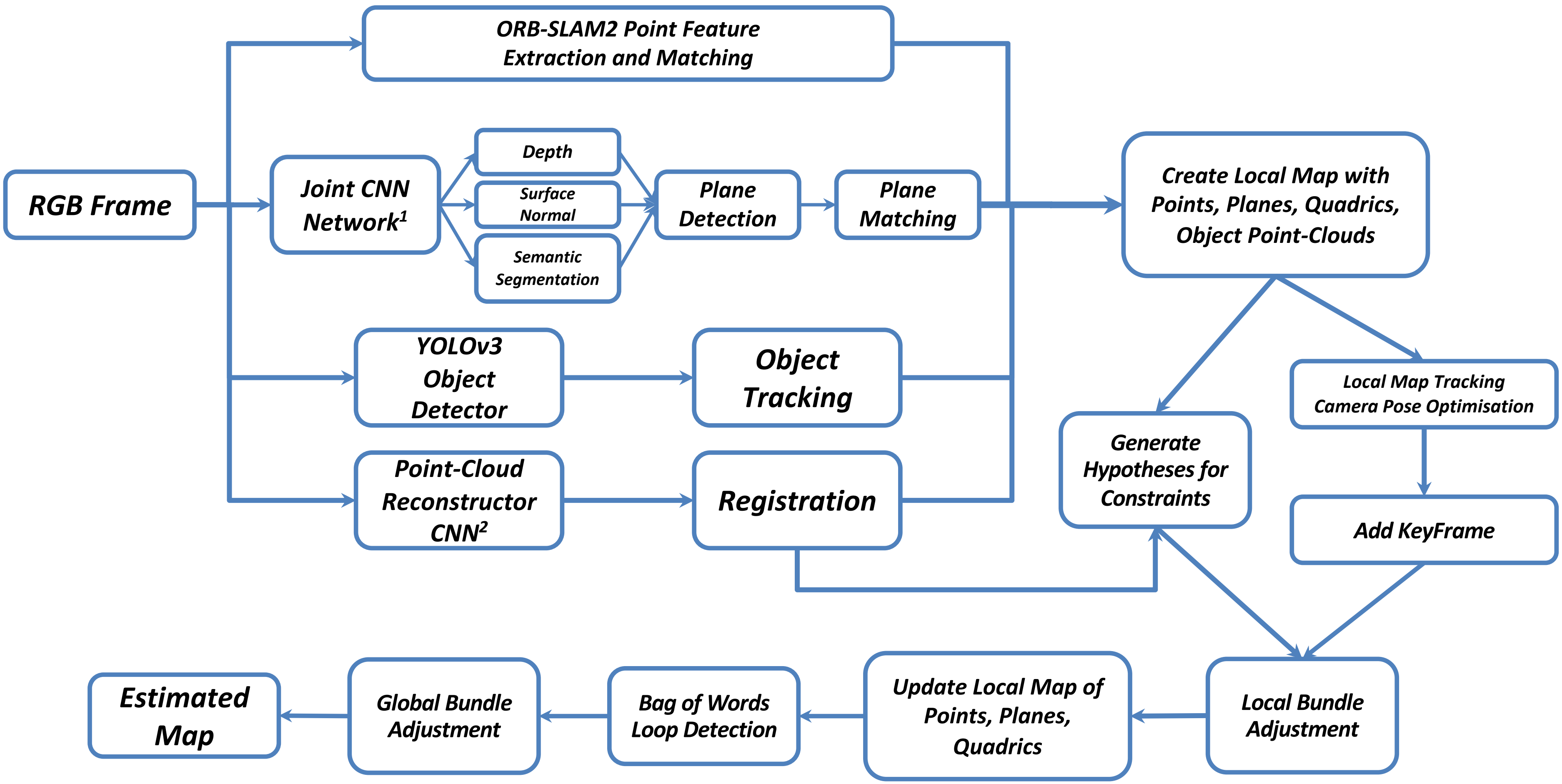
Multi-view Deform-depth Pairs



Results

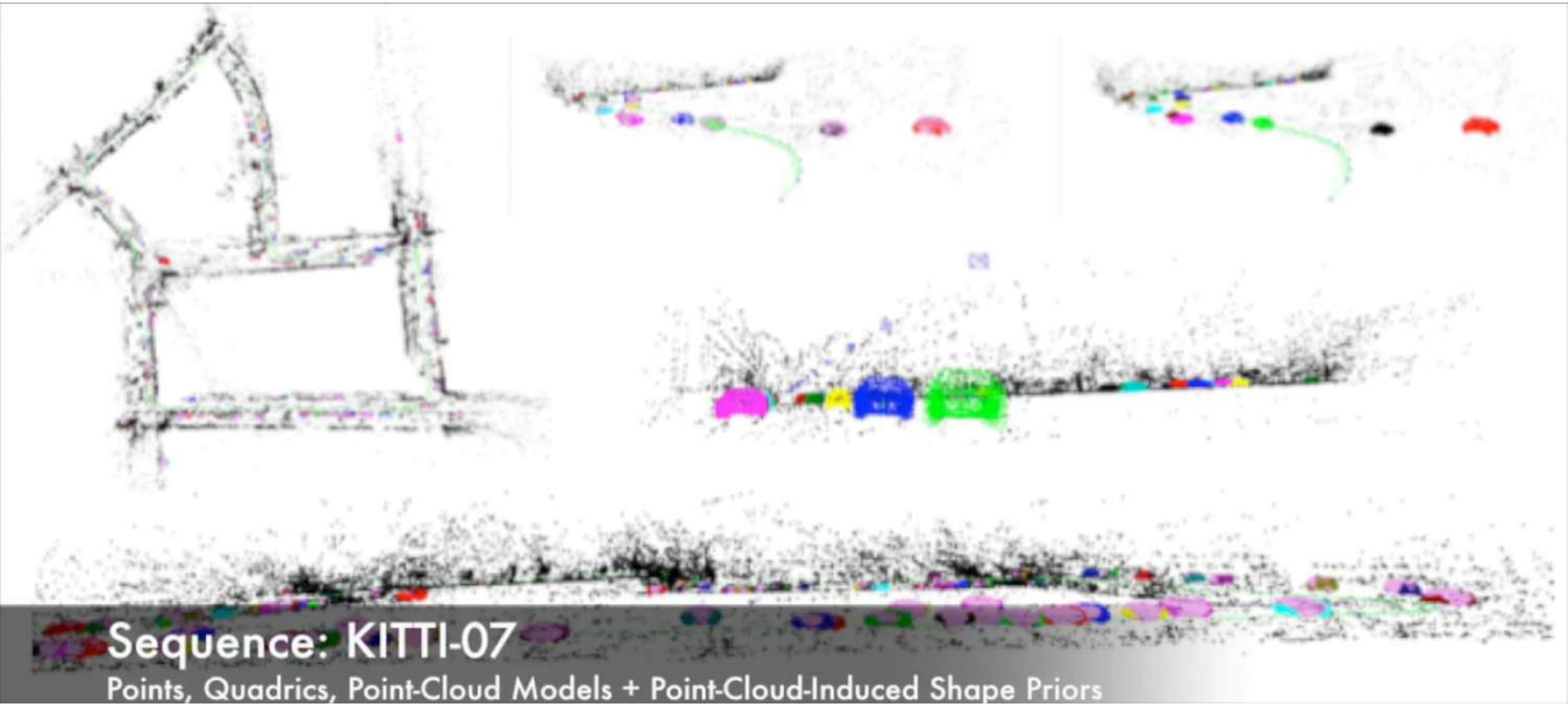


Pipeline of the system



[1]V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, "Real-time joint semantic segmentation and depth estimation using asymmetric annotations," arXiv, 2018.
[2]K. Li, T. Pham, H. Zhan, and I. Reid, "Efficient Dense Point Cloud Object Reconstruction using Deformation Vector Fields", ECCV 2018.

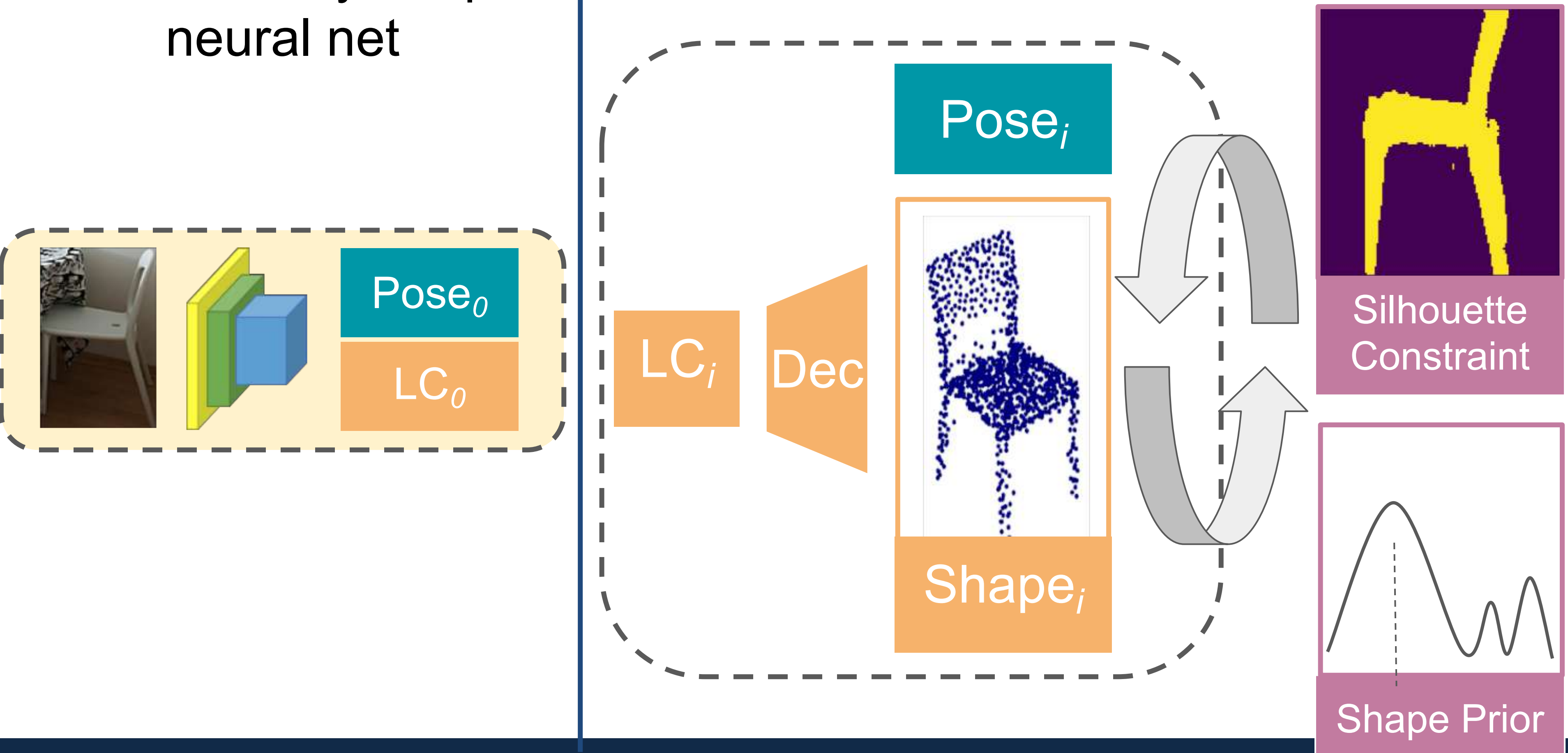
Object-based SLAM



Optimizable Object Reconstruction from a Single View

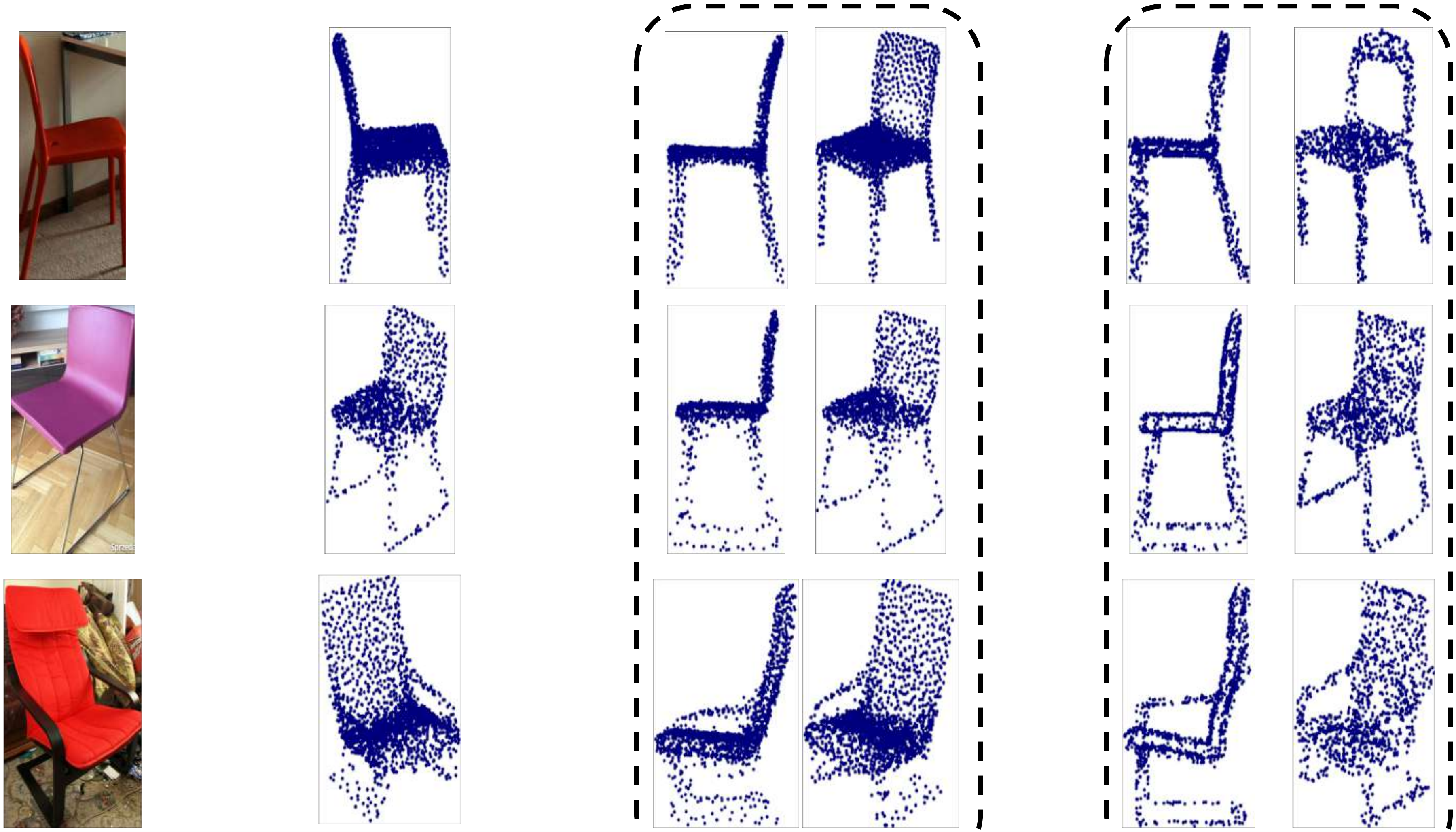
Initialized by deep neural net

Optimization



Qualitative Results on Shape and Pose

Input Image | Aligned View of Reconstruction | 2 Views of Reconstruction | 2 Views of Ground-truth



Conclusions (lessons so far)

- Deep networks can capture semantics and even geometry
  - They should not be a replacement for geometry, but a complement to it
  - Very good at extracting info and relations that we find hard to model explicitly / analytically
  - Provide a better/stronger prior than smoothness for scene regularisation
  - Provide better features for matching
  - Enable richer semantics – even in real-time – which can help with reconstruction
- Working towards bringing all of the above together into a single system
  - Not there yet, but watch this space…

AUSTRALIAN CENTRE FOR
ROBOTIC
VISION

Thank you. Questions?…

Australian Government
Australian Research Council

QUT

THE UNIVERSITY of ADELAIDE

THE AUSTRALIAN NATIONAL UNIVERSITY

MONASH University

CSIRO

UNIVERSITY OF OXFORD

Inria
INVENTORS FOR THE DIGITAL WORLD

Imperial College London

ETH zürich

Georgia Tech