深度数据驱动的重建与交互

3D Reconstruction and Human-Computer Interaction Driven by Depth Data

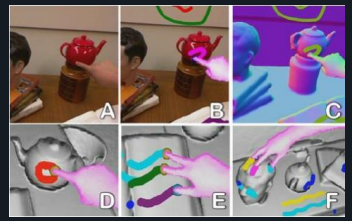
报告人：许威威， 浙江大学CAD&CG国家重点实验室百人计划研究员

国家自然科学基金优秀青年基金获得者

- Background
  - Fast development of commercial RGBD cameras
  - Various applications of digitalized indoor scenes

Holoportation

Holoportation



Kinect Fusion

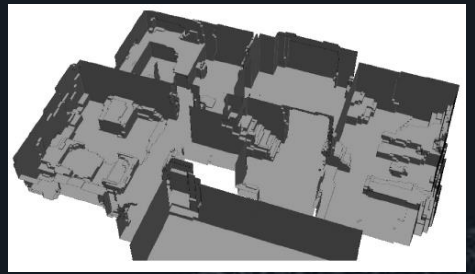
Virtual Try-on

Limitations of Existing Reconstruction Methods

- Existing representations of geometry from depth data
  - Point clouds
  - Signed distance fields
  - Axis-aligned plane proxy
- Lack of semantics
  - Not suitable to applications that require semantic information

[Du et al. 2011]

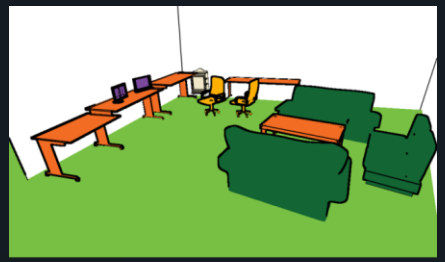
[Izadi et al. 2011]



[Furukawa et al.2009]

Our Goal

- Semantic modeling



High level applications

...

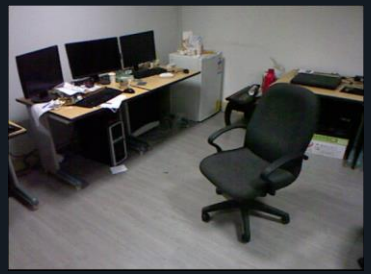
[Yu et al. 2011] [Merrell et al. 2011]

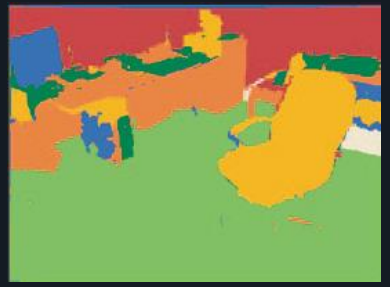
Challenges for Semantic Modeling

- Object segmentation (detection)
  - Automatic methods:
    - Accuracy issue
    - Generalization capability issue
  - Interactive methods:
    - Interaction efforts
- Geometry reconstruction
  - Severe occlusions in indoor scenes
  - Partial and noisy depth data

Our Key Idea

- Combine user interaction and automatic algorithm





Automatic

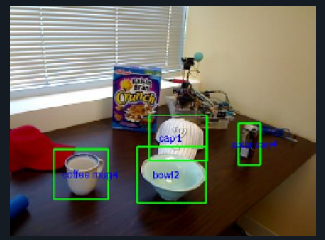
Improved

Our Key Idea

- Search a 3D model database to find models that best approximate the scene geometry

Related Work: Indoor Scene Images Segmentation and Labeling

- CRF model
  - [Xiong and Huber 2010]
  - [Anand et al. 2011]
  - [Silberman and Fergus 2011]
  - [Koppula et al. 2011]
  - [Koppula et al. 2011]
- Object detection
  - [Janoch et al. 2011]
  - [Lai et al. 2010]
- Interactive segmentation
  - [Li et al. 204]



Related Work: Indoor Scene Modeling

- Point cloud
  - [Fox et al. 1999]
  - [Whitaker et al. 1999]
- Image-based modeling
  - [Furukawa and Ponce 2010]
  - [Furukawa et al. 2009a]
  - …
- RGBD Camera
  - [Izadi et al. 2011]
  - [Henry et al. 2012]
  - [Du et al. 2011]

Interactive Context-aware Image Segmentation and Labeling

- Labeling with CRF model

Data term Compatibility term

$$E(C) = \sum_i E_1(C_i : x_i) + \lambda \sum_{i,j} E_2(C_i, C_j)$$

Progressively updated to make the CRF model context aware

C: Image Labeling
x_i: Pixel

Data Term

Appearance term Geometry term

$$E_1(c_i : x_i) = E_a(c_i : x_i^a) + E_g(c_i : x_i^g)$$

$$E_g(c_i : x_i^g) = -\log((1-\alpha_g)P_t(c_i|x_i^g) + \alpha_g P_c(c_i|x_i^g))$$

NYU image database

Previous segmentation result

α_g: weight to blend P_t and P_c

Data Term – Geometry Term

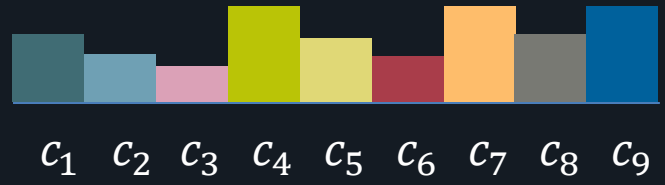
- Local geometry feature $x_i{}^g$
  - Height: h_i
  - Size: s_i
  - Orientation: θ_i

$$P_t(c_i|x_i{}^g) = P_t(c_i|h_i)P_t(c_i|s_i)P_t(c_i|\theta_i)$$

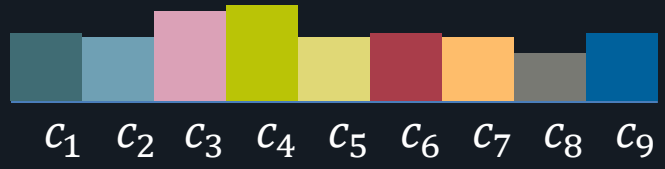
Data Term – Model Updating

- Geometry term updating

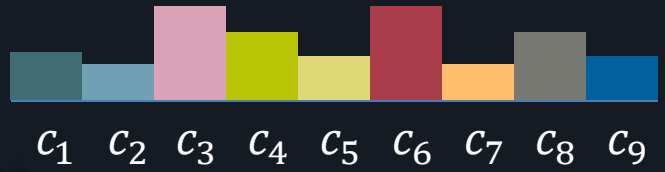
$P_c(c_i|h_i)$



Updated $P_t(c_i|h_i)$



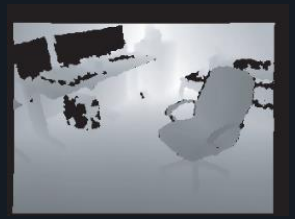
Previous $P_t(c_i|h_i)$



Compatibility Term

- $E_2(c_i, c_j) = \delta[c_i \neq c_j]\mathrm{sim}(\mathbf{f}_i, \mathbf{f}_j)$
  - $\mathbf{f}_i = [r, g, b, d]^T$
    - Concatenation of the RGB values and depth value at pixel i
  - $\mathrm{sim}(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\frac{\|\mathbf{f}_i - \mathbf{f}_j\|^2}{2\sigma^2})$
    - Similarity between two pixels
    - σ: average distances between the features

Experiment Results
- Model Updating in Segmentation

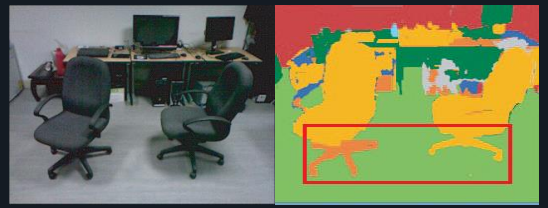


First frame

automatic segmentation

Segmentation result updated according to ***user strokes***

Second frame

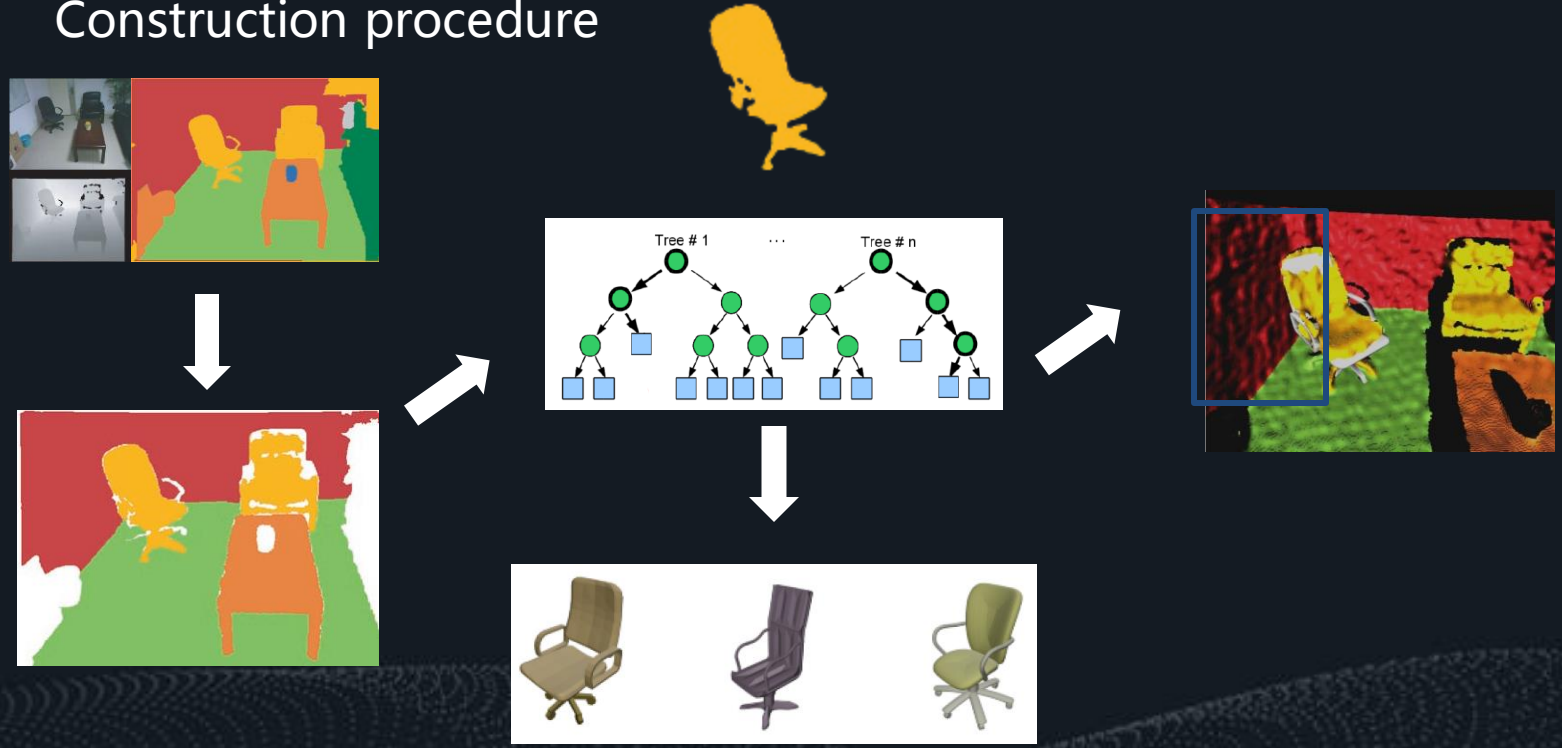


automatic segmentation

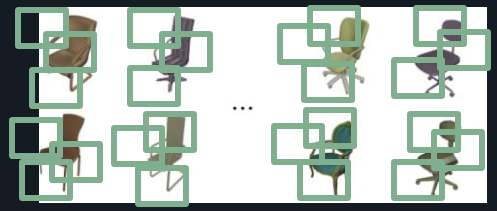
Segmentation result using ***updated appearance and geometry model***

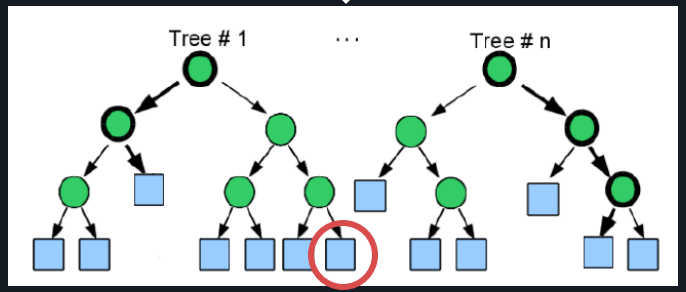
Data-Driven Construction of Indoor Scenes

- Construction procedure



Matching with Random Regression Forest - Training

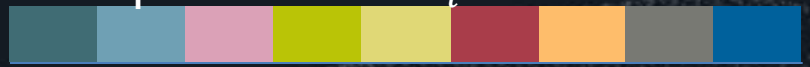




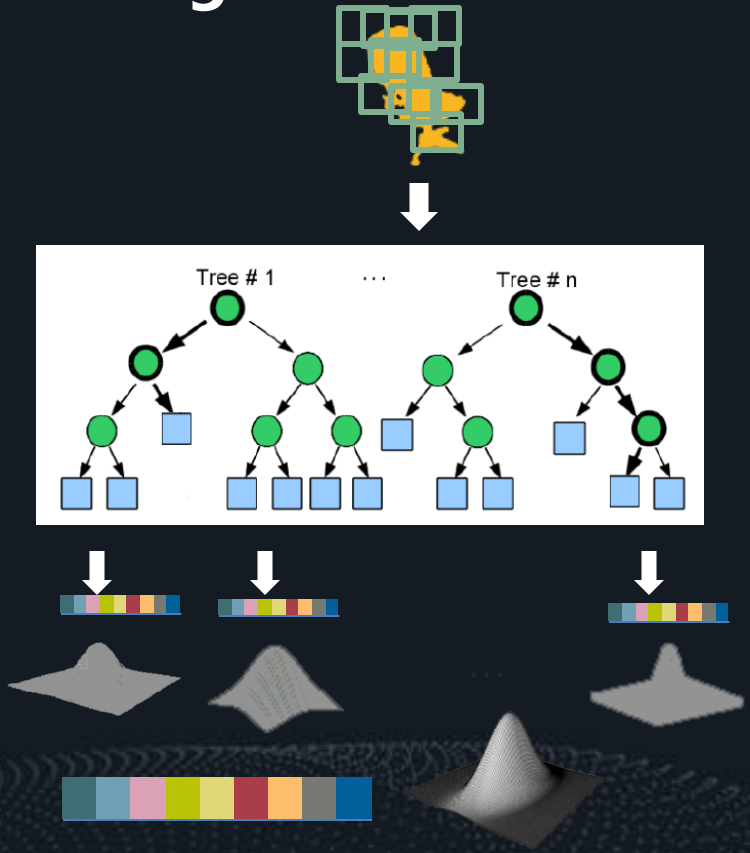
- Patch: $\widehat{P_i} = (\mathbf{I}_i, \theta_i)$
  - $\mathbf{I}_i$: **geometry** features
  - $\theta_i = \{\theta_{yaw}, \theta_{pitch}, \theta_{roll}, t, \mathbf{m}_i\}$

$\{\theta_{yaw}, \theta_{pitch}, \theta_{yaw}, t\}$

Sample number N_i for each class

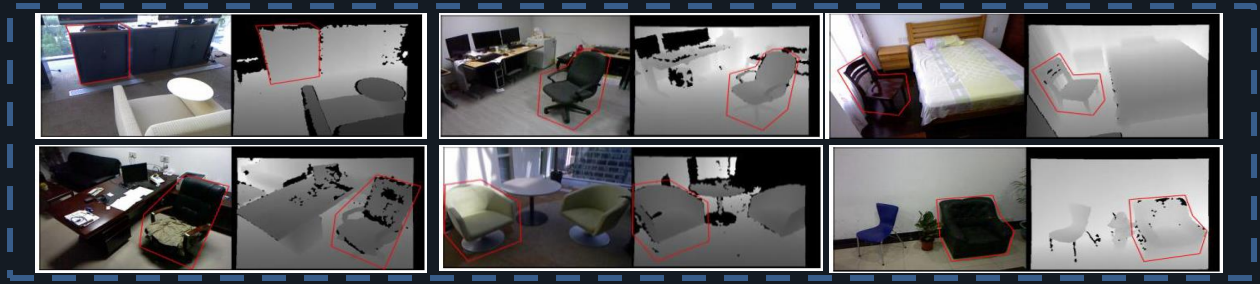


Matching with Random Regression Forest - Testing

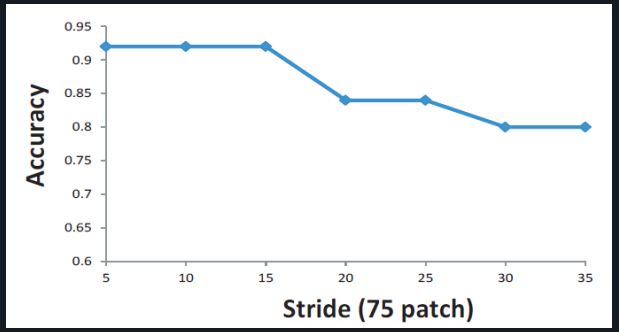


- Dense sampling
- Average all candidate votes on the leaves to get the model label distribution:
- $p(\mathbf{m}|\mathbf{O}) = \frac{1}{K \times N} \sum_{i=1}^{n} \sum_{j=1}^{K} p(\mathbf{m}_j | \widehat{P}_i)$
- Cluster $\{\theta_{yaw}, \theta_{pitch}, \theta_{yaw}, t\}$ to remove noise

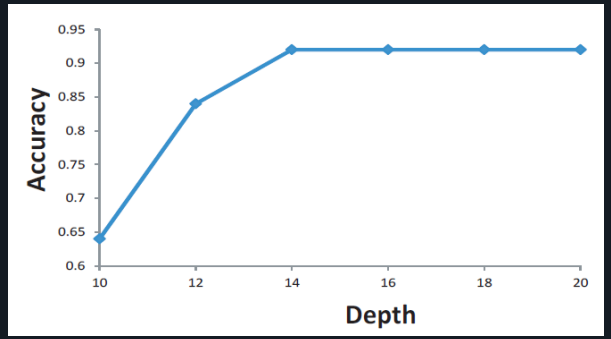
Experiment Results - Model Matching Accuracy



Examples of segmented objects

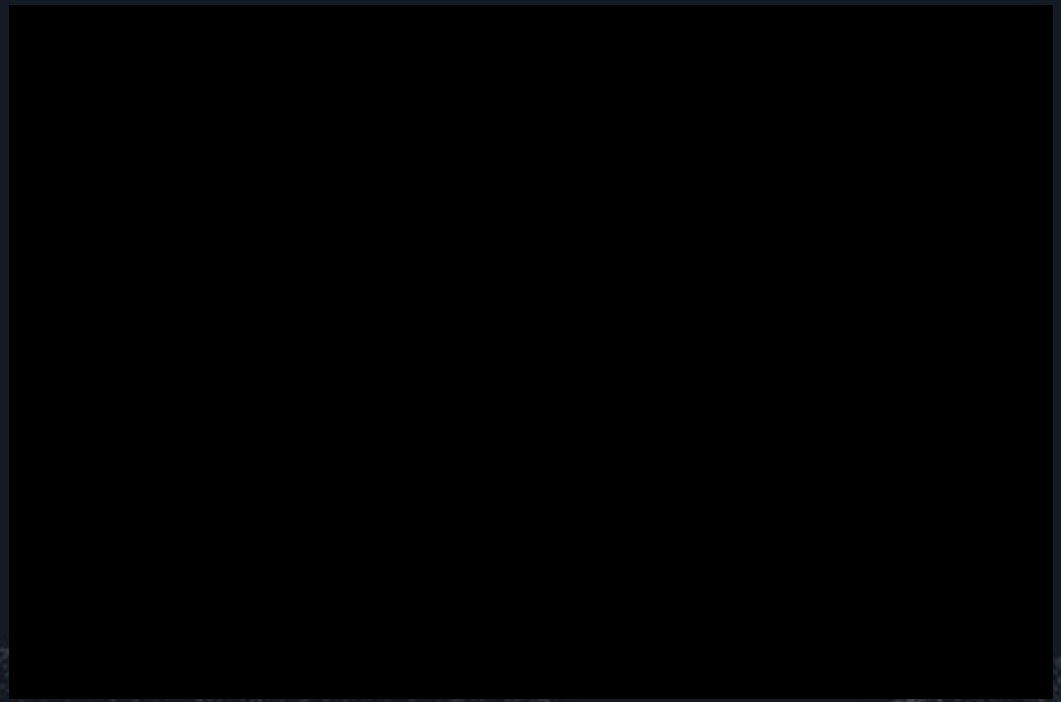


Accuracy as function of testing stride



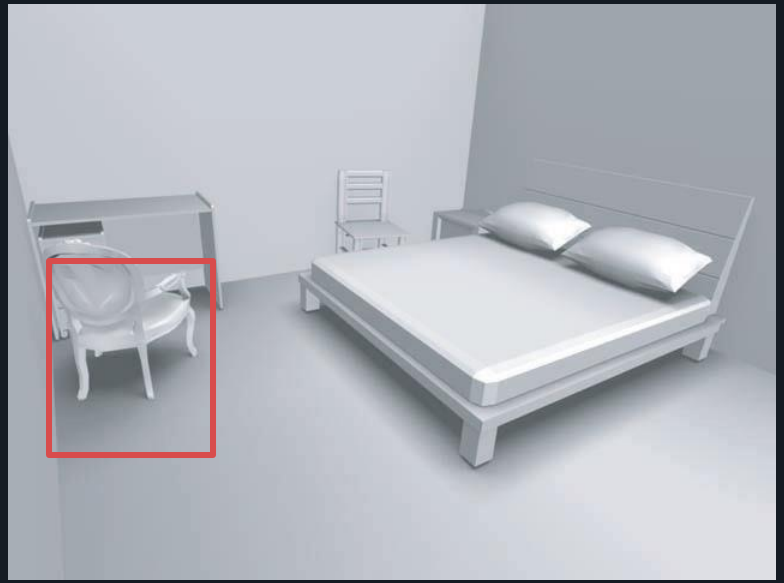
Accuracy as function of tree depth

Experiment Results – Modeling Processing



Experiment Results – More Result

Experiment Results – Failure Case



Online Structure Analysis for Real-time Indoor Scene Reconstruction

Yizhong Zhang *
Weiwei Xu †
Yiying Tong ‡
Kun Zhou *
*Zhejiang Univ.
†Hangzhou Normal Univ.
‡Michigan State Univ.

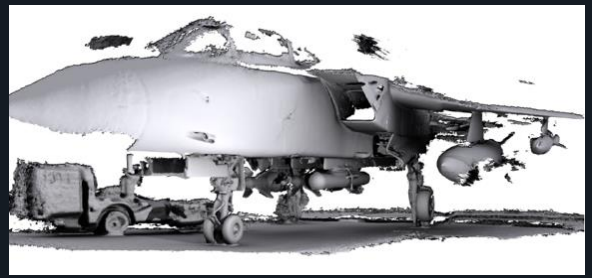
Our Goal

Our Goal

My lab, over 100m^2, scanned in 70 minutes

Related works

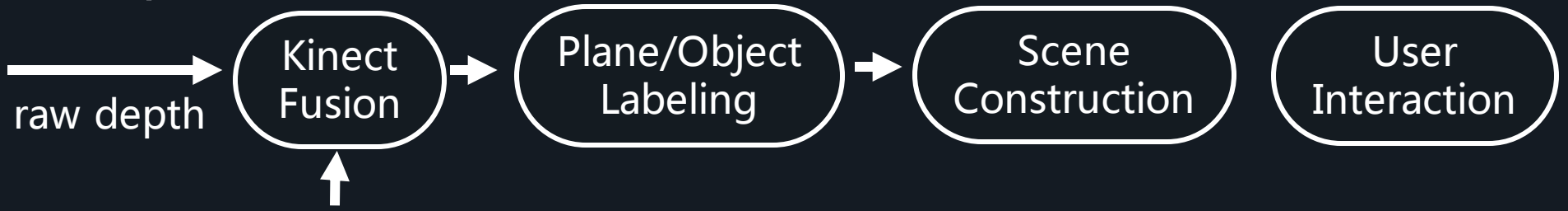
[Nießner et al. 2013]

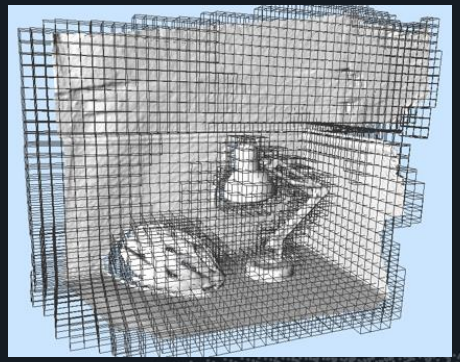


[Chen et al. 2013]

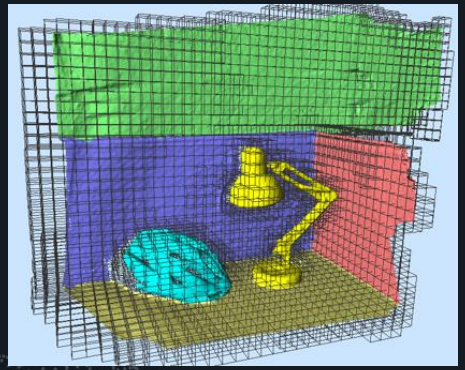
[Zhou et al. 2013]

Pipeline

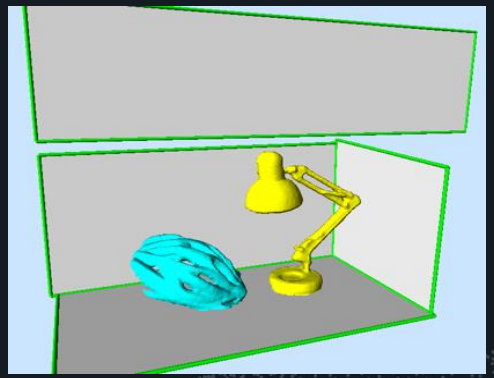




KinectFusion

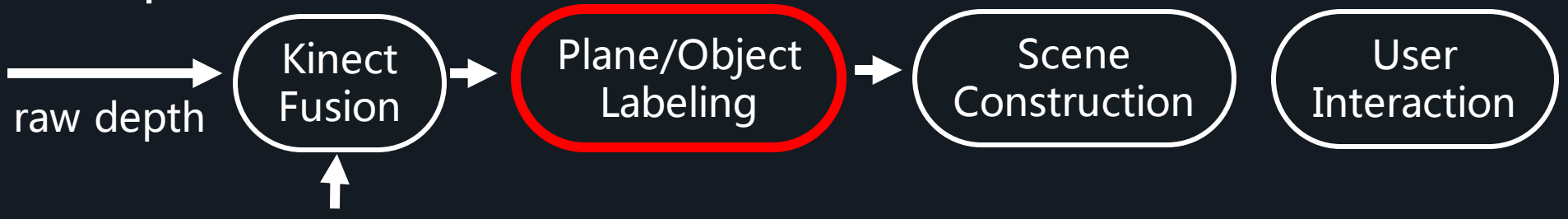


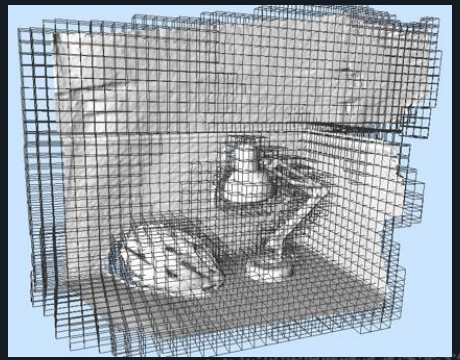
Labeled Volume



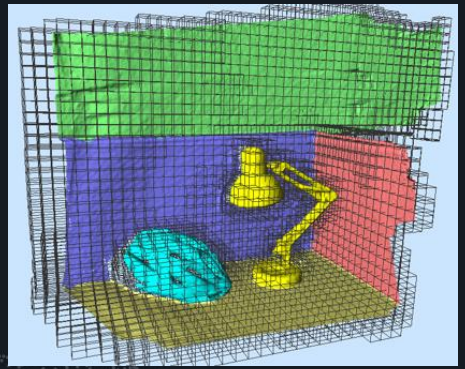
Structured Scene

Pipeline

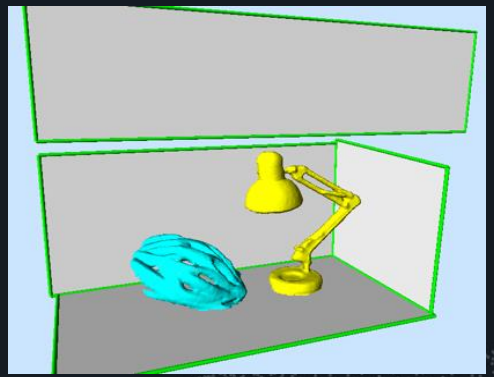




KinectFusion

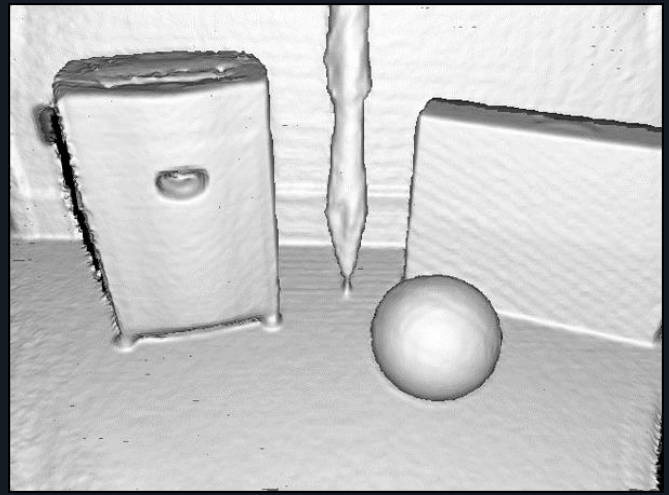


Labeled Volume

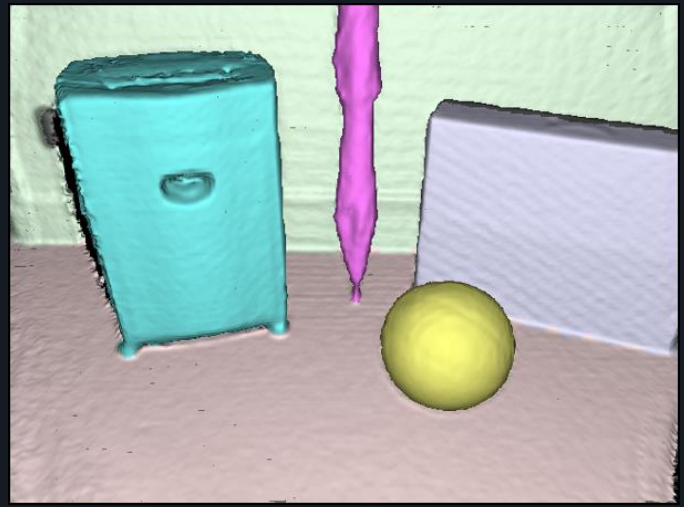


Structured Scene

Labeling

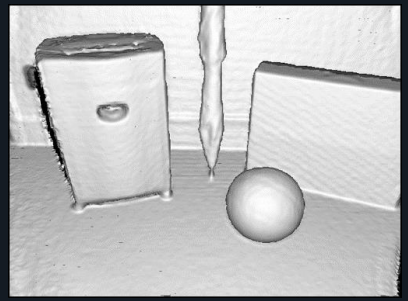


Before Labeling

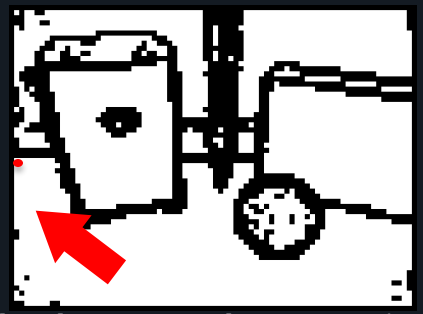


After Labeling

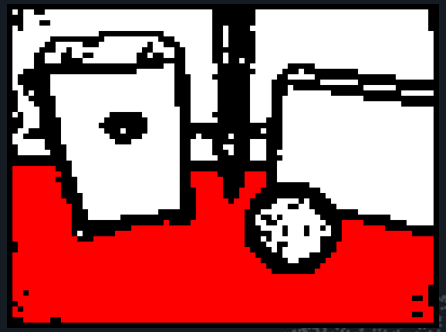
Plane Detetection



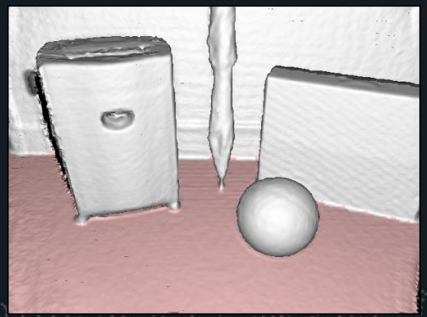
Ray casting depth



Label non-planar pixels

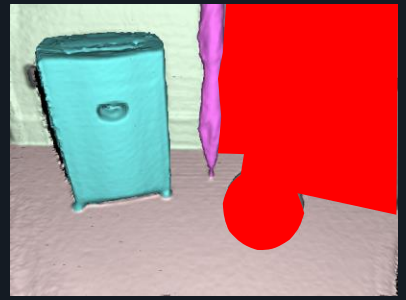


Flood fill



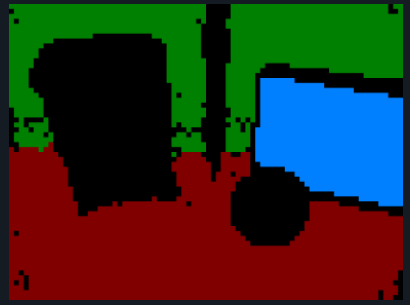
Plane detected

Labeling



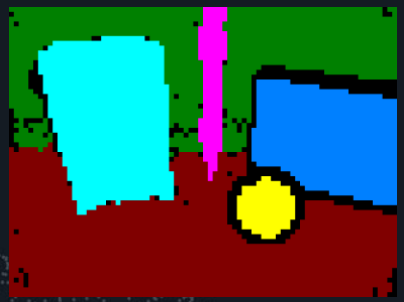
Before Labeling

Label Existing Planes

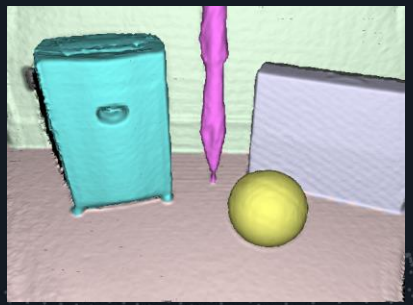


Label New Planes

Label Existing Objects



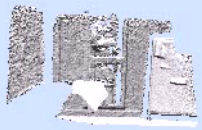
Label New Objects



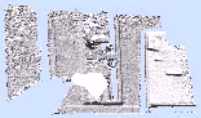
After Labeling

Drifting Reduction

25 meters corridor, capture at 30fps

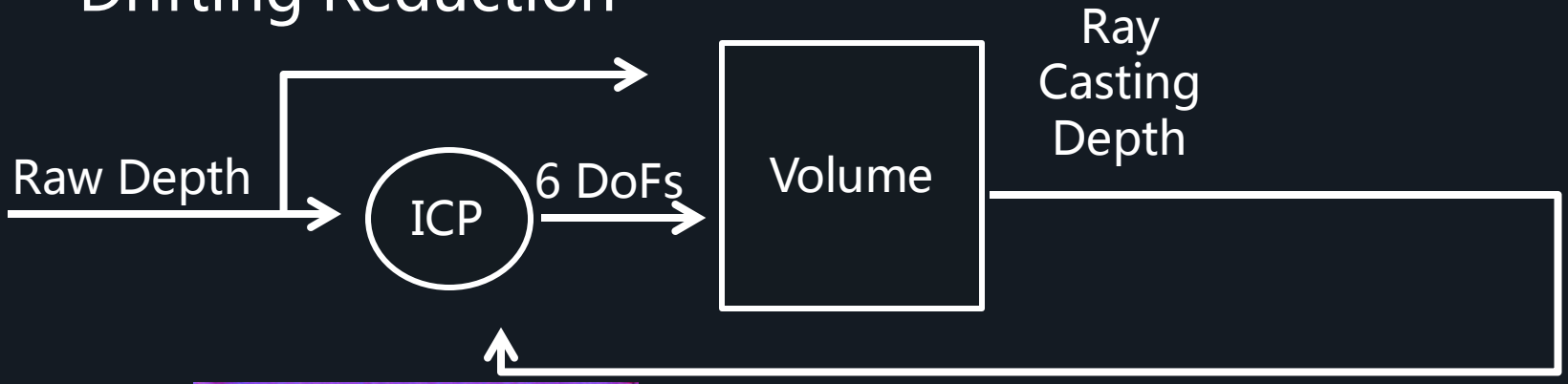


KinectFusion



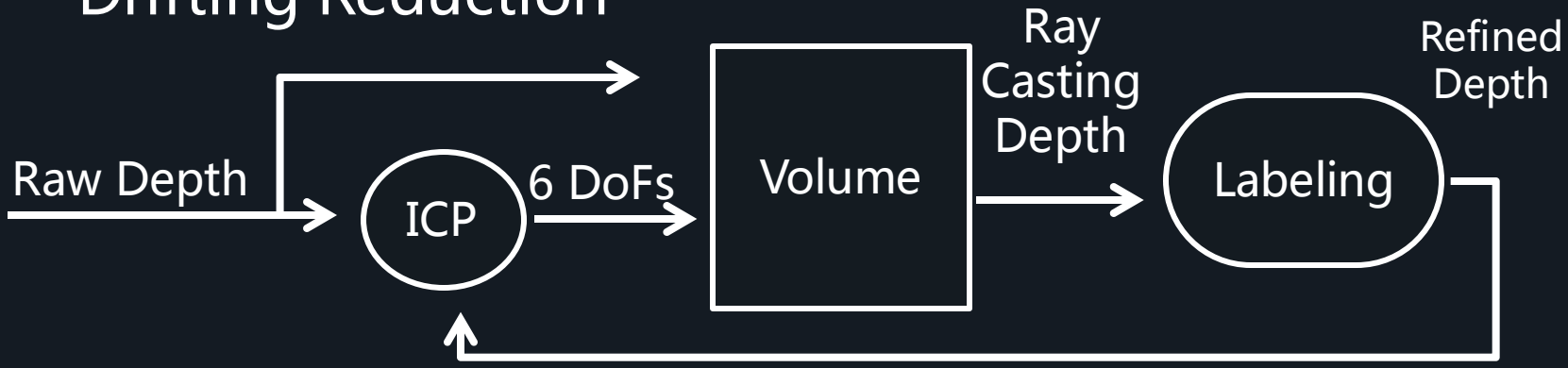
Our Method

Drifting Reduction

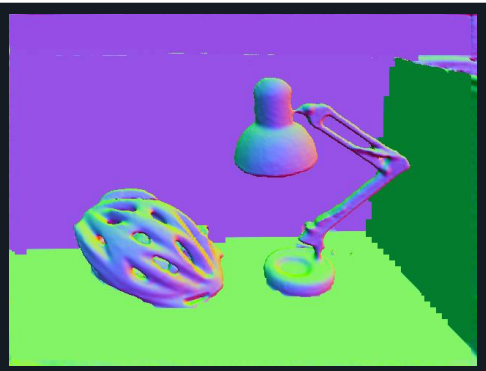


Ray Casting Depth

Drifting Reduction

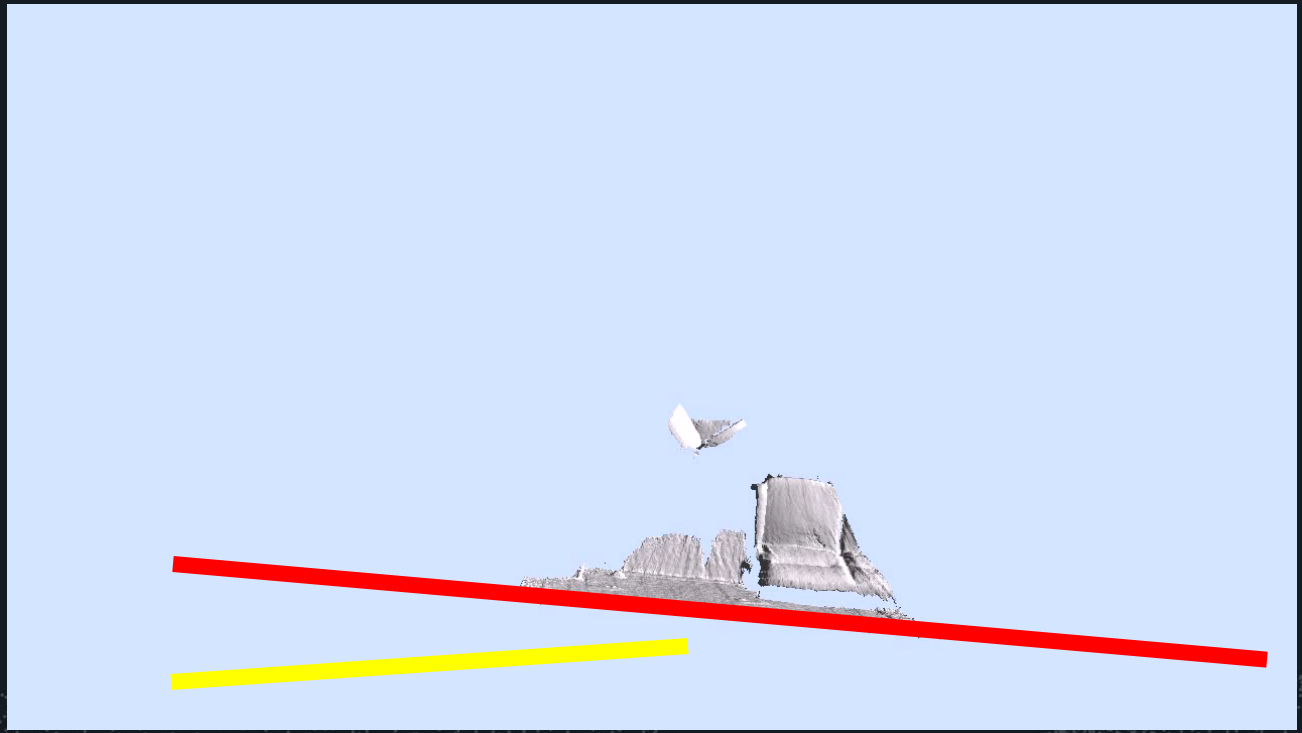


Ray Casting Depth

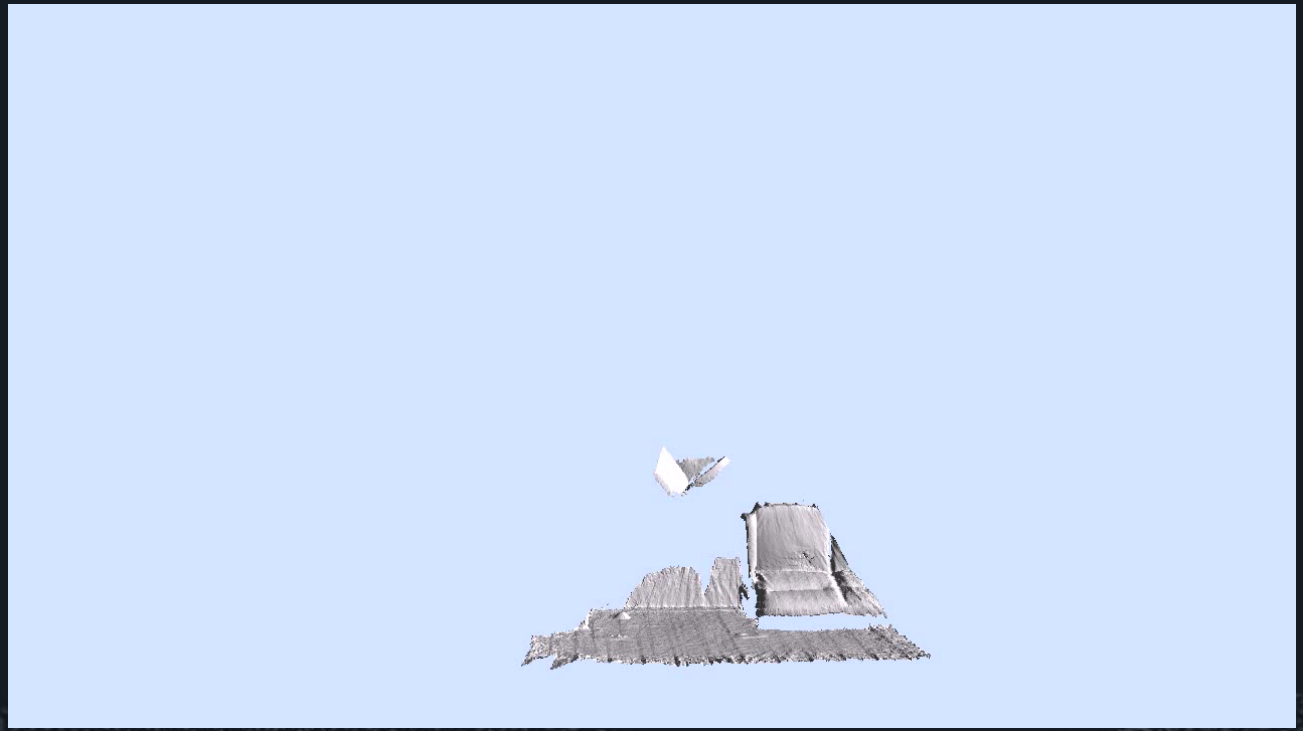


Refined Depth

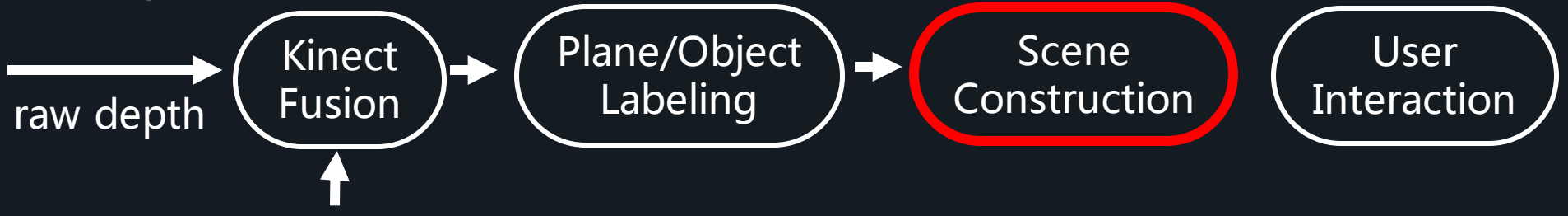
Drifting Reduction

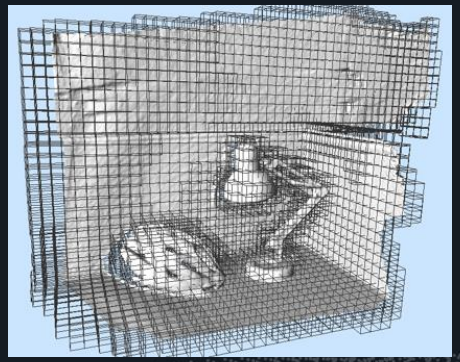


Drifting Relief

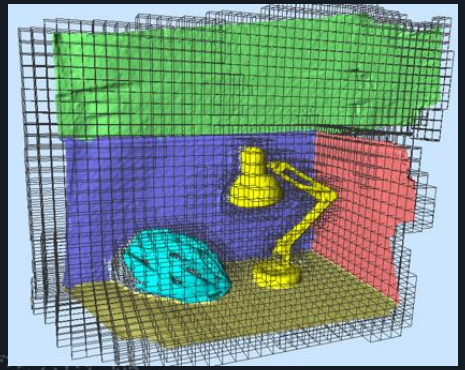


Pipeline

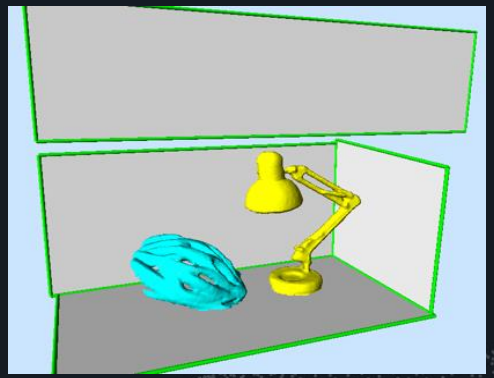




KinectFusion

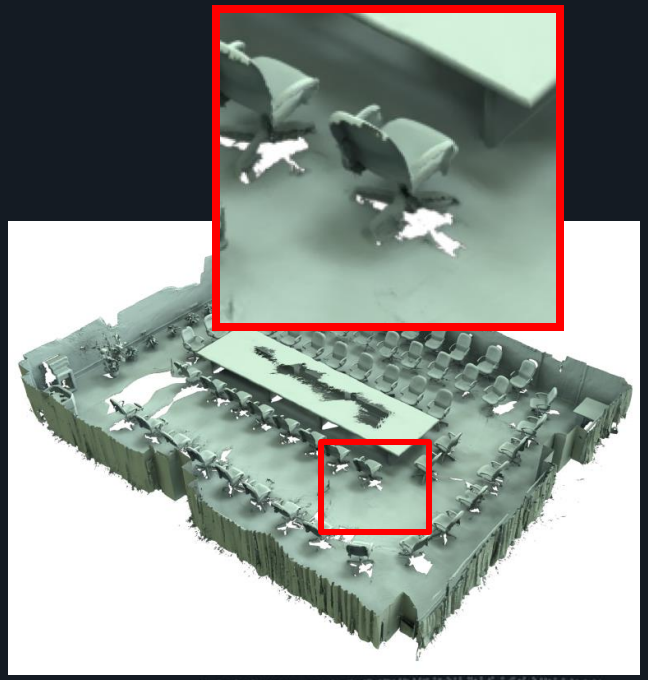


Labeled Volume

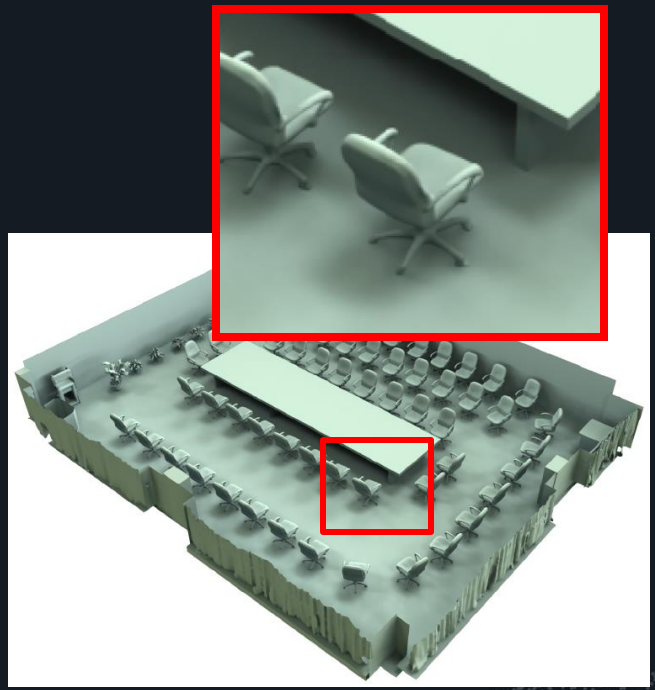


Structured Scene

Scene Construction

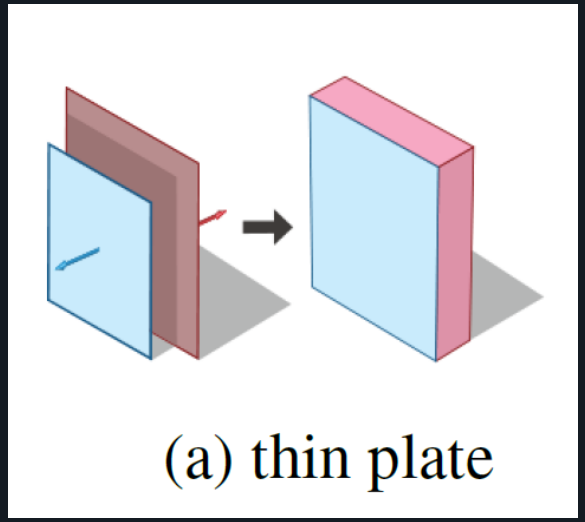


Raw Data

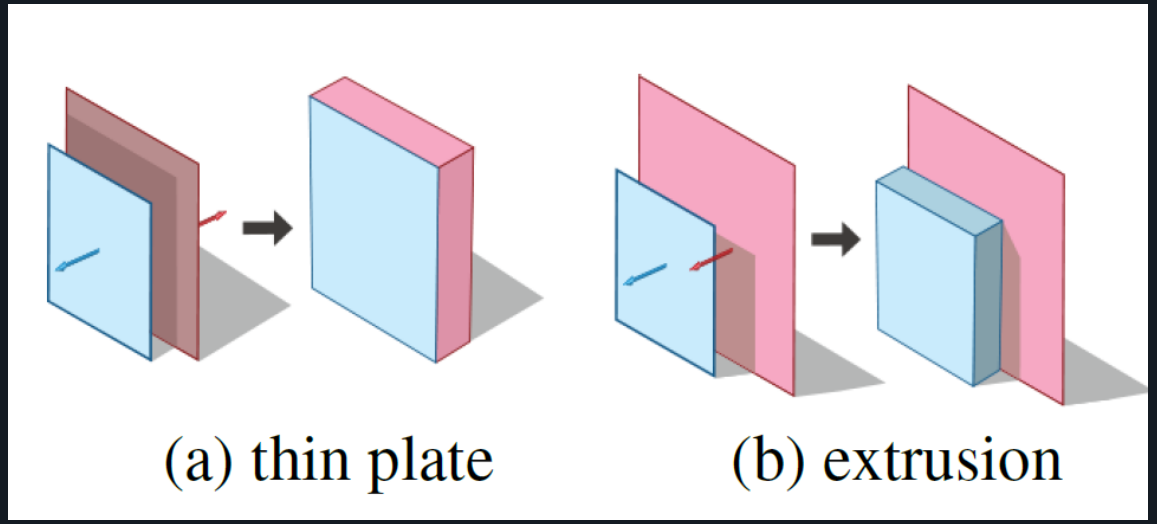


Constructed Scene

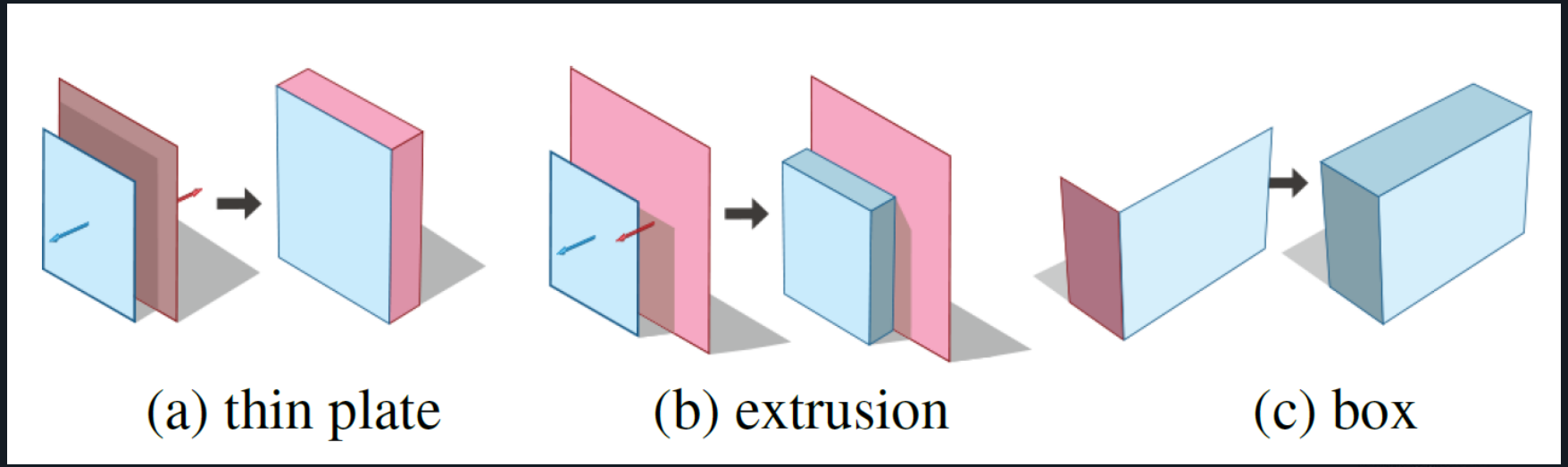
Rectilinear Structure Heuristics



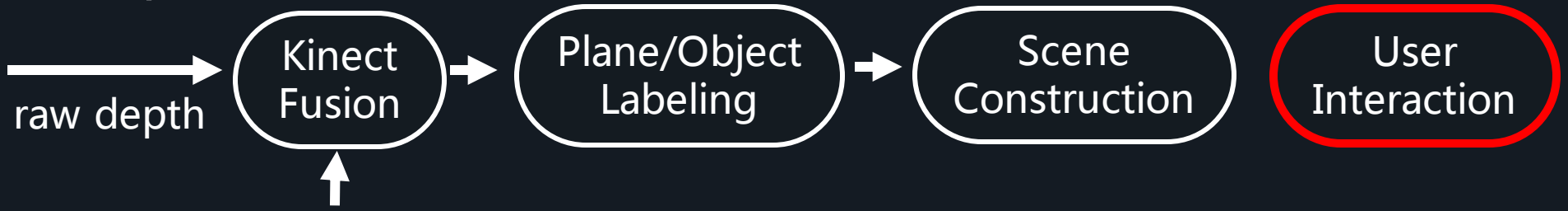
Rectilinear Structure Heuristics

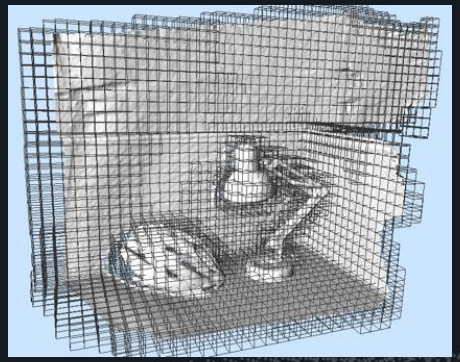


Rectilinear Structure Heuristics

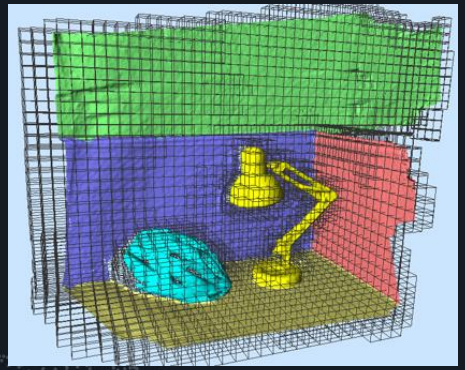


Pipeline

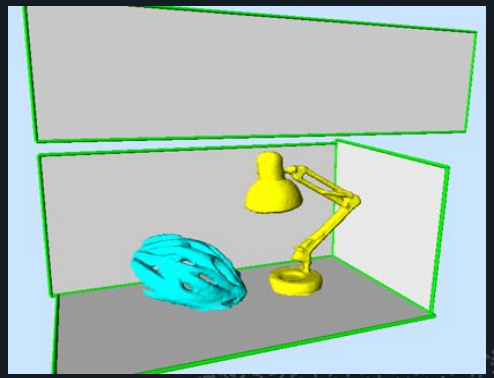




KinectFusion

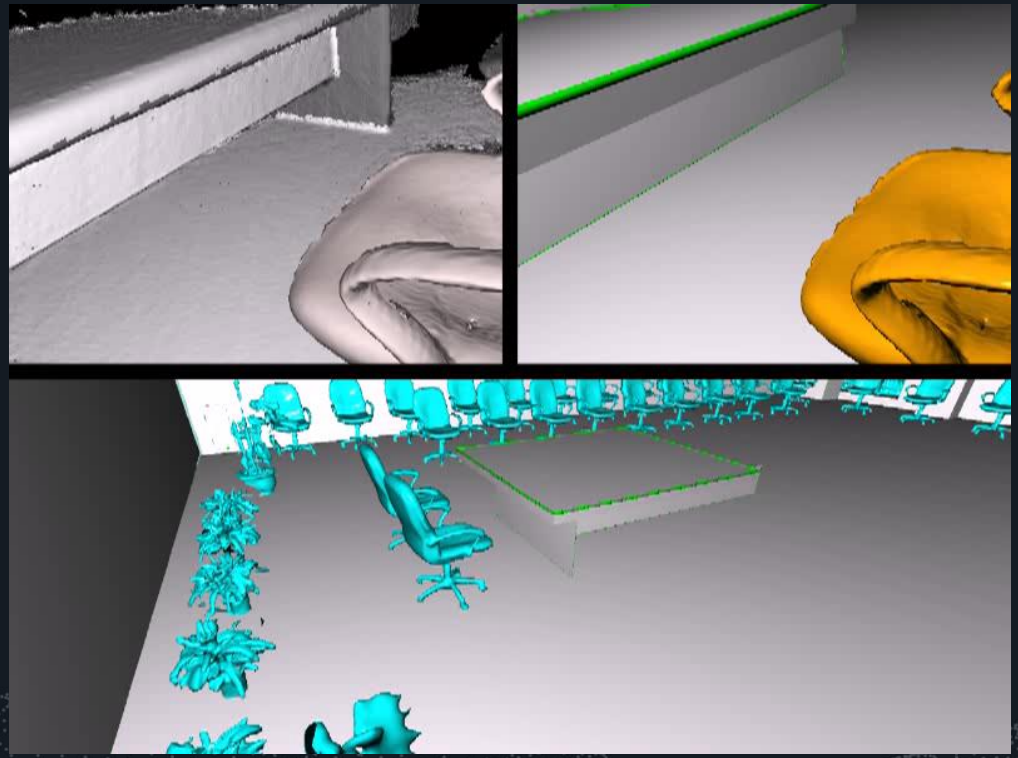


Labeled Volume



Structured Scene

User interaction



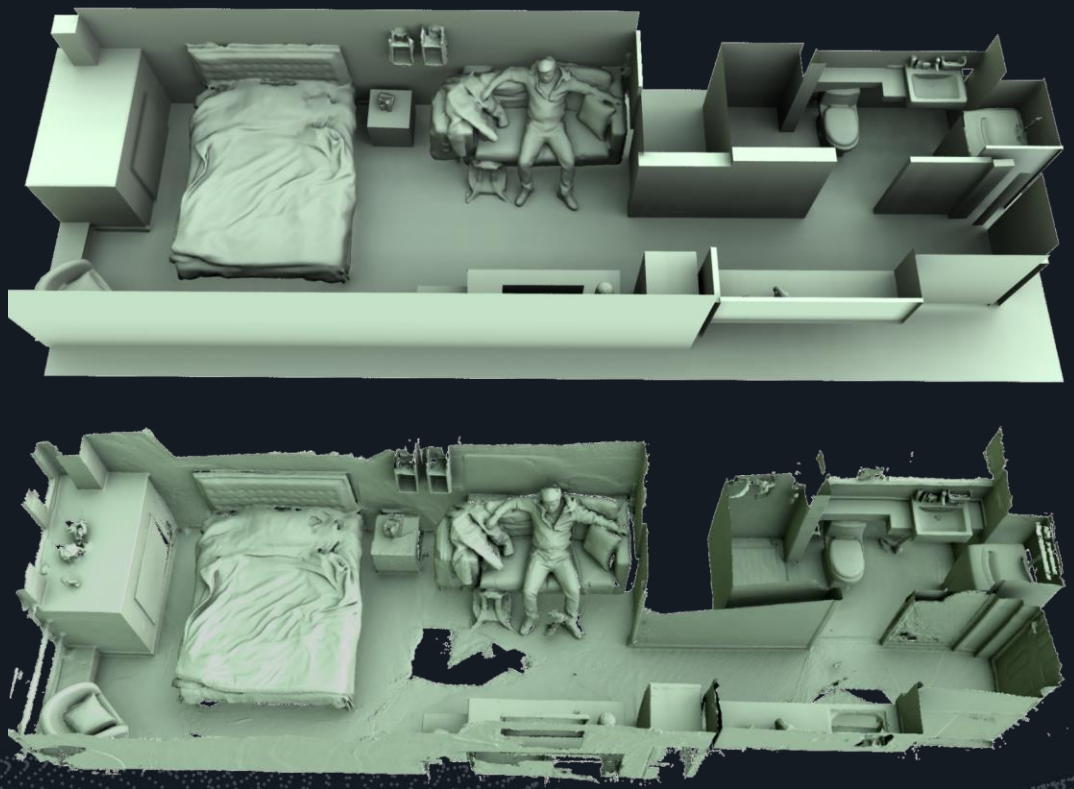
Result

Meeting room, 140m^2, 56 duplicate chairs, 40 minutes

Result

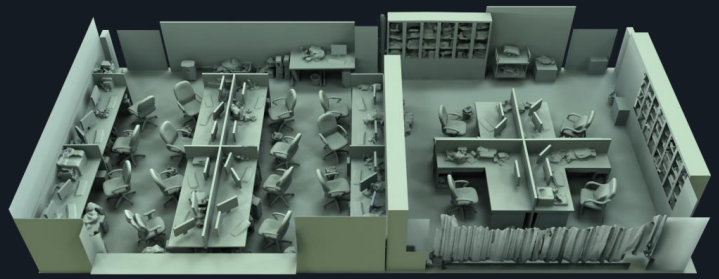
Apartment, multiple rooms, 15 minutes

Result



Apartment, multiple rooms, 15 minutes

Discussion



Conclusion

- An real-time indoor-scene reconstruction system
- Online segmentation
- Improved accuracy

飞天·智能

APSARA INTELLIGENCE