Single-Image Piece-wise Planar 3D Reconstruction via Associative Embedding

Shenghua Gao

@ShanghaiTech University

A revisit of 3D reconstruction

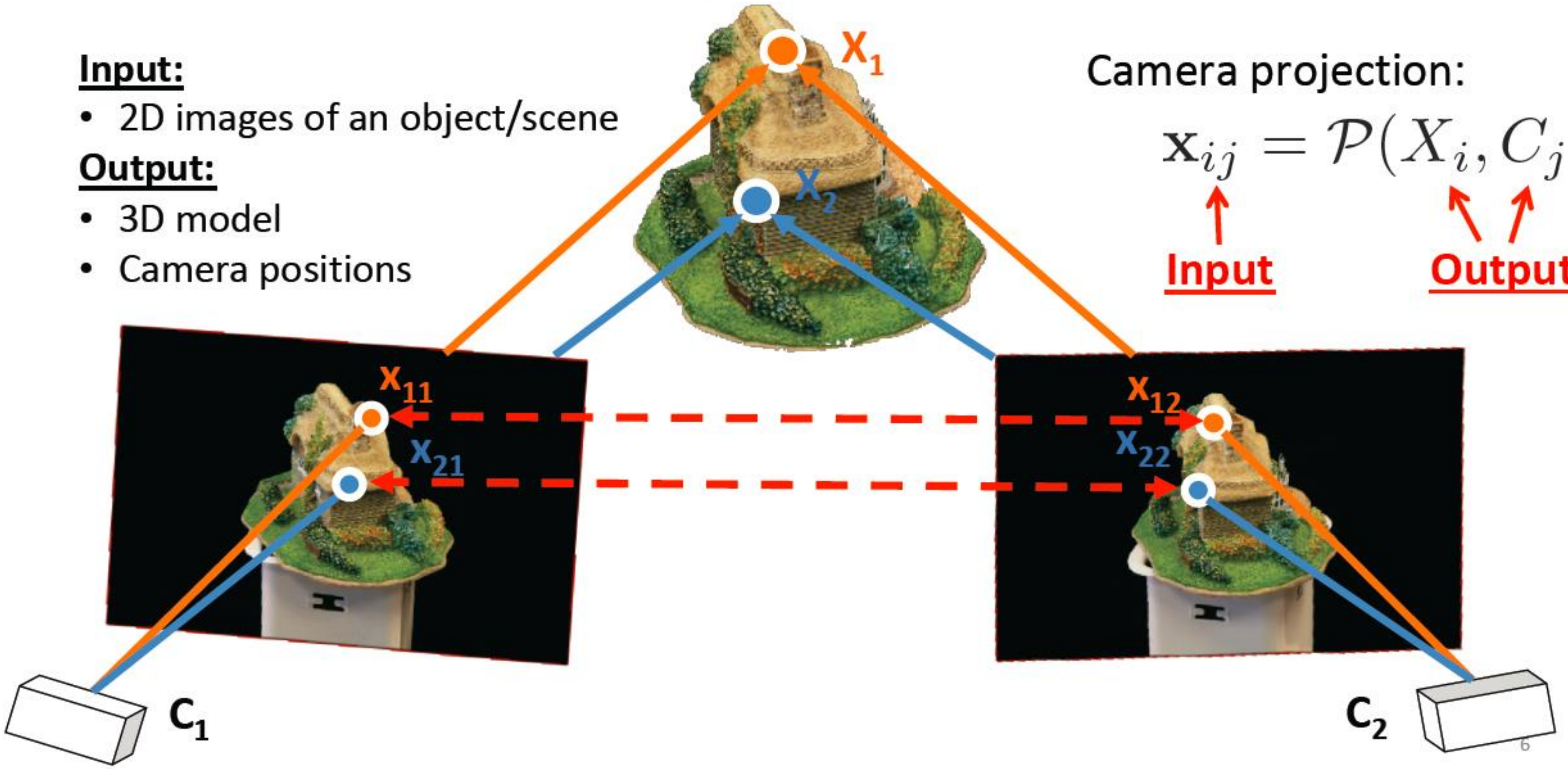
Input:
- 2D images of an object/scene

Output:
- 3D model
- Camera positions

Camera projection:

$$\mathbf{x}_{ij} = \mathcal{P}(X_i, C_j)$$

Input Output



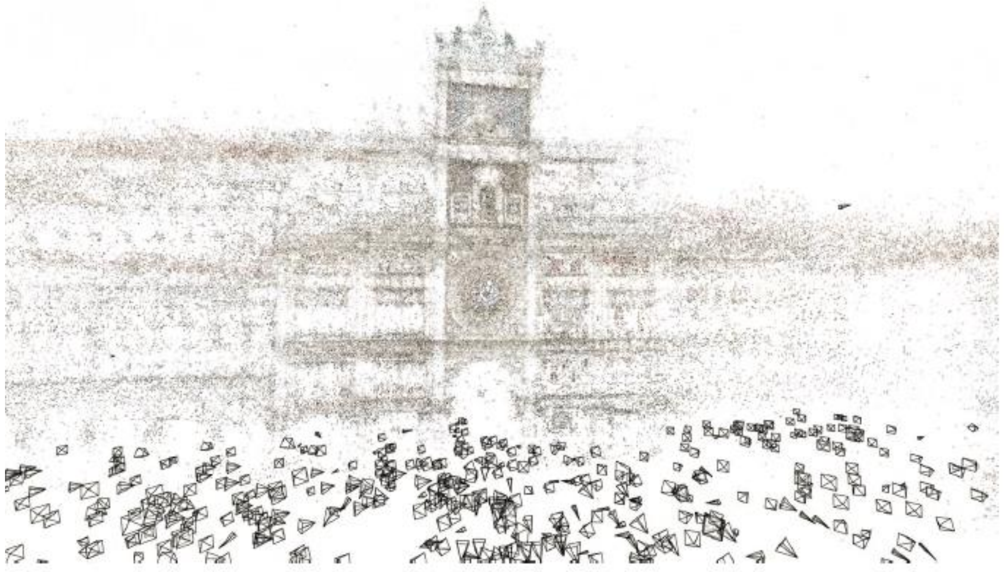
Credited to Z. Zhou @ Penn State

A revisit of 3D reconstruction

- 3D reconstruction is an optimization problem!
- Objective function of reprojection

$$\min \sum_{ij} ||X_{ij} - P(\mathrm{X}_i, C_j)||$$

- Current methods can well handle…
  - Millions of images
  - Varying camera poses
  - Different lighting conditions
  - Image noises

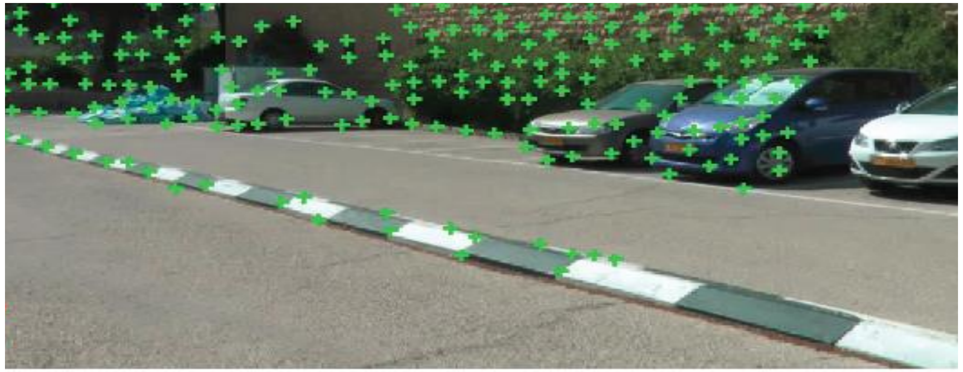


San Marco Square, 14,079 images, 4,515,157 points

Building Rome in a Day. Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz and Richard Szeliski. Communications of the ACM, 2011.

But in practice, feature point based matching approaches still fail

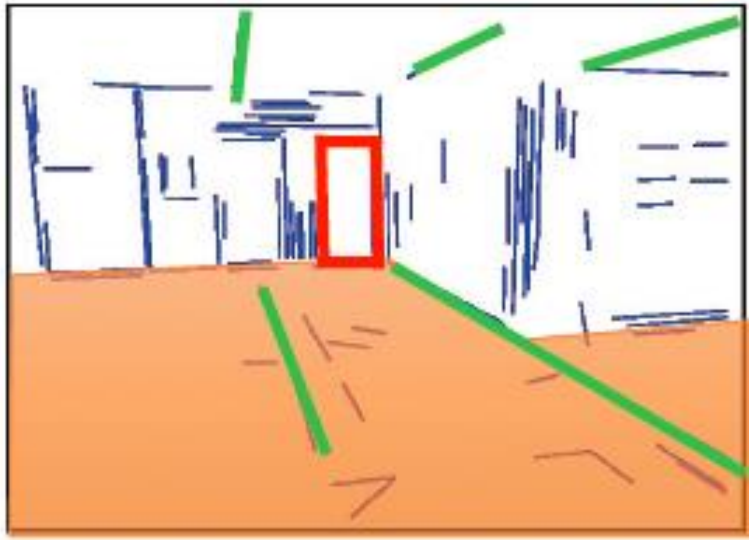
Repetitive and symmetric patterns



Textureless areas

Dynamic scenes

The role of structure in human 3D perception



- Structure: **spatial relationships** among multiple points, lines, patches, etc.
- Human perceives 3D space by recognizing **many types of structure** in the

scene: *parallelism, planar surfaces, regular shapes, repetitive patterns, symmetry, self-symmetry,…*

Geometric Reasoning for Single Image Structure Recovery.
David C. Lee, Martial Hebert, and Takeo Kanade. CVPR 2009.

Our work: Structure learning for 3D Vision

Structure learning

- Line detection
- Plane detection
- Room layout estimation

global structure

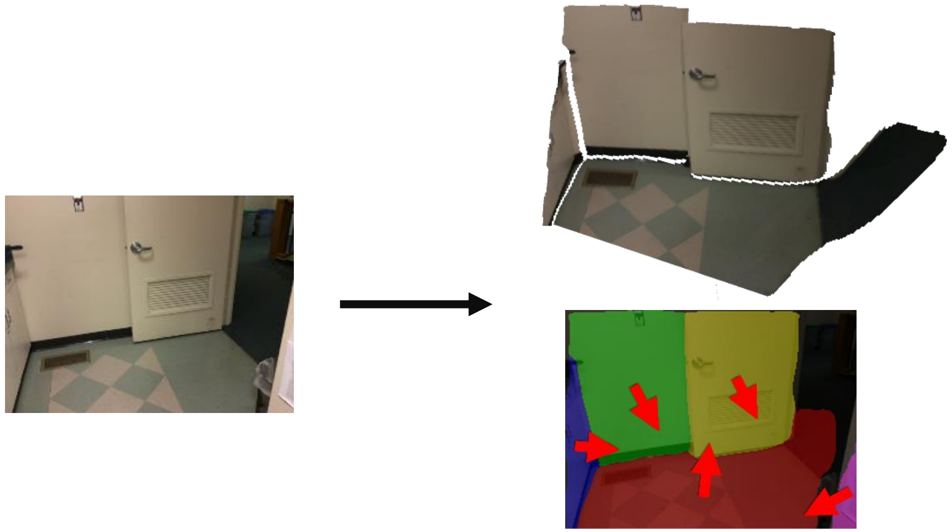
vs.

local structure

Challenges in feature matching

- Feature mismatching
- Textureless regions
- Dynamic objects

Line detection



Plane detection

Room layout estimation

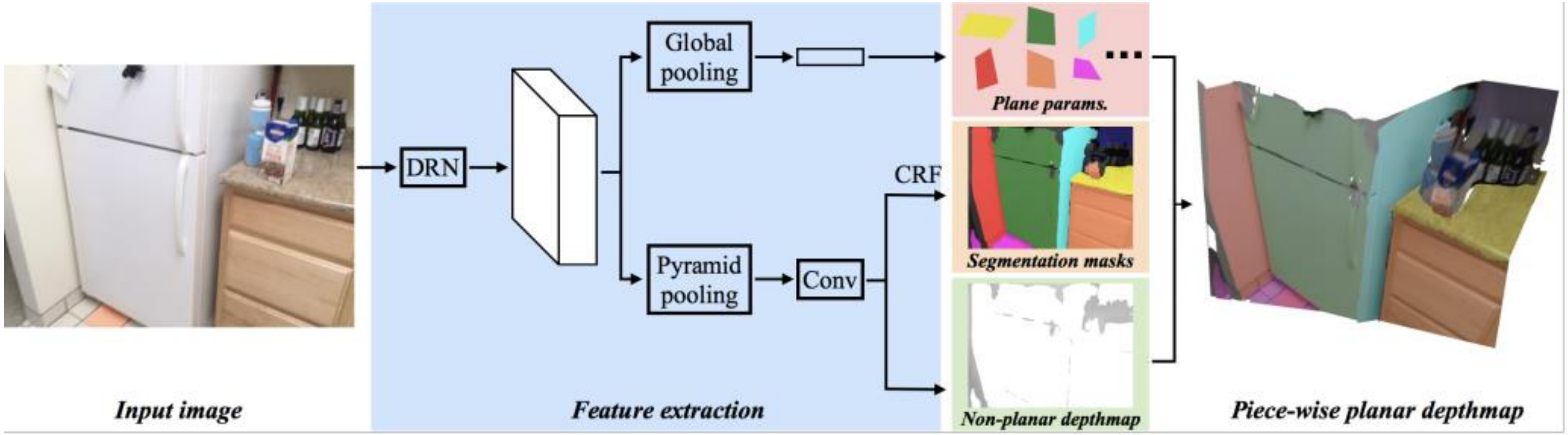
Single-Image Piece-wise Planar 3D Reconstruction via Associative Embedding

Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, **Shenghua Gao**

3D Plane reconstruction

3D plane = 2D segmentation map + 3D parameters

PlaneNet-A top-down approach

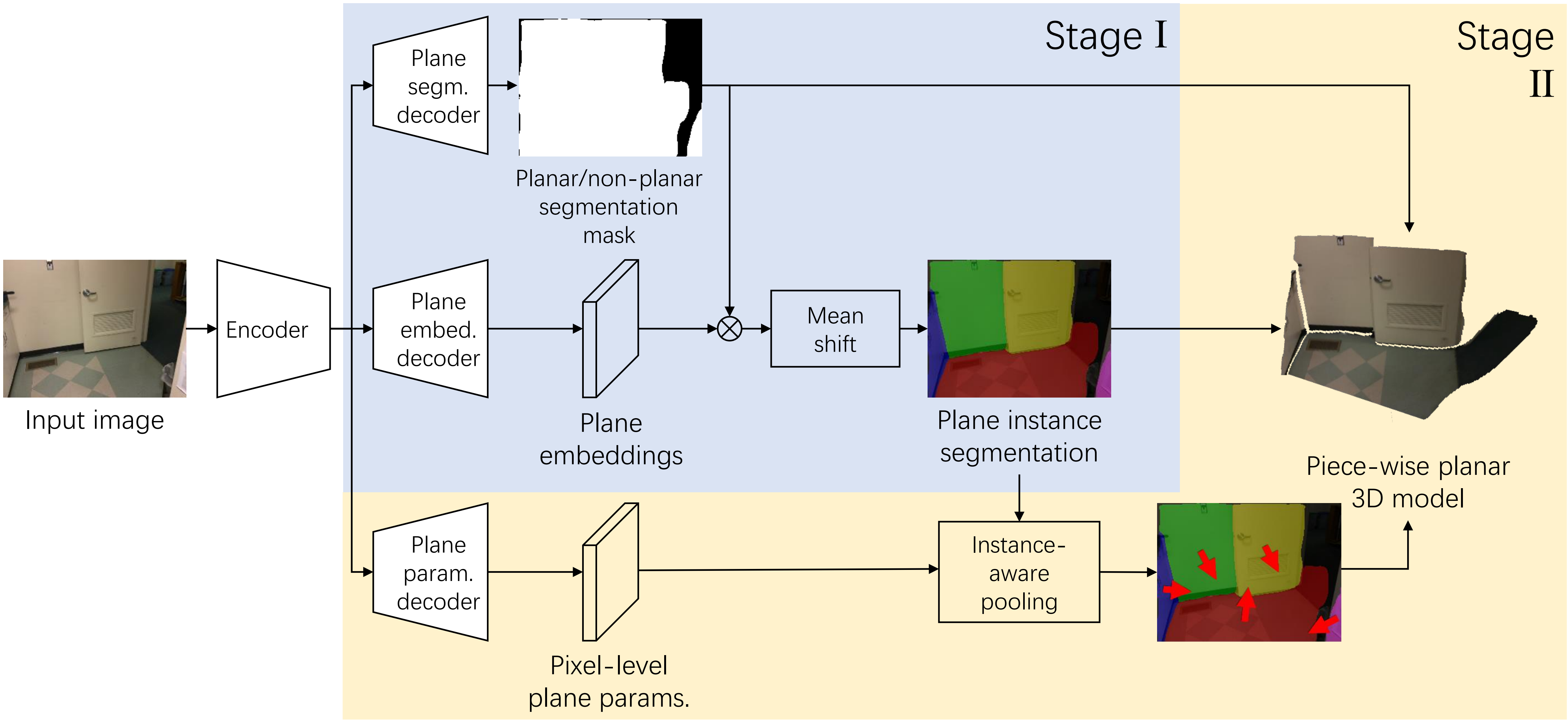


The number of planes is given;
The order of plane is given;
Impractical!

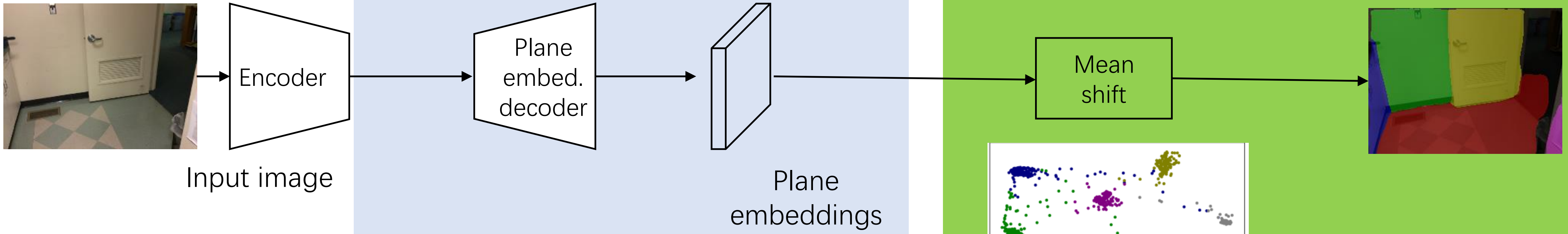
Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. **Planenet: Piece-wise planar reconstruction from a single rgb image**. In CVPR, pages 2579–2588, 2018.

Single-Image Piece-wise Planar 3D Reconstruction via Associative Embedding

- A bottom-up approach
- No prior about the number of planes and their orders in each scene
- Efficient – real time estimation


Stage I
Stage II
Plane segm. decoder
Planar/non-planar segmentation mask
Input image
Encoder
Plane embed. decoder
Plane embeddings
Mean shift
Plane instance segmentation
Piece-wise planar 3D model
Plane param. decoder
Pixel-level plane params.
Instance-aware pooling

Embedding and clustering



$$L_E = L_{pull} + L_{push}, \quad (2)$$

where

$$L_{pull} = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c} \sum_{i=1}^{N_c} \max\left(\| \mu_c - x_i \| - \delta_{\mathrm{v}}, 0 \right), \quad (3)$$

$$L_{push} = \frac{1}{C(C-1)} \sum_{c_A=1}^{C} \sum_{c_B=1}^{C} \max\left(\delta_{\mathrm{d}} - \| \mu_{c_A} - \mu_{c_B} \|, 0 \right)$$

Efficient meanshift

Shift a small number of anchors in embedding space

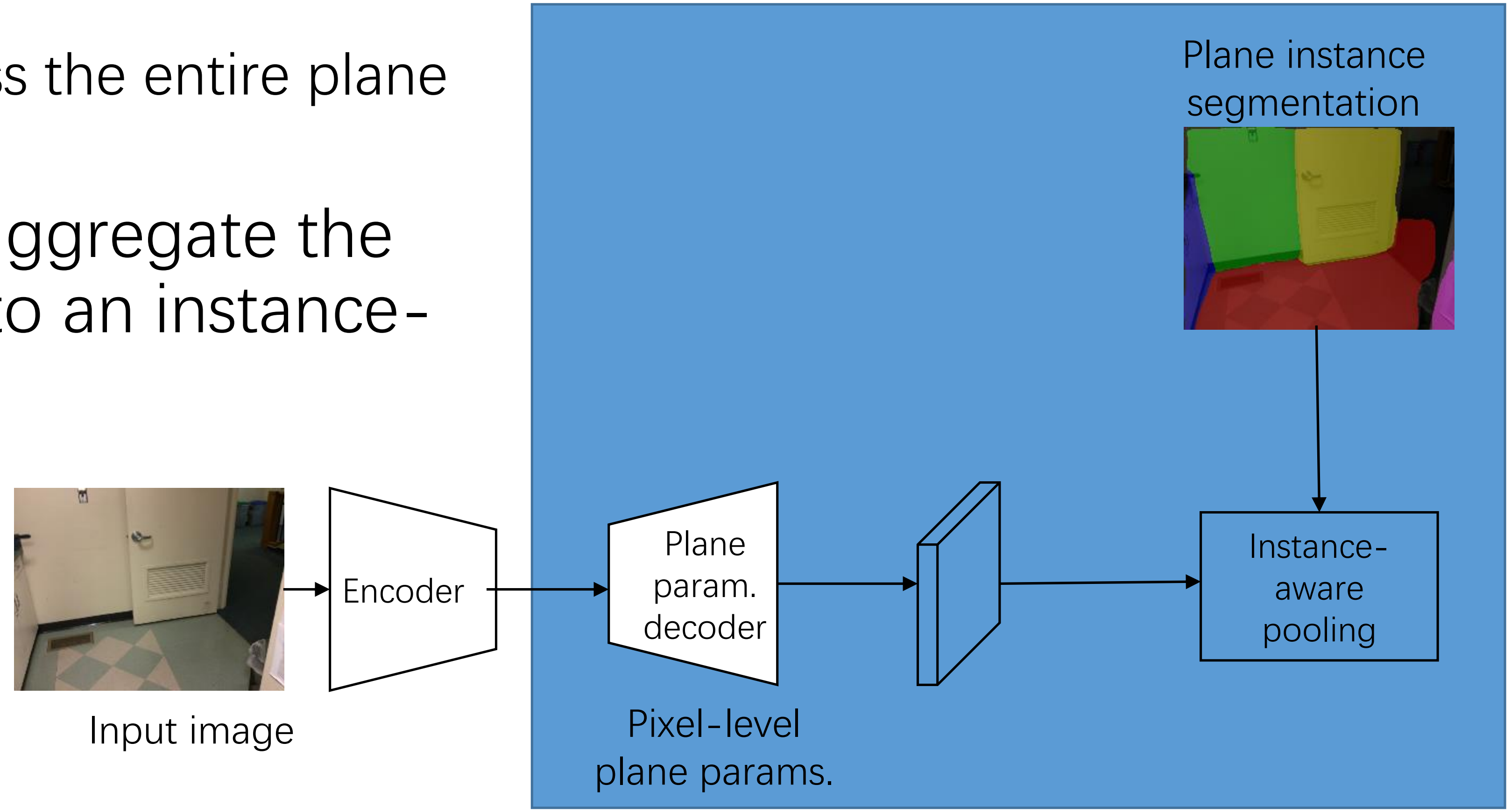
Assign each pixel to the nearest anchor.

Instance-aware pooling

- Pixel level parameter supervision is not sufficient
  - inconsistent outputs across the entire plane instance.
- Instance-aware pooling aggregate the pixel-level parameters into an instance-level parameter

$$n_j = \frac{1}{Z_j}\sum_{i=1}^{N} S_{ij}\cdot n_i,$$

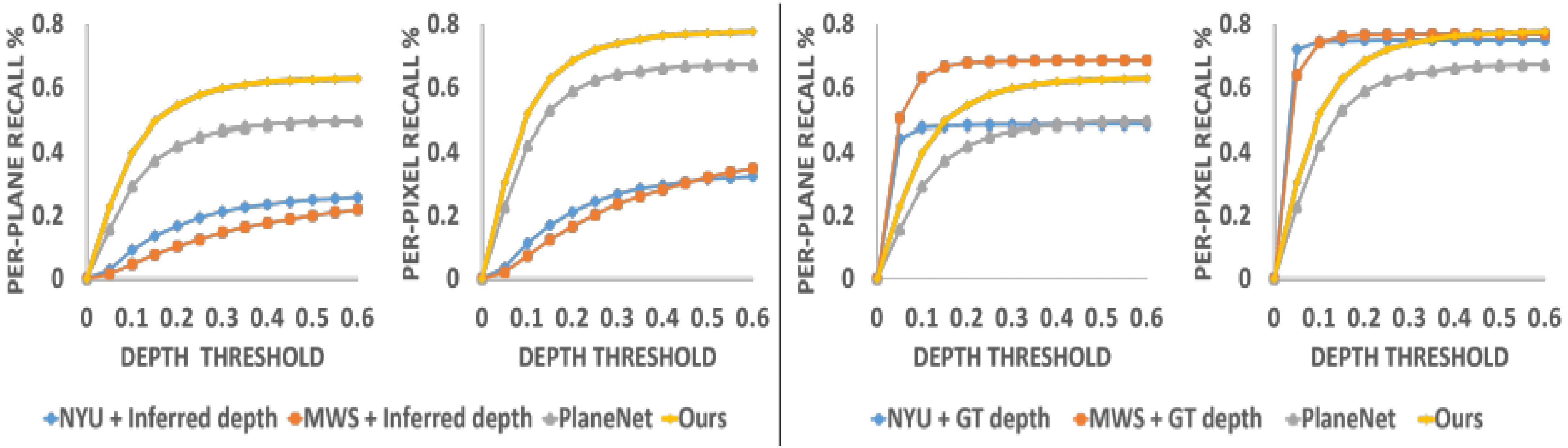
$$L_{IP} = \frac{1}{N\tilde{C}}\sum_{j=1}^{\tilde{C}}\sum_{i=1}^{N} S_{ij}\cdot \|n_j^T Q_i - 1\|,$$



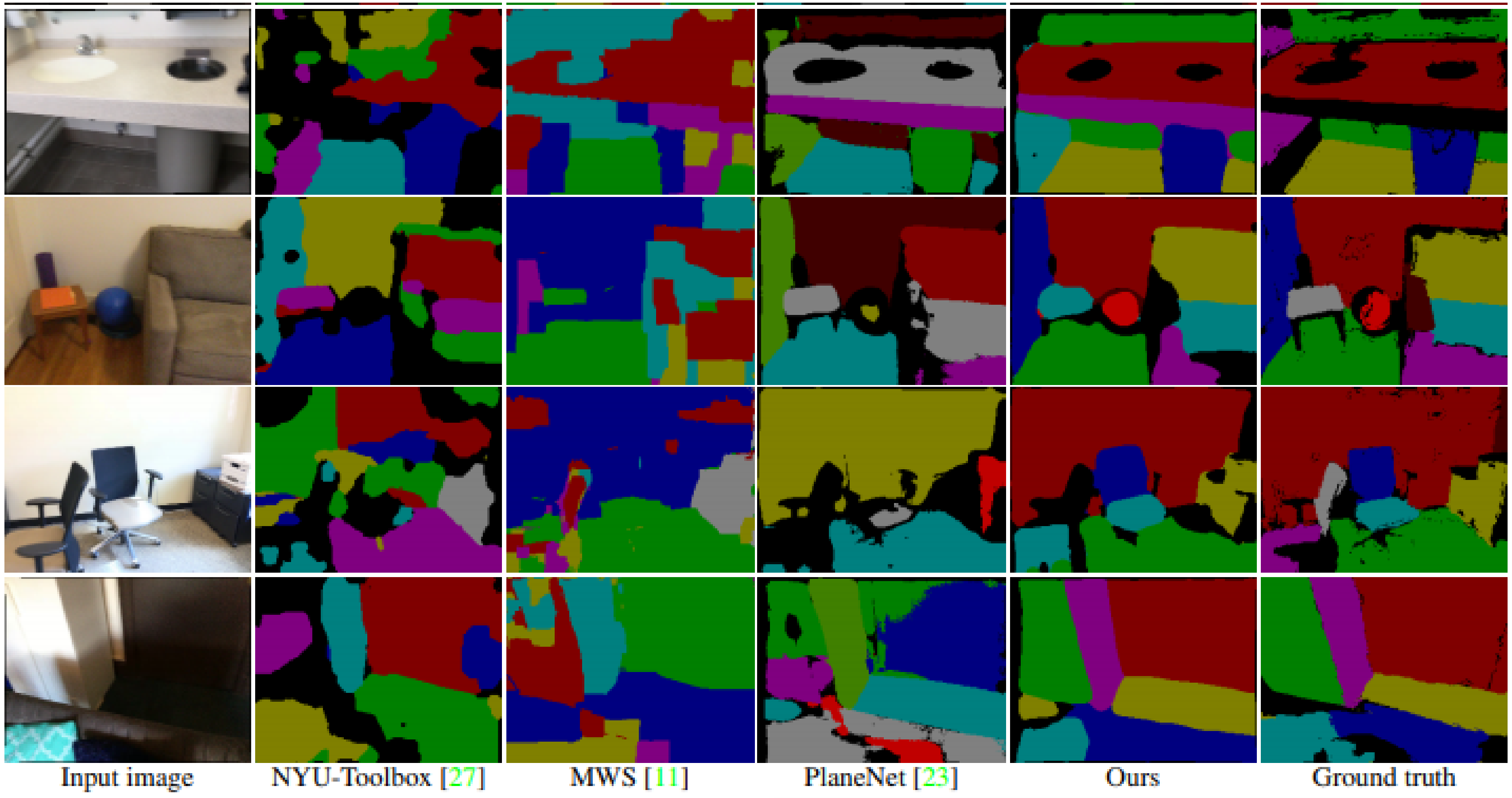
Evaluation

- Metrics
  - Plane recall- the percentage of correctly predicted ground truth planes;
  - Pixel recall - the percentage of pixels within the correctly predicted planes.
- A plane is considered correctly predicted if
  - The planes has more than 0.5 intersection-over-union (IOU) score;
  - the mean depth difference over the overlapping region is less than a threshold, which varies from 0.05m to 0.6m with an increment of 0.05m.

Evaluation with ScanNet



Plane instance segmentation results on the ScanNet dataset.



Evaluation with NYUv2

Table 2: Plane instance segmentation results on the NYUv2 test set.

Method	RI ↑	VI ↓	SC ↑
GT Depth + NYU-Toolbox [27]	0.875	1.284	0.544
PlaneNet [23]	0.723	1.932	0.404

RI: Rand index
VI: Variation of information;
SC: segmentation covering.

Table 3: Comparison of depth prediction accuracy on the NYUv2 test set.

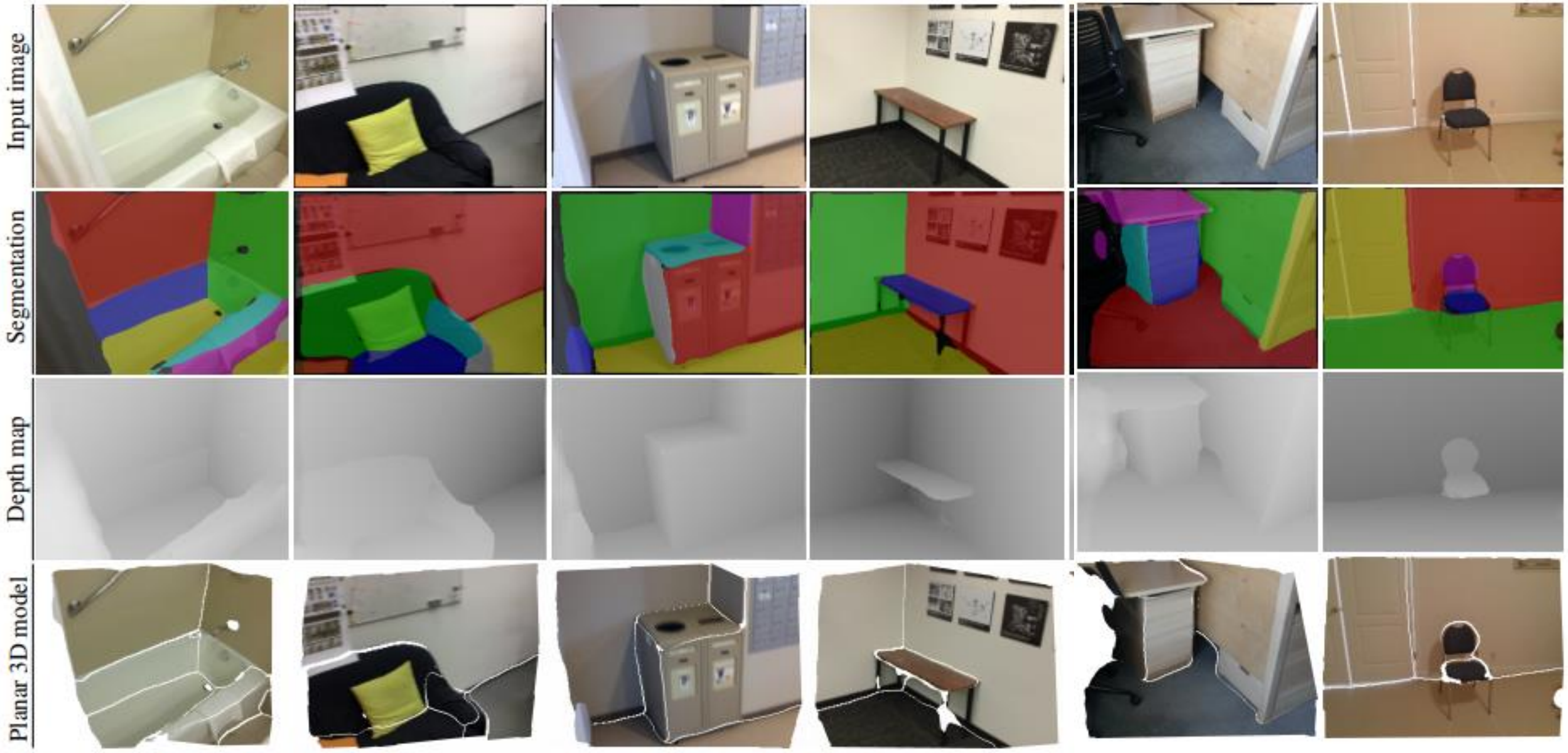
Method	Lower the better					Higher the better		
	Rel	Rel(sqr)	$\log_{10}$	RMSE_{iin}	$\text{RMSE}_{\log}$	1.25	1.25^2	1.25^3
Eigen-VGG [8]	0.158	0.121	0.067	0.639	0.215	77.1	95.0	98.8
SURGE [28]	0.156	0.118	0.067	0.643	0.214	76.8	95.1	98.9
FCRN [18]	0.152	0.119	0.072	0.581	0.207	75.6	93.9	98.4
PlaneNet [23]	0.142	0.107	0.060	0.514	0.179	81.2	95.7	98.9
Ours (depth-direct)	**0.134**	**0.099**	**0.057**	**0.503**	**0.172**	**82.7**	**96.3**	**99.0**
Ours	0.141	0.107	0.061	0.529	0.184	81.0	95.7	**99.0**

Speed

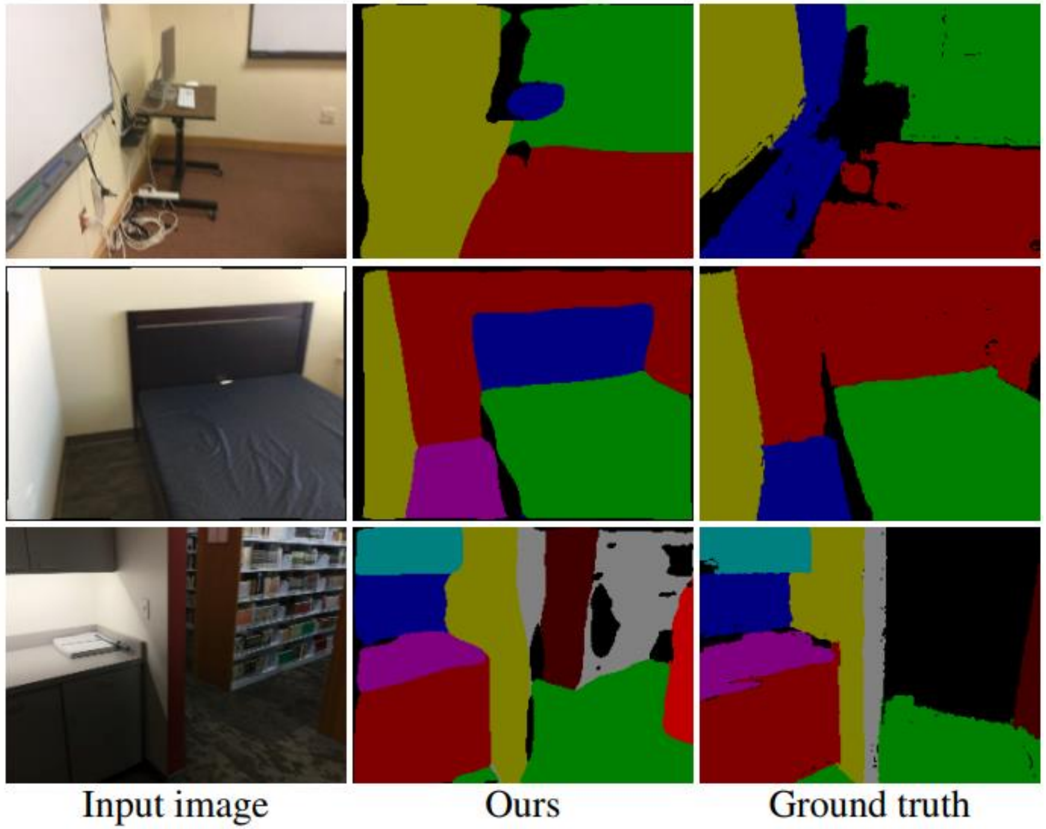
Table 1: Runtime comparison (* denotes CPU time).

Method	NYU-Toolbox [27]	MWS [11]	PlaneNet [23]	Ours
FPS	0.14*	0.05*	1.35	**32.26**

Results Visualization

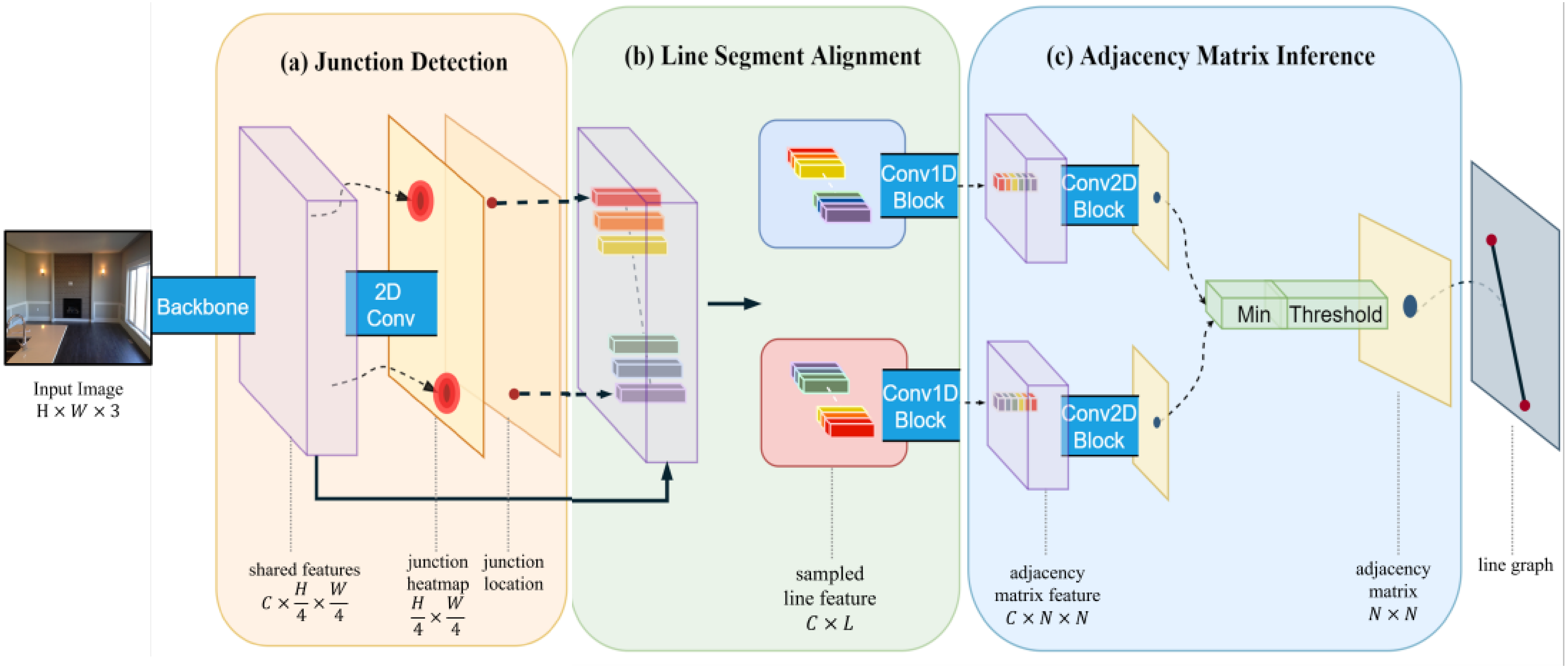


Failure cases



PPGNet: Learning Point-Pair Graph for Line Segment Detection

Ziheng Zhang, Zhengxin Li, Ning Bi, **Shenghua Gao**



(a) Junction Detection
(b) Line Segment Alignment
(c) Adjacency Matrix Inference
Backbone
2D Conv
Conv1D Block
Conv1D Block
Conv2D Block
Conv2D Block
Min
Threshold
Input Image
H × W × 3
shared features
C × H/4 × W/4
junction heatmap
H/4 × W/4
junction location
sampled line feature
C × L
adjacency matrix feature
C × N × N
adjacency matrix
N × N
line graph

Summary

- Structure learning would contribute to 3D reconstruction
- Bottom-up based solution for piece-wise planar 3D reconstruction
- Application of structure information for robotics, image/video editing, scene understanding, VR, AR, etc..

Thank you

gaoshh@shanghaitech.edu.cn