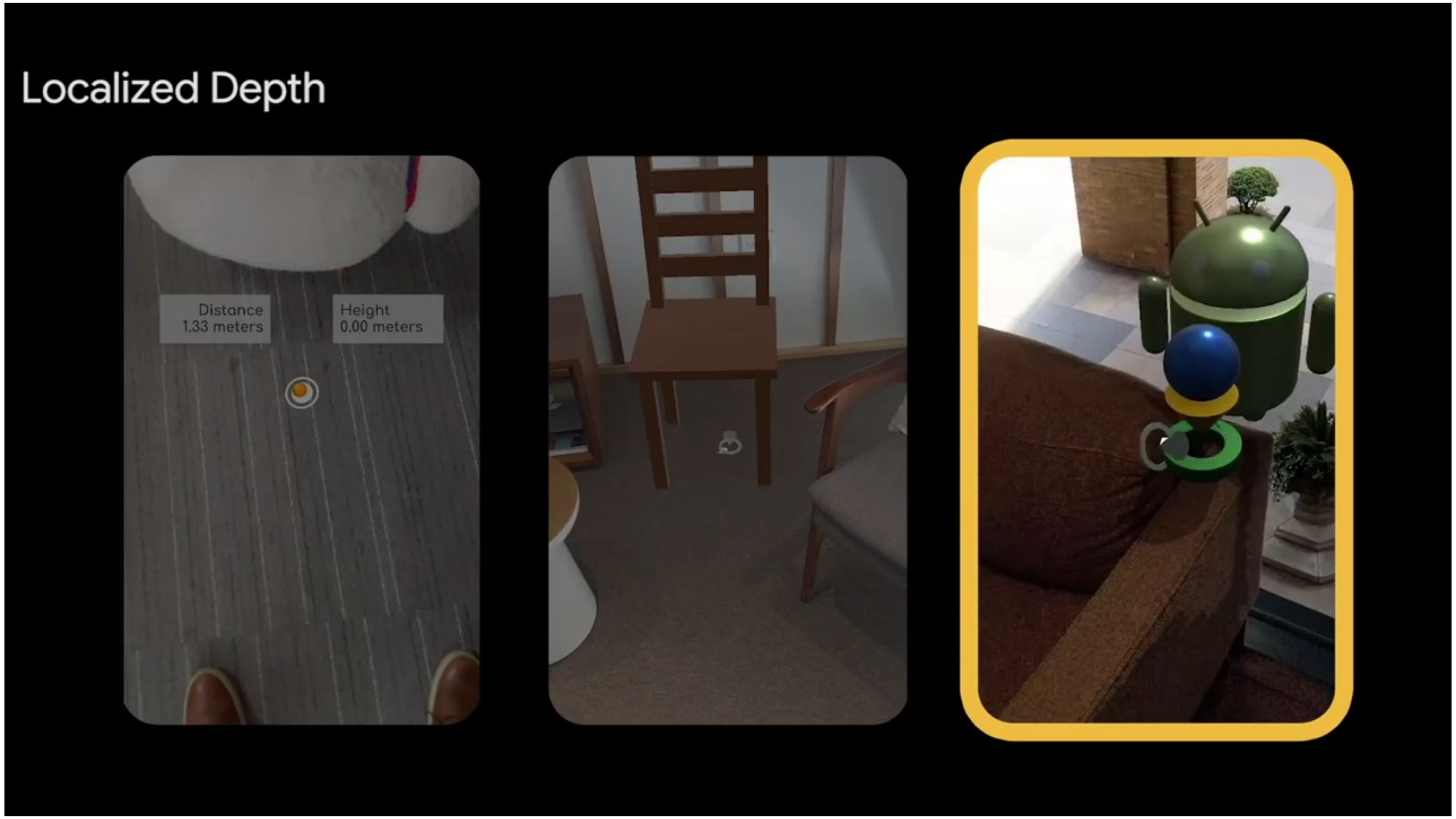
PlaneMVS: 3D Plane Reconstruction from Multi-view Stereo

Jiachen Liu, Pan Ji, Nitin Bansal, Changjiang Cai, Qingan Yan, Xiaolei Huang, Yi Xu

CVPR 2022

Background

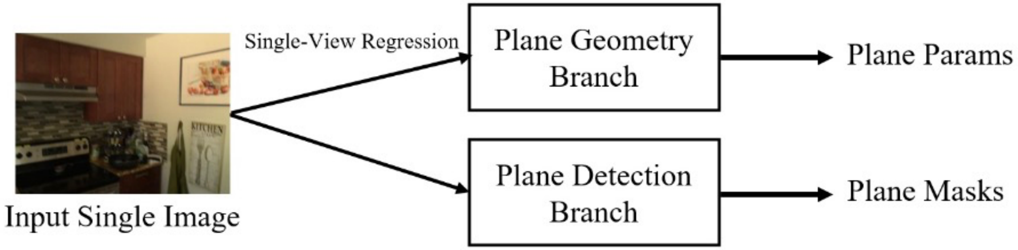
- 3D piece-wise plane reconstruction would be a key for AR applications.



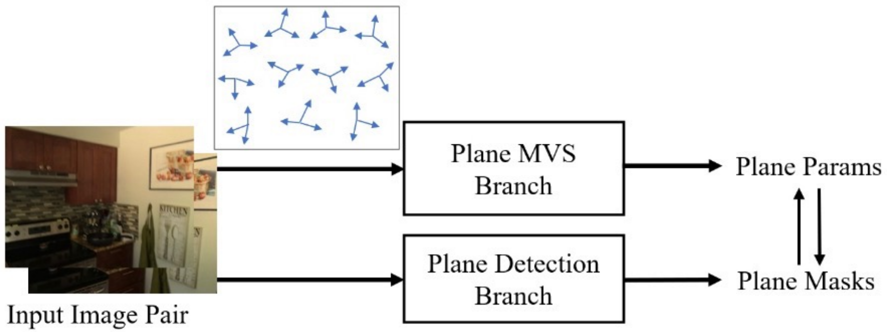
A video demo from Google DepthLab[1]

[1] Du, Ruofei, et al. "DepthLab: Real-Time 3D Interaction With Depth Maps for Mobile Augmented Reality." *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 2020

Motivation: Single-view v.s. Multi-view Reconstruction



(a) Single-view plane reconstruction framework



(c) The proposed multi-view plane reconstruction framework

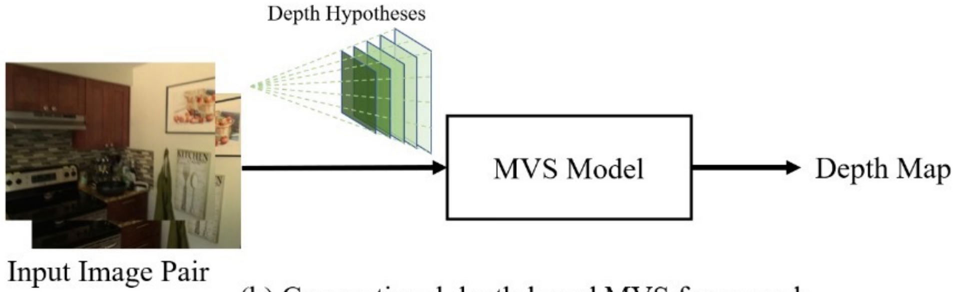
Single-view methods:

- Rely on single-view regression on geometry
- Suffer from depth scale ambiguity
- Perform well on plane detection

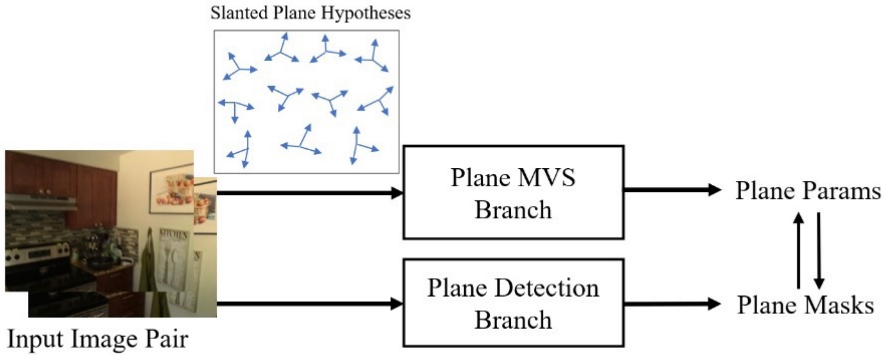
Multi-view stereo methods:

- Reconstruct planes from multi-view geometry
- Resolve depth scale ambiguity
- Inherit single-view plane detection

Motivation: Depth Hypothesis v.s. Slanted Plane Hypothesis



(b) Conventional depth-based MVS framework



(c) The proposed multi-view plane reconstruction framework

Depth-based MVS:

- Assume frontal-parallel plane hypothesis
- Output a depth map

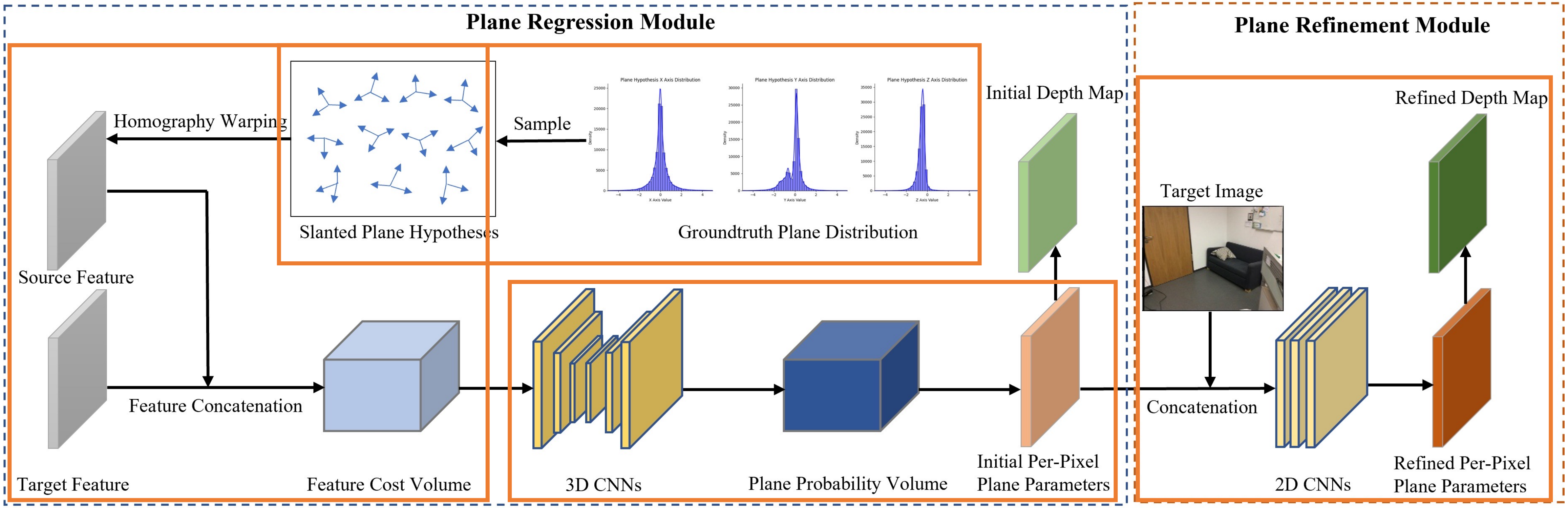
Slanted-plane-based MVS:

- Use slanted planes as hypothesis
- Output a 3-channel map of plane parameters

Based on plane homography

$$H_i(\boldsymbol{n_i}, e_i) \sim \boldsymbol{K}(\boldsymbol{R} - \frac{\boldsymbol{t}\boldsymbol{n_i}^T}{e_i})\boldsymbol{K}^{-1}$$

Our Proposed Framework



Our proposed PlaneMVS framework

Instance-level Plane Reconstruction

- Plane instance-aware soft pooling loss:

$$\boldsymbol{p_t} = \frac{\sum_{i=1}^{N} \sigma_i \cdot \boldsymbol{p_i}}{\sum_{i=1}^{N} \sigma_i} \qquad \mathbf{D}_i = -\frac{\mathbb{1}_i}{\boldsymbol{p_t}^T \boldsymbol{K}^{-1} \mathbf{x}_i} \qquad L_{sp} = \|\mathbf{D} - \mathbf{D}^*\|_1$$

- Final losses with learnable uncertainty[1]:

$$L = \sum_{i}^{N_D} \omega_{D_i} L_{D_i} + \sum_{j}^{N_M} \omega_{M_j} L_{M_j} + \omega_{sp} L_{sp}$$

- Apply convex upsampling[2]:

Learn an 8x8x3x3 grid for each pixel $\longrightarrow$ Weighted combination over neighbors

[1] Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).
[2] Teed, Zachary, and Jia Deng. "Raft: Recurrent all-pairs field transforms for optical flow." *European conference on computer vision*. Springer, Cham, 2020.

Experimental Results

ScanNet

Method	Depth Metrics							Detection Metrics					
	AbsRel↓	SqRel↓	RMSE↓	RMSE_log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑	$AP^{0.2m}$↑	$AP^{0.4m}$↑	$AP^{0.6m}$↑	$AP^{0.9m}$↑	AP↑	mAP↑
PlaneRCNN [31]	0.164	0.068	0.284	0.186	0.780	0.953	0.989	0.310	0.475	0.526	0.546	0.554	0.452
MVSNet [58]	0.105	0.040	0.232	0.145	0.882	0.972	0.993	-	-	-	-	-	-
DPSNet [19]	0.100	0.035	0.215	0.135	0.896	0.977	0.994	-	-	-	-	-	-
NAS [26]	0.098	0.035	0.213	0.134	0.905	0.979	0.994	-	-	-	-	-	-
ESTDepth [33]	0.113	0.037	0.219	0.147	0.879	0.976	0.995	-	-	-	-	-	-
PlaneMVS-pixel (Ours)	0.091	0.029	0.194	0.120	0.920	0.987	0.997	0.448	0.535	0.556	0.560	**0.564**	**0.466**
PlaneMVS-final (Ours)	**0.088**	**0.026**	**0.186**	**0.116**	**0.926**	**0.988**	**0.998**	**0.456**	**0.540**	**0.559**	**0.562**	**0.564**	**0.466**

Generalizability (trained on ScanNet)

7-Scenes

Method	AbsRel↓	$\delta < 1.25$↑
PlaneRCNN [31]	0.221	0.640
MVSNet [58]	0.162	0.766
DPSNet [19]	0.159	0.788
NAS [26]	0.154	0.784
ESTDepth [34]	**0.153**	0.786
Ours	0.158	**0.793**
Ours-FT	0.113	0.890

TUM-RGBD

Method	AbsRel↓	SqRel↓	$\delta < 1.25$↑
PlaneRCNN [31]	0.243	0.105	0.655
Ours	0.143	0.07	0.795
Ours-FT	**0.120**	**0.054**	**0.851**

Ablation Study

Quantitative

Method	Depth Metrics							Detection Metrics				
	AbsRel↓	SqRel↓	RMSE↓	RMSE_log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑	$AP^{0.2m}$↑	$AP^{0.4m}$↑	$AP^{0.6m}$↑	$AP^{0.9m}$↑	AP↑
Baseline	0.170	0.074	0.305	0.200	0.746	0.944	0.990	0.288	0.458	0.519	0.545	0.551
+ Soft-pooling loss	0.119	0.042	0.234	0.148	0.871	0.979	0.995	0.380	0.520	0.549	0.557	0.561
+ Loss term uncertainty	0.089	0.027	0.190	0.119	0.922	0.987	0.997	0.449	0.535	0.556	0.560	0.562
+ Convex upsampling	**0.088**	**0.026**	**0.186**	**0.116**	**0.926**	**0.988**	**0.998**	**0.456**	**0.540**	**0.559**	**0.562**	**0.564**

Table 3. Ablation study on the components of our proposed method.

Method	AbsRel↓	SqRel↓	$\delta < 1.25$↑	$AP^{0.2m}$↑	AP↑
Fronto-MVS	0.094	0.033	0.917	0.433	0.548
Ours	**0.088**	**0.026**	**0.926**	**0.456**	**0.564**

Table 4. Ablation study: slanted v.s. fronto-parallel plane.

Qualitative

Image w/o soft-pooling loss w/ soft-pooling loss

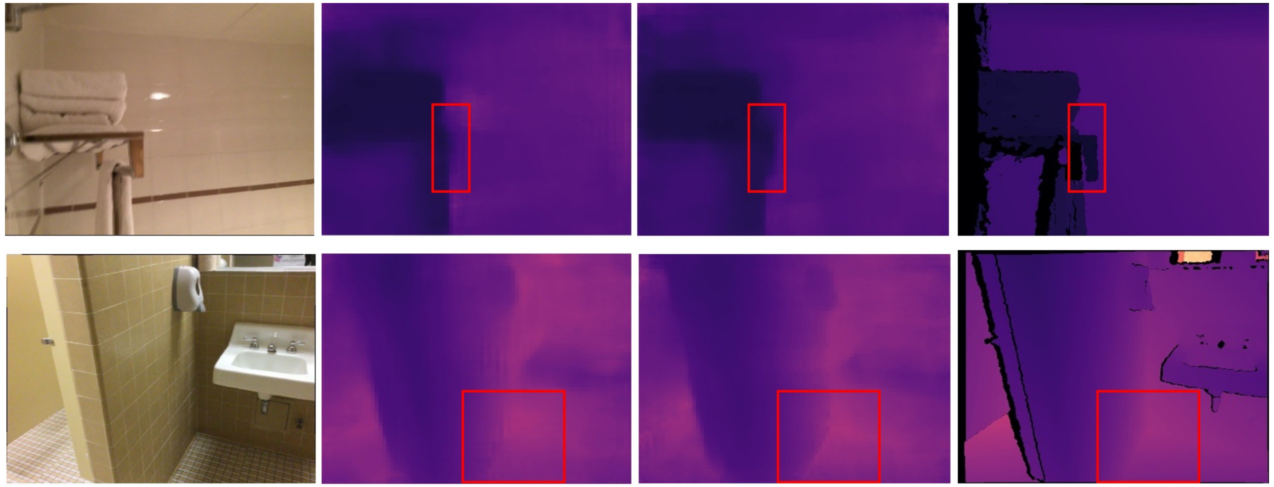
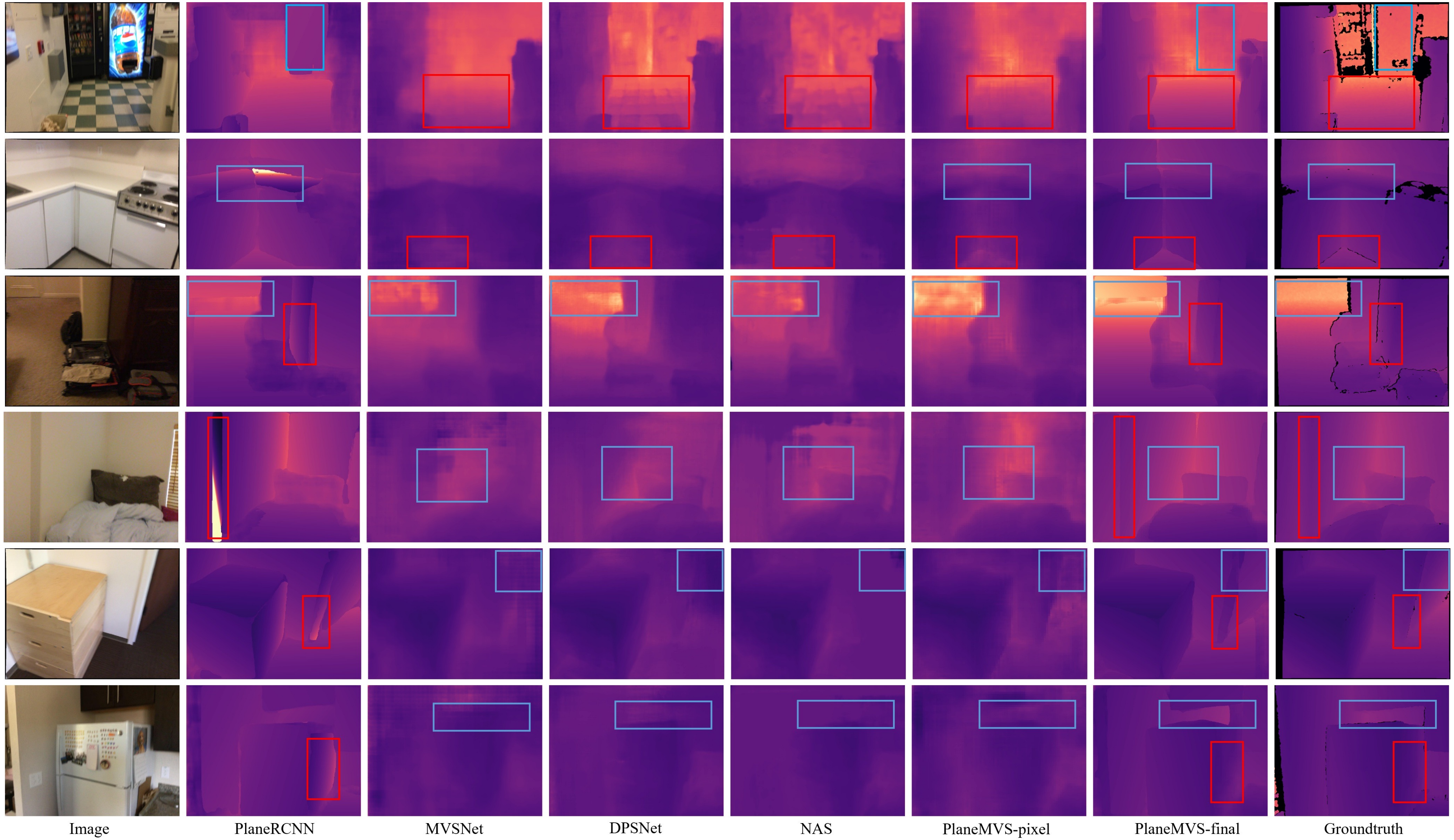
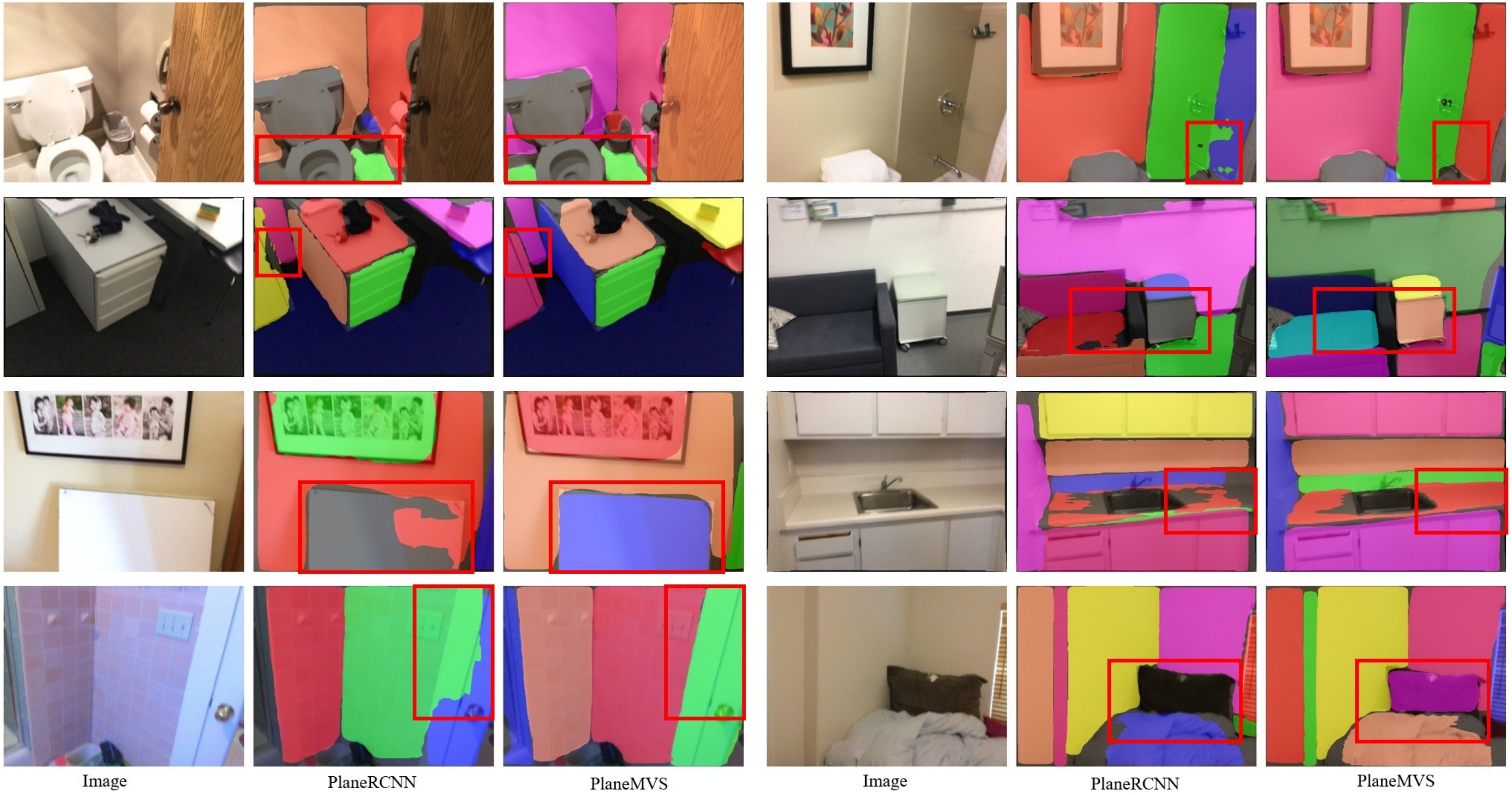


Image Bilinear Upsampling Convex Upsampling Groundtruth

Qualitative Results – ScanNet Planar Depth



Qualitative Results – ScanNet Planar Segmentation



Qualitative Results – 7Scenes

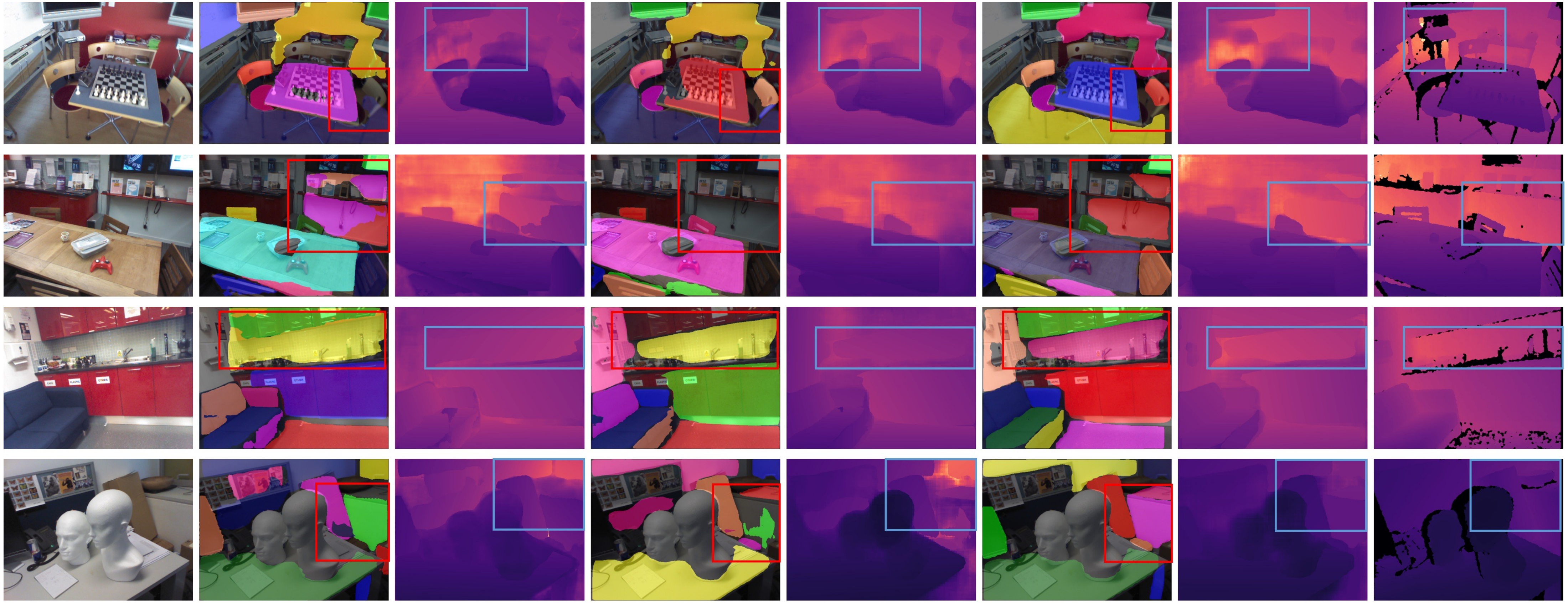


Image	PlaneRCNN-det	PlaneRCNN-depth	PlaneMVS-det	PlaneMVS-depth	PlaneMVS-FT-det	PlaneMVS-FT-depth	GT-depth

Qualitative Results – TUM-RGBD

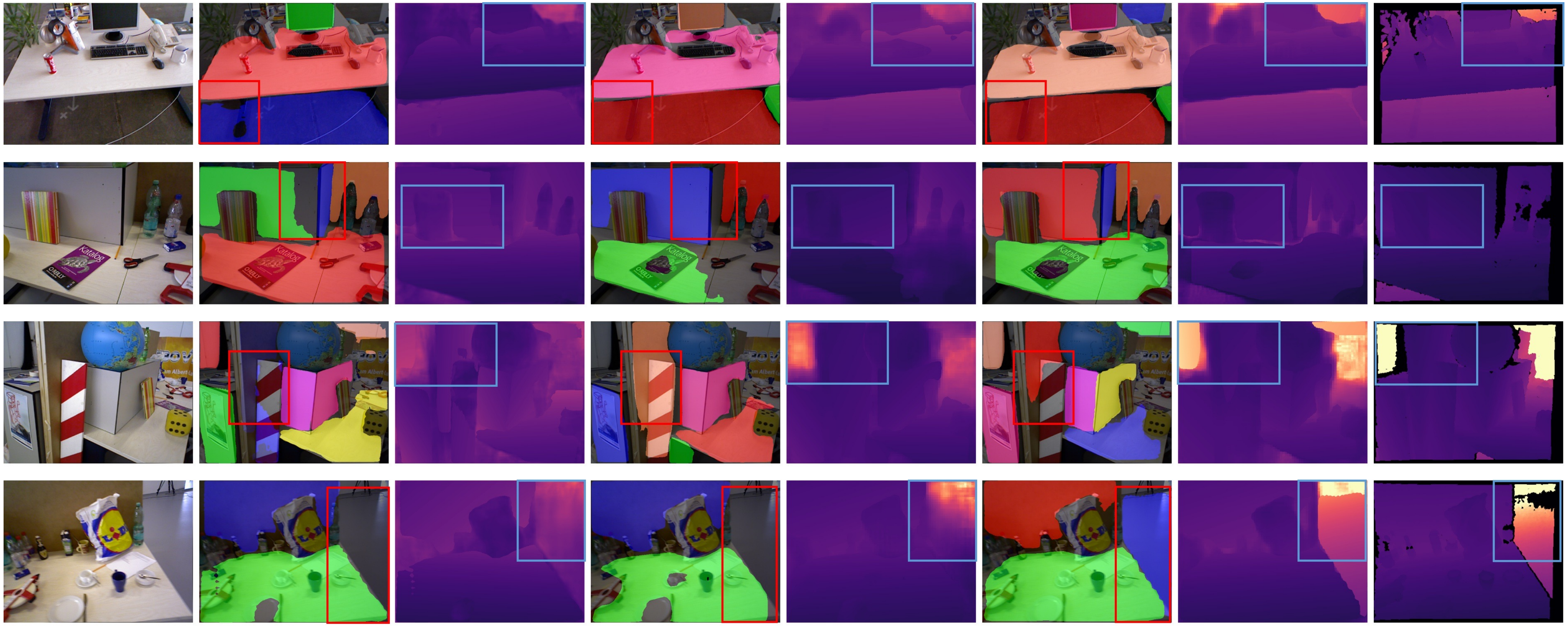


Image PlaneRCNN-det PlaneRCNN-depth PlaneMVS-det PlaneMVS-depth PlaneMVS-FT-det PlaneMVS-FT-depth GT-depth

Conclusion and Future Work

- We present an end-to-end MVS framework to reconstruct 3D planes on multiple posed images.
- Our core contributions lie in the proposed **slanted plane hypotheses** and the **soft pooling loss** to associate the plane detection and the plane MVS modules.
- In the future, we would like to explore the possibility to extend the proposed framework to videos, for temporal consistent 3D plane reconstruction.

Thank you!