DeepViewpoints: Hyper-Rays with Harnomics Encoding for 6DoF Viewpoint Learning (Supplementary Material)

Zhixiang Min Juan Carlos Dibene Enrique Dunn
Stevens Institute of Technology

1. Additional Experiments.

1.1. Ablation Study on Semantic Labels.

In Table.1, we study the effect of aggregating GT semantic labels into our input point cloud datum, as an input feature to PointNet. We found including semantic labels only marginally improves the AP when learning with all data, but their inclusion is crucial in learning semantic purposes.

Split	w/ Semantic Label		w/o Semantic Label	
	Loc. AP	View AP	Loc. AP	View AP
Bed	68.50	55.94	68.91	38.08
Toilet	92.78	83.57	70.20	44.47
All	86.37	56.70	84.67	55.08

Table 1: **Ablation Study on Semantic Labels.** We study the effect of using GT semantic labels as point cloud input features.

1.2. KITTI Dataset.

We additionally test our method on the KITTI odometry dataset [2], where the environment and viewpoint pose have very different distribution than the indoor ScanNet dataset. We train on sequences 00, 02 and test on sequence 05. We fuse each lidar scan with adjacent scans to fill the holes and down sample the points using voxel filter as our point cloud input. Note although KITTI has less variation at the pitch and roll axes, we still treat it as 6DoF viewpoint learning for generality. Due to the lack of semantic label and mesh, we only use [4] with depth and point saliency statistics as baseline. The GT H.R.P. baseline is also attached as reference for the environment scale.

KITTI Results. In Table.2, we report location and viewpoint selection precision similar to the ScanNet setup. For the baseline method, we additionally restrict its motion model to 4DoF using GT average height, pitch and roll. Both methods recover most locations and viewpoints on the KITTI dataset, while our method shows higher precision. The viewpoint selection precision is usually lower since it is difficult to determine the driving direction (i.e. forward/backward). In Fig.1, we visualize the selected viewpoints using lines indicating oriented camera pose by composing all scans together. The resulting viewpoints nicely follow the driving direction of the road.

	Location (<2m)			View (<2m & <30°)		
Method	Precision	Recall	AP	Precision	Recall	AP
GT H.R.P.	-	-	4.70	-	-	0.88
Adrian *et al.* [4]	37.15	96.99	38.12	17.73	84.34	20.04
+ 4DoF	69.30	95.24	71.23	35.44	89.27	37.29
Ours	**76.42**	**98.08**	**78.52**	**43.08**	**96.13**	**45.74**

Table 2: **Viewpoint selection precision on KITTI.**

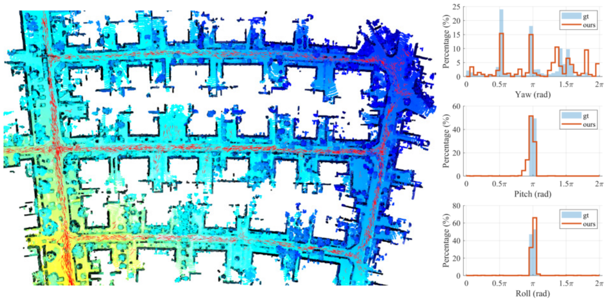


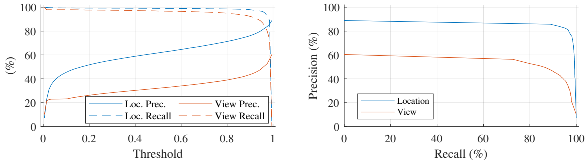
Figure 1: **Results on KITTI.** We compose results on multiple lidar segments to a street BEV. The selected viewpoints are visualized as red lines with same orientation. The map color indicates the changes in y-axis (height).

1.3. Additional results on ScanNet.

In Fig.4,5,6, we show additional results where the three rows from top to bottom are GT, ours and Kyle *et al.* [3]. The visualization of location score is same to the main paper. The right grid of viewpoint images are randomly selected viewpoint images of each method at the optimal threshold. The viewpoint images selected by our method exhibit good diversity and realism.

1.4. PR curve.

In Fig.2, we report the PR curve and performances over different thresholds on ScanNet. We evenly sample 100 thresholds and estimate precision/recall for each scene. The



(a) Performance over Thresholds.

(b) PR Curve

Figure 2: **Detailed Performance on ScanNet.**

final reported precision/recall is the average over all scenes.

2. Supportive Details.

2.1. Hyper-parameters.

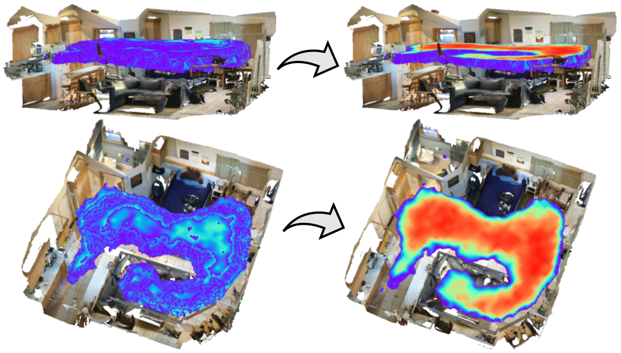
In Table.3, we detail the default value of our hyper-parameters. The hyper-parameters are commonly applied to both ScanNet and KITTI datasets except for the maximum distance γ_{max}.

Symbol	Value	Eq.Ref	Description
P	4096	Eq.(6)	number of input points
D	128	Eq.(7,8,9)	descriptor dimension
H_1	8	Eq.(9)	length field (fourier series) degree
H_2	2	Eq.(7)	optic-ray direction field ($\mathbb{S}^2$) degree
H_3	2	Eq.(8)	hyper-ray direction field ($\mathbb{S}^3$) degree
V_2	64	Eq.(12)	Voronoi 2-sphere resolution
V_3	512	Eq.(18)	Voronoi 3-sphere resolution
λ	0.5	Eq.(9)	view cropping visibility
η_r	20°	Eq.(24)	virtual roll-axis FoV
γ_{max}	10 meters	Eq.(9)	maximum distance for ScanNet
γ_{max}	80 meters	Eq.(9)	maximum distance for KITTI

Table 3: **Hyper-parameter Table.**

2.2. Baseline implementation details.

We implement Kyle *et al*. [3] and Adrian *et al*. [4] as our baselines. Our implementation only replicates their method for scoring a viewpoint, and integrates their scoring model under the same viewpoint sampling and selection model as our method. Since both baselines require viewpoint images as input, we first render and store all required viewpoints images on disk. In order to have same sampling density as our method (0.2m grid with 4096 rotations per location), the number of viewpoint hypotheses easily exceeds 100M images per scene, which is not a reasonable amount to render and store. To avoid this, we cache panoramas at every location and crop the panoramas into perspective images to create viewpoint images during inference time. The rendering takes a week using Open3D [5] on a 32 core CPU, while the inference of both methods on ScanNet takes 1-2 days with pytorch GPU acceleration. The images we render for both baseline methods are GT semantic label images, since both methods only model the semantics and discard RGB color. For Kyle *et al*. [3], we build a $48 \times 64 \times 32$ (height, width, depth) histogram for each semantic class in ScanNet (40 for total) during the training. During the inference, we score the viewpoint image by integrating each pixel's corresponding bin value. For Adrian *et al*. [4], we build histograms for depth statistics, semantic statistics and mesh saliency, where we rate viewpoint images using histograms during the inference. Due to the different semantic statistics for each scene, we normalize the score of both baselines for each scene separately. We sum up all viewpoints of the same location as their location score. Since Adrian *et al*. [4] does not analyze the structure of pixels on the image, it cannot distinguish the roll of a viewpoint. Hence, the "+fix gravity" in our main paper Table.1 setup manually fixes this issue by assuming all viewpoints align to a known gravity direction. Kyle *et al*. [3] models the image pixel structure using a pixel-wise histograms, hence it roughly captures correct roll angle as visualized in main paper Fig.5. However, the histogram shows a certain amount of outliers with incorrect roll angle, indicating the difficulty to recover the correct camera pose from image capture.



Heat-Volumes

After Clipping

Figure 3: **Heat-Volume and Clipping.** We clip the heat-volume to the height of maximal accumulated score for better visualization.

2.3. Regarding the heat-volume visualization.

It is worth noting that the heat-map in our location score visualizations are all 3D heat-volumes. The heat-volumes are built by linearly interpolating the score of each location. However, directly rendering the solid volumes will prevent its internal structure from being visualized as shown in Fig.3. Hence, we clip the top of the volume to reveal its interior, where the volume is clipped to the height of maximal accumulated score.

2.4. Harmonics Polynomial Lookup Table.

For convenience, we attach both 2-/3-sphere harmonics polynomials up to 2 degrees in Table.4. They are computed using the software provided by [1]. The tables take unit vector input $[x_1, x_2, x_3]$ for 2-sphere and $[x_1, x_2, x_3, x_4]$ for 3-sphere in Cartesian coordinate.

l	m	expression
0	0	1
1	-1	$\sqrt{3}x_2$
	0	$\sqrt{3}x_3$
	1	$\sqrt{3}x_1$
2	-2	$\sqrt{15}x_1x_2$
	-1	$\sqrt{15}x_2x_3$
	0	$\frac{\sqrt{5}}{2}(1-3x_2^2)$
	1	$\sqrt{15}x_1x_3$
	2	$\frac{\sqrt{15}}{2}(1-x_2^2-2x_3^2)$

(a) 2-sphere

k	l	m	expression
0	0	0	1
1	0	0	$2x_2$
	1	-1	$2x_3$
		0	$2x_4$
		1	$2x_1$
2	0	0	$1-4x_2^2$
	1	-1	$2\sqrt{6}x_2x_3$
		0	$\sqrt{2}(1-x_2^2-3x_3^2)$
		1	$2\sqrt{6}x_2x_4$
	2	-2	$2\sqrt{6}x_3x_4$
		-1	$2\sqrt{6}x_1x_2$
		0	$\sqrt{6}(1-x_2^2-x_3^2-2x_4^2)$
		1	$2\sqrt{6}x_1x_3$
		2	$2\sqrt{6}x_1x_4$

(b) 3-sphere

Table 4: **Spherical Harmonics Lookup Tables.**

References

[1] Sheldon Axler, Paul Bourdon, and Ramey Wade. *Harmonic function theory*, volume 137. Springer Science & Business Media, 2013. 2

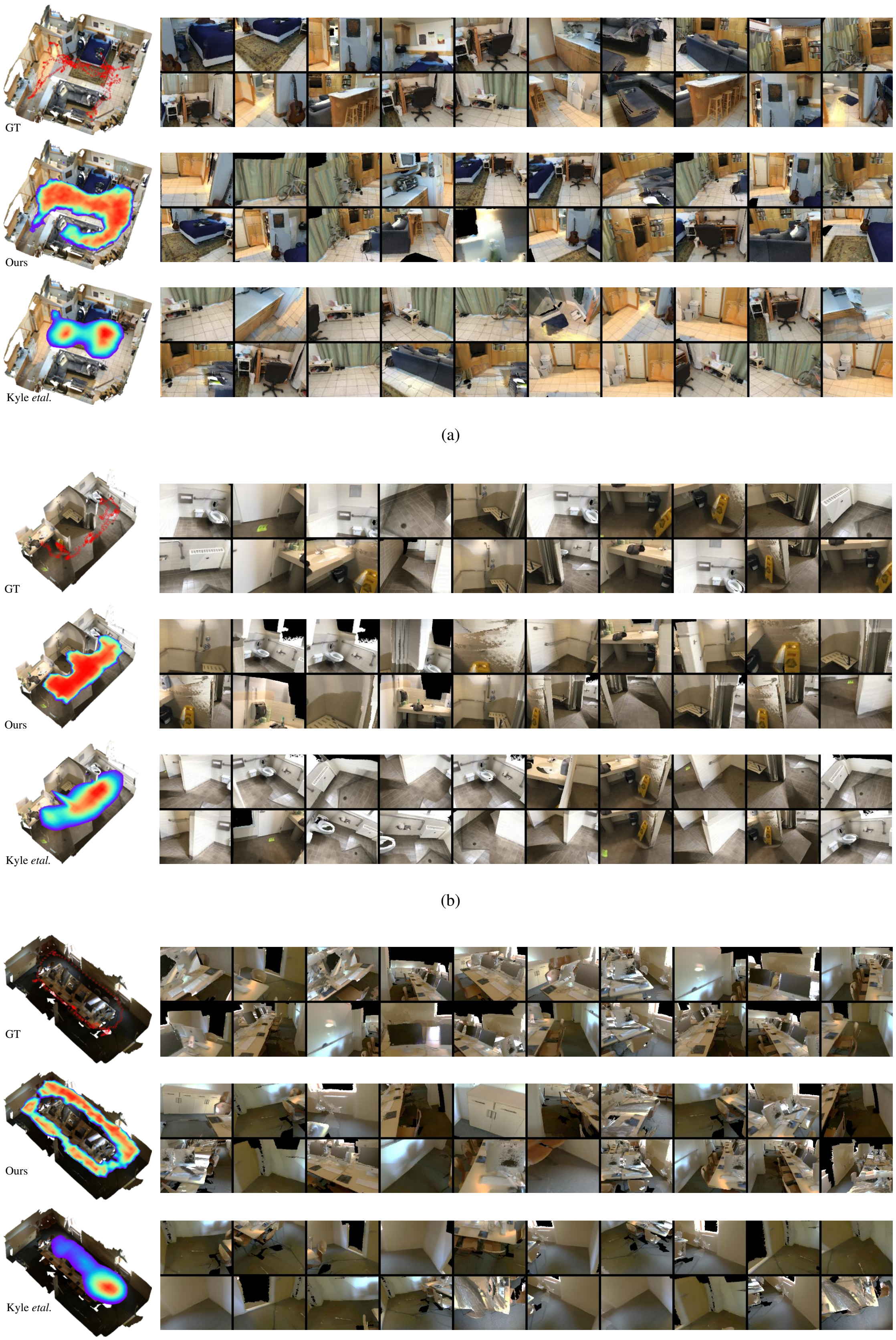
[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[3] Kyle Genova, Manolis Savva, Angel X Chang, and Thomas Funkhouser. Learning where to look: Data-driven viewpoint set selection for 3d scenes. *arXiv preprint arXiv:1704.02393*, 2017. 1, 2, 4, 5, 6

[4] Adrian Secord, Jingwan Lu, Adam Finkelstein, Manish Singh, and Andrew Nealen. Perceptual models of viewpoint preference. *ACM Transactions on Graphics (TOG)*, 30(5):1–12, 2011. 1, 2

[5] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 2

3 Pages Left ↓



(a)

(b)

(c)

Figure 4: **Results on ScanNet (Group 1).** The left 3D figures visualize the location score using 3D heat-volumes clipped to the height of maximal accumulated score. The right pictures are randomly picked viewpoints at an optimal threshold gives best F1 score. Each group of rendered viewpoint pictures corresponds to a different method. From top to bottom is GT, ours and Kyle *et al*. [3].

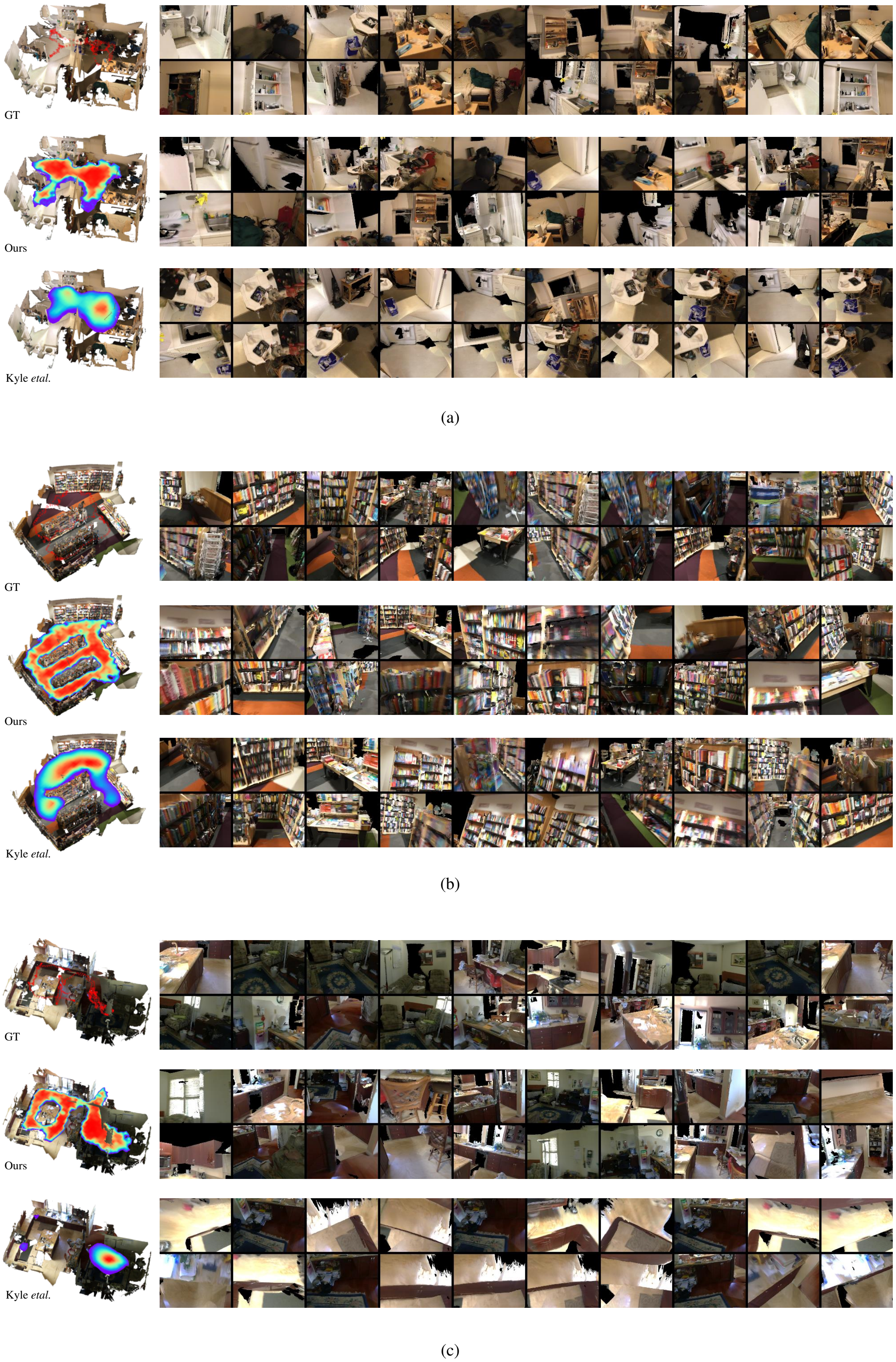
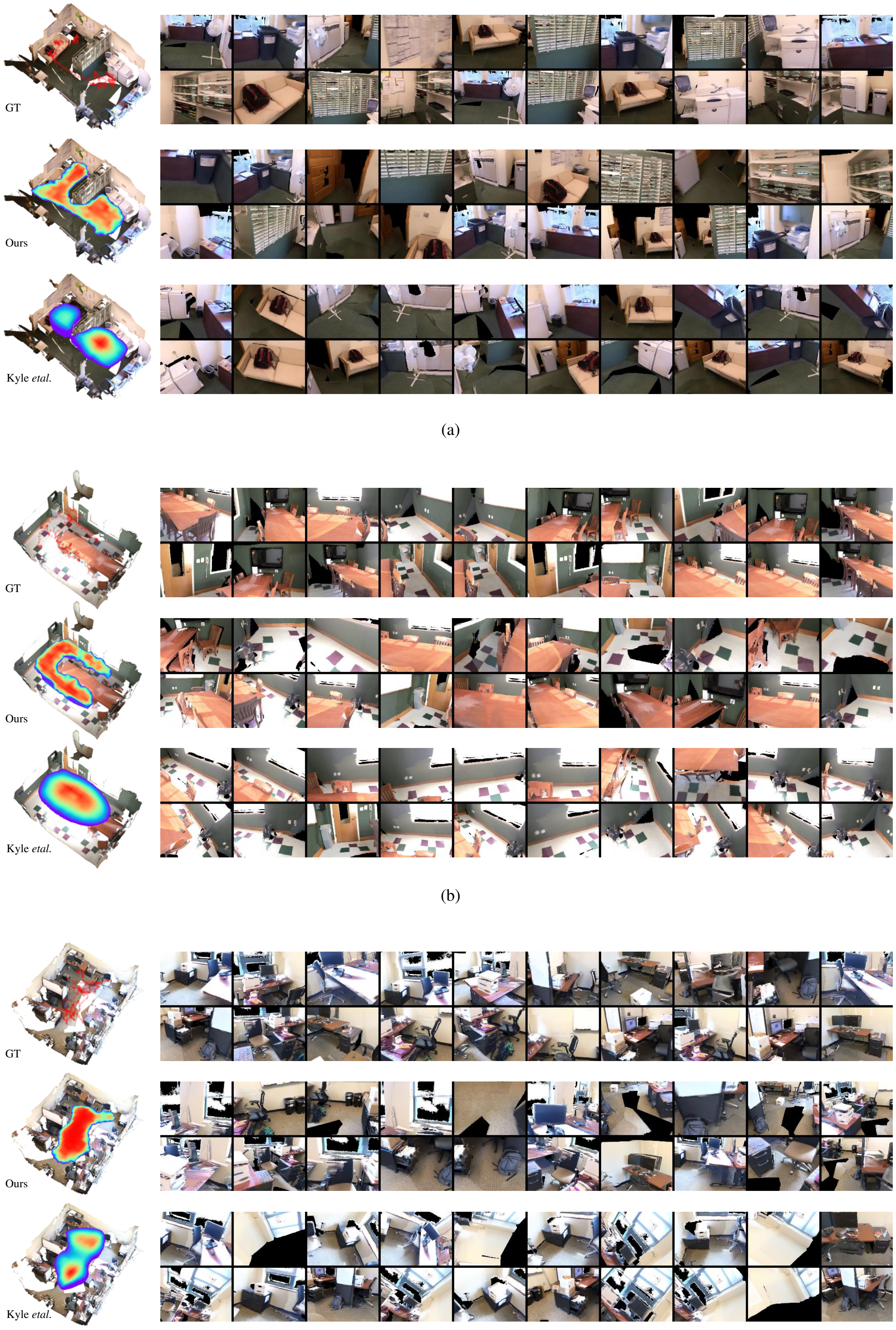


Figure 5: **Results on ScanNet (Group 2).** The left 3D figures visualize the location score using 3D heat-volumes clipped to the height of maximal accumulated score. The right pictures are randomly picked viewpoints at an optimal threshold gives best F1 score. Each group of rendered viewpoint pictures corresponds to a different method. From top to bottom is GT, ours and Kyle *et al*. [3].



(a)

(b)

(c)

Figure 6: **Results on ScanNet (Group 3).** The left 3D figures visualize the location score using 3D heat-volumes clipped to the height of maximal accumulated score. The right pictures are randomly picked viewpoints at an optimal threshold gives best F1 score. Each group of rendered viewpoint pictures corresponds to a different method. From top to bottom is GT, ours and Kyle *et al*. [3].