3D Room Layout Estimation from a Cubemap of Panorama Image via Deep Manhattan Hough Transform

Yining Zhao[1], Chao Wen[2], Zhou Xue[2], Yue Gao[1]
[1] BNRist, THUIBCS, BLBCI, KLISS, School of Software, Tsinghua University
[2] Pico IDL, ByteDance

Tsinghua University · ECCV TEL AVIV 2022 · ByteDance · Pico

Overview

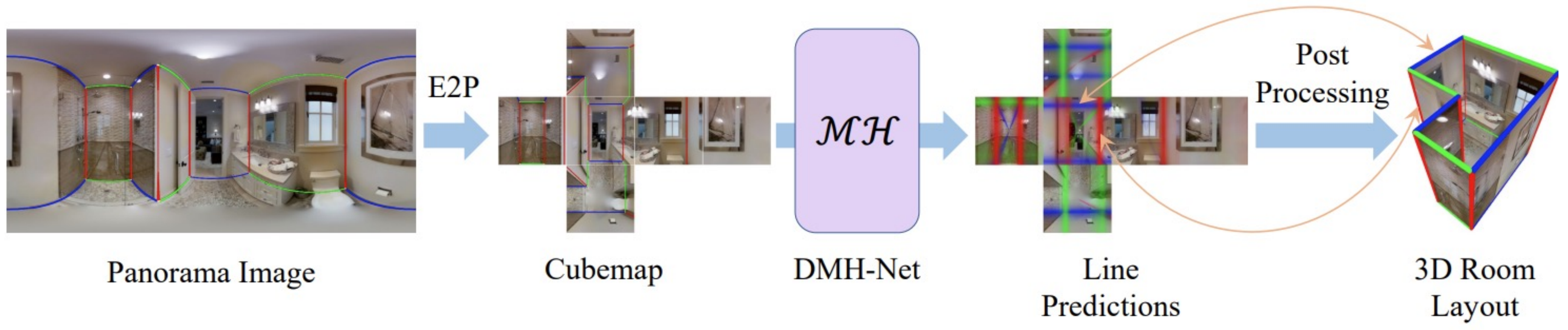
Motivation.

- 3D room layout can be compactly described.
- Detecting wireframe locally is challenging with occlusion and few appearance clues.

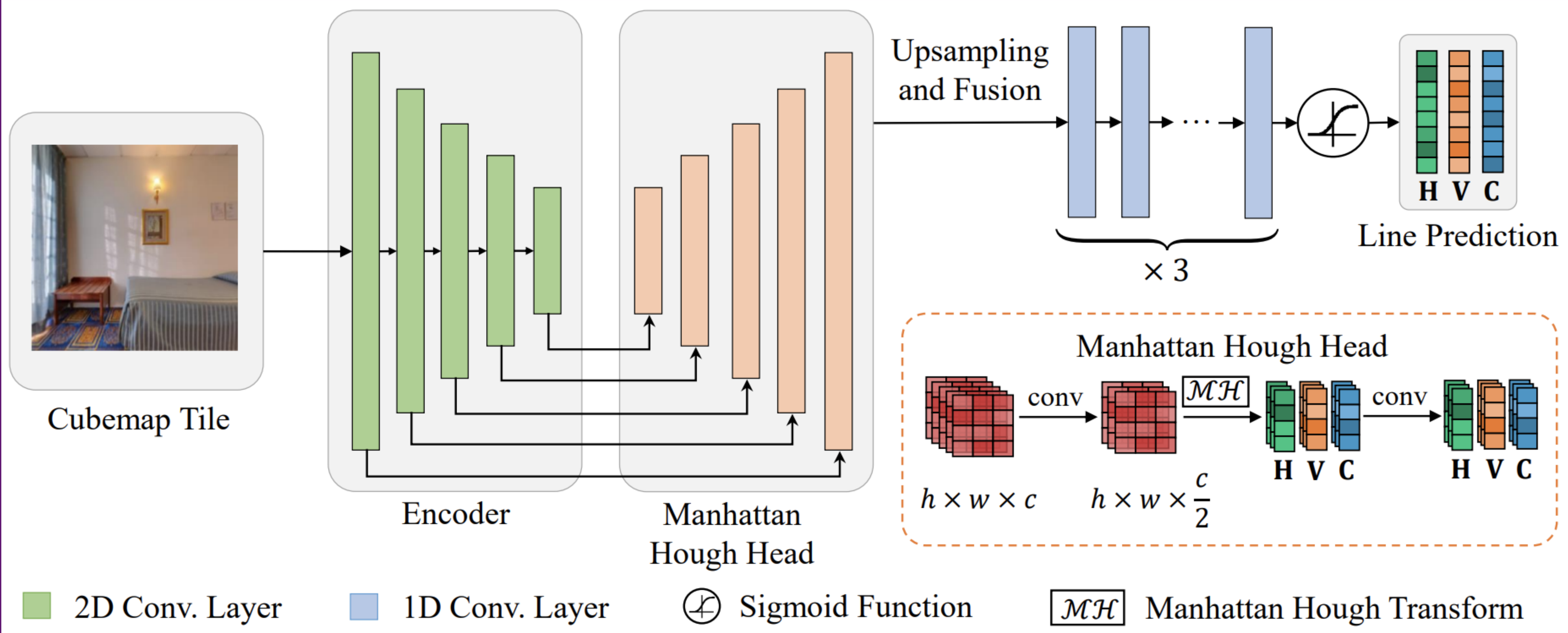
Contributions.

- We introduce **Manhattan** world assumption through **Deep Hough Transform** to capture the long-range pattern.
- We propose a novel framework estimating layouts on each distortion-free cubemap tile individually.
- We predict Manhattan lines with explicit geometric meaning.

Pipelines.



Network Architecture



Encoder. Model agnostic. Compatible with DRN or ResNet.

Manhattan Hough Head. Employ DMHT and 2D conv. to get feature in Hough space.

Upsampling, Fusion & Prediction. Fusion multi-scale feature using upsampling and convolution. Represent line prediction as probability.

Loss. Binary cross entropy loss for each types of lines:

$$\mathcal{L} = \mathcal{L}_{bce}(\mathbf{H}, \mathbf{H}^*) + \mathcal{L}_{bce}(\mathbf{V}, \mathbf{V}^*) + \mathcal{L}_{bce}(\mathbf{C}, \mathbf{C}^*)$$

Experiments

Dataset.

- Cuboid room: PanoContext & Stanford 2D-3D.
- Non-cuboid room: Matterport 3D.

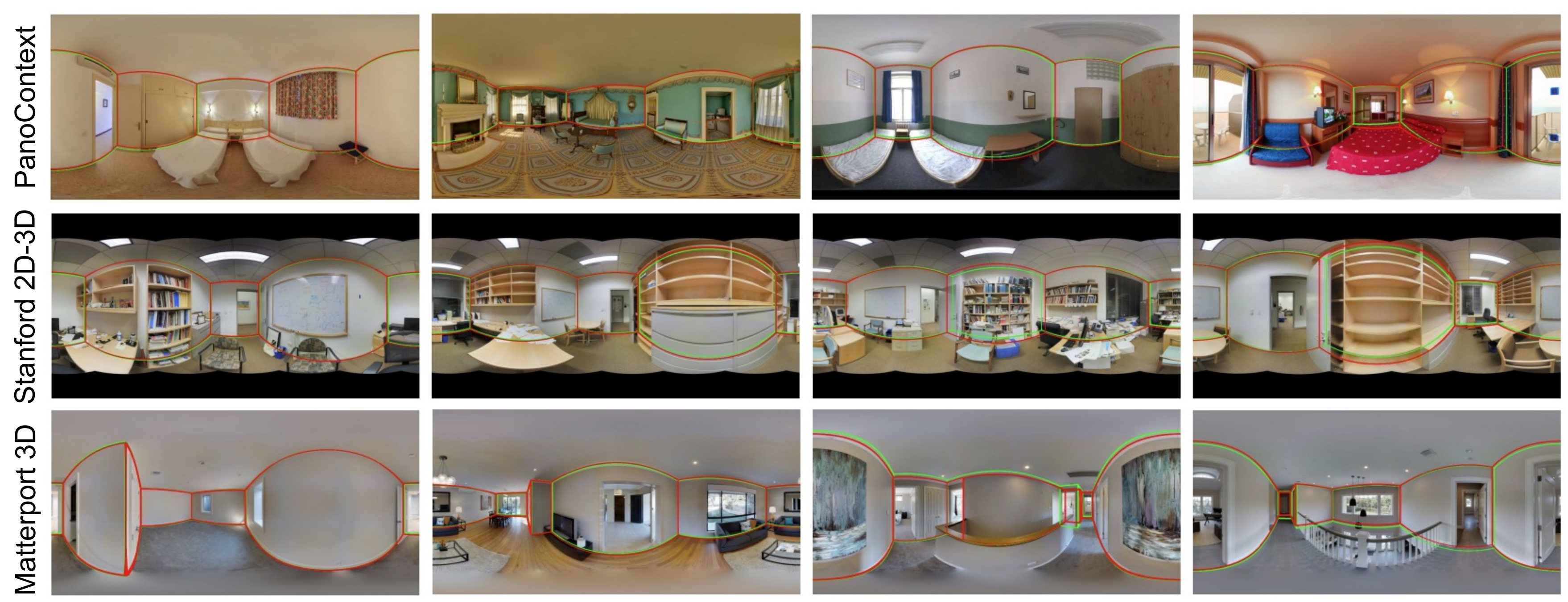
Cuboid room results.

Method	PanoContext			Stanford 2D-3D		
	3DIoU↑	CE↓	PE↓	3DIoU↑	CE↓	PE↓
PanoContext [44]	67.23	1.60	4.55	-	-	-
LayoutNet [47]	74.48	1.06	3.34	76.33	1.04	2.70
DuLa-Net [40]	77.42	-	-	79.36	-	-
CFL [11]	78.79	0.79	2.49	-	-	-
HorizonNet [34]	82.17	0.76	2.20	79.79	0.71	2.39
AtlanteNet [11]	-	-	-	82.43	0.70	2.25
LED^2-Net [36]	82.75	-	-	83.77	-	-
DMH-Net (Ours)	**85.48**	**0.73**	**1.96**	**84.93**	**0.67**	**1.93**

Non-cuboid room results.

Metrics	3DIoU ↑					2DIoU ↑					δ_i ↑				
# of corners	4	6	8	10+	Overall	4	6	8	10+	Overall	4	6	8	10+	Overall
LayoutNet-v2 [48]	81.35	72.33	67.45	63	75.82	84.61	75.02	69.79	65.14	78.73	0.897	0.827	0.877	0.8	0.871
DuLa-Net-v2[48,40]	77.02	78.79	71.03	63.27	75.05	81.12	82.69	74	66.12	78.82	0.818	0.859	0.823	0.741	0.818
HorizonNet$^+$[48,34]	81.88	82.26	71.78	68.32	79.11	84.67	84.82	73.91	70.58	81.71	0.945	0.938	0.903	0.861	0.929
AtlantaNet[27]	82.64	80.1	71.79	**73.89**	**81.59**	85.12	82.00	74.15	**76.93**	**84.00**	**0.950**	0.815	**0.911**	**0.915**	**0.945**
HoHoNet[35]	82.64	82.16	73.65	69.26	79.88	85.26	84.81	75.59	70.98	82.32	-	-	-	-	-
LED^2-Net[36]	84.22	**83.22**	**76.89**	70.09	81.52	86.91	**85.53**	**78.72**	71.79	83.91	-	-	-	-	-
DMH-Net (Ours)	**84.39**	80.22	66.15	64.46	78.97	**86.94**	82.31	67.99	66.2	81.25	0.949	**0.951**	0.838	0.864	0.925

Qualitative results. Green: ground truth. Red: our prediction.



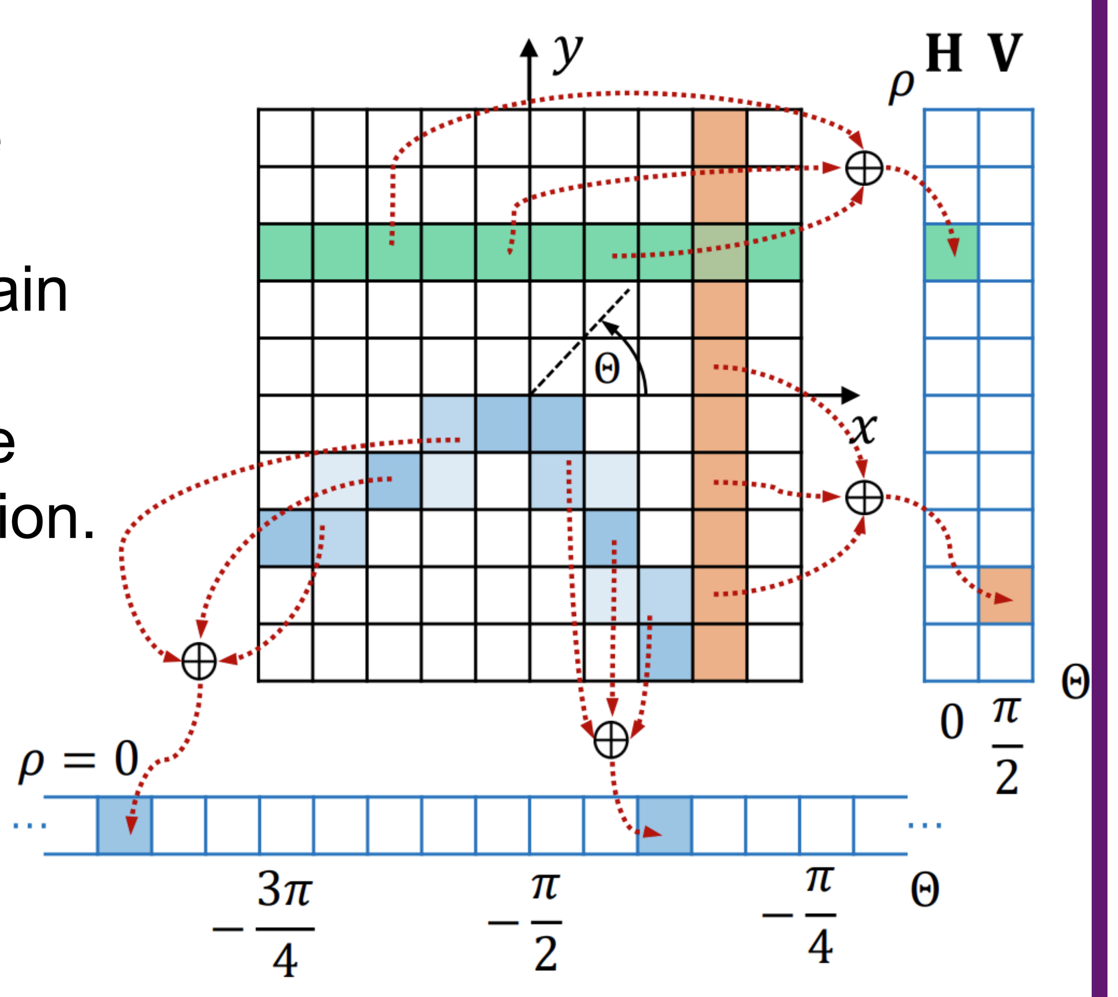
Code, Model Available here.

Deep Manhattan Hough Transform (DMHT)

Observation. After alignment, wireframe lines in cubemap should be horizontal, vertical or passing the center of the tile.

Objective. Detect and distinguish wireframe lines in each cubemap tile using Hough features.

Method. For 3 types of lines, we perform **Hough voting** on feature maps to obtain vectors indicating the confidence score of line existence at each position.

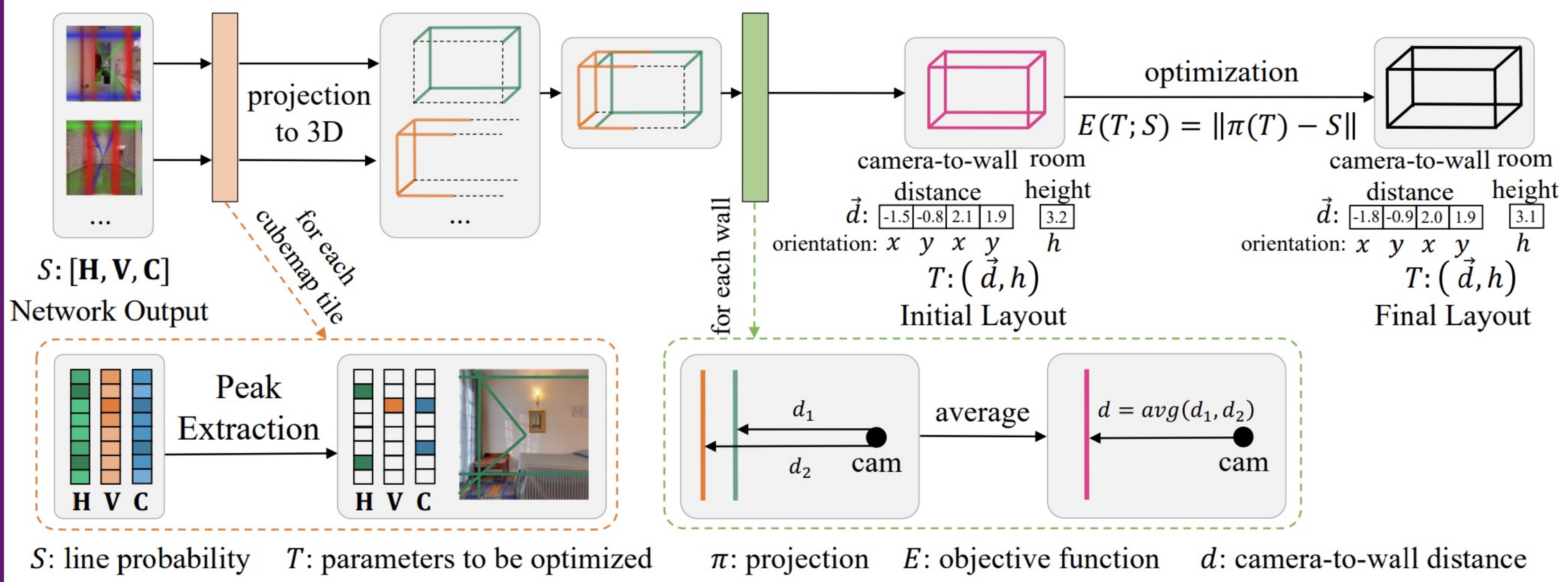


$$\mathbf{C}(\theta) = \mathcal{MH}_C(\theta) = \mathcal{H}(0, \theta)$$

$$\mathbf{V}(\rho) = \mathcal{MH}_V(\rho) = \mathcal{H}(\rho, \tfrac{\pi}{2})$$

$$\mathbf{H}(\rho) = \mathcal{MH}_H(\rho) = \mathcal{H}(\rho, 0)$$

Post processing



Representation. For an n-corner room, $n+1$ parameters are used to represent the layout: n distances from the camera to each wall, and the height of the room.

Initialization. (1) Generate partial wireframes by projection to 3D for each cubemap tile. (2) Average multiple line proposals for each wall.

Optimization. Convert layout parameters to wireframe, transformed onto each tile then maximize the overall probability via SGD.