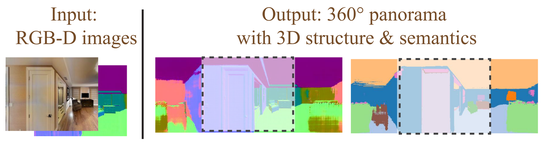
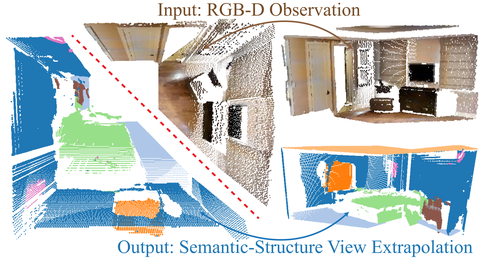
Im2Pano3D: Extrapolating 360° Structure and Semantics Beyond the Field of View

Shuran Song[1] Andy Zeng[1] Angel X. Chang[1] Manolis Savva[1] Silvio Savarese[2] Thomas Funkhouser[1]

Princeton University[1] Stanford University[2]

Introduction





In this paper, we explore the task of **semantic-structure view extrapolation.** Given a partial observation of the room in the form of an RGB-D image, the goal for this task is to predict both 3D structure and semantics for a full panoramic view of the same scene.

Key Idea

Indoor environments are often **highly structured**. By learning over the statistics of many typical scenes, the model should be able to leverage **strong contextual cues** inside the image to predict what is beyond the field of view.

Prior Work: Color Only

Hard to be used directly to support high level planning.

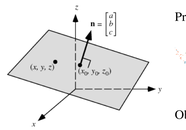
Learning to Look Around
[Jayaraman and Grauman]

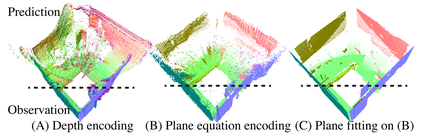
Image Inpainting
[Pathak et. al]

User-guided view extrapolation [Zhang et al.]

Data Representation

Whole room panorama representation: We use a sky-box-like multi-channel panorama to represent 3D scenes. The views are circularly connected, hence, observing the inner two views is equivalent to observing the outer two views of its shifted panorama.

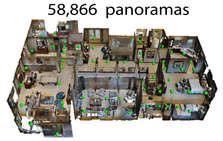




3D structure representation: We propose to represent the 3D surface using its plane equation, which decomposes the 3D surface into its surface normal (a,b,c) and plane distance to the origin (p). Top shows predictions with different encodings.

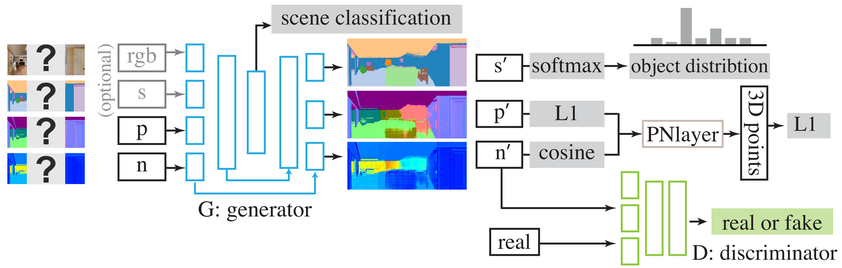
Training Data

Synthetic Houses (SUNCG):
58,866 panoramas



Real Houses (Matterport3D):
5,315 panoramas

Im2Pano3D Network



Network architecture. The network uses a multi-stream autoencoder structure. A PN-layer is used to ensure consistency between normal and plane distance predictions.

Network losses. To provide meaningful supervision we use multiple loss functions which consider pixel level accuracy and global contextual consistency.

Evaluation

models type+loss	semantics PoG	exist	size	emd	IoU	acc.	incept.	3D surface (m) mean	med.	0.2‰	0.5‰	1‰	normals (degree) mean	med.	11.25‰	22.5‰	30‰
pn2pn+A	-	-	-	-	-	-	-	***0.320***	***0.119***	***67.6***	***81.4***	***91***	38.5	5.5	70.3	74.5	76
d2d+A	-	-	-	-	-	-	-	0.353	0.148	63.1	79.6	90.1	59.0	41.2	12.7	29.3	38.9
pns2pns+S	0.379	0.613	0.653	1.184	0.313	***0.728***	0.375	0.416	0.227	51.8	74.3	88.9	32.5	7.6	62.3	72.2	76.0
pns2pns+A	0.370	0.681	0.750	1.269	0.318	0.719	0.452	0.343	0.15	63.3	80.4	90.9	37.7	4.4	72.2	76.0	77.4
pns2pns+A+S	***0.382***	***0.710***	***0.754***	***1.204***	***0.330***	0.716	***0.463***	0.339	0.151	64.0	80.8	91	***36.9***	***4.6***	***73.0***	***76.4***	***77.8***

Ablation studies on SUNCG (Named by: input2output)

models type	train	PoG	exist	size	semantics emd	IoU	acc.	incept.	3D surface (m) mean	med.	0.2‰	0.5‰	1‰	normals (degree) mean	med.	11.25‰	22.5‰	30‰
human	-	0.303	0.650	1.474	0.943	0.203	0.522	-	0.661	0.449	29.1	57.7	78.7	49.9	17.4	51.2	58.2	60.8
avg all	m	0.131	0.228	1.574	2.007	0.098	0.498	-	0.925	0.685	12.6	37.8	67.9	46.2	41.8	3.1	17.5	31.4
avg type	m	0.155	0.260	1.265	2.089	0.107	0.508	-	0.905	0.668	13.8	39.6	69.6	45.8	40.4	4.5	20.7	34.0
nn	m	0.126	0.531	1.901	2.820	0.078	0.302	-	1.286	0.898	15.8	33.6	56.4	65.1	58.1	23.8	31.2	34.9
inpaint	s+m	0.145	0.488	1.407	1.984	0.082	0.347	0.183	0.867	0.591	19.3	46.3	72.3	59.5	50.4	23.3	32.8	37.9
rgbpn2pns	s	0.185	0.56	1.589	1.729	0.129	0.378	0.233	0.609	0.365	32.3	63.4	82.5	47.2	20.8	43.6	54.7	59.4
rgbpn2pns	m	0.245	0.542	0.933	1.535	0.174	***0.566***	0.394	0.603	0.361	37.4	63.7	82.1	**39.1**	22.4	34.9	52.6	60.4
rgbpn2pns	s+m	***0.275***	***0.616***	***0.936***	***1.487***	***0.208***	***0.566***	***0.402***	***0.524***	***0.280***	***43.6***	***69.5***	***85.5***	43.6	***19.0***	***42.9***	***57.2***	***62.8***

Comparing to baseline methods on Matterport3D.

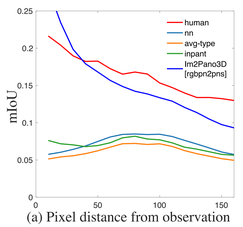
s: semantic
d: depth
p: plane distance
n: normal.
A: adversarial loss,
S: scene attribute loss.

Baselines:
avg: average distribution
nn: nearest neighbor
impaint: [Pathak et al]

Training set:
m: matterport3d
s: suncg
s+m: both

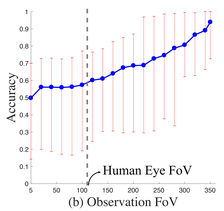
Experiments

How far can we predict?



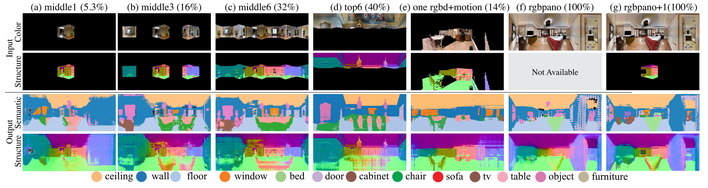
(a) Pixel distance from observation

Accuracy v.s. FOV



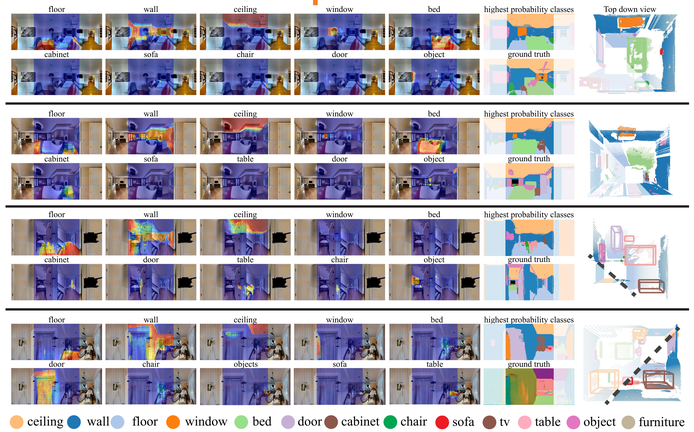
(b) Observation FoV

Input Camera Configuration

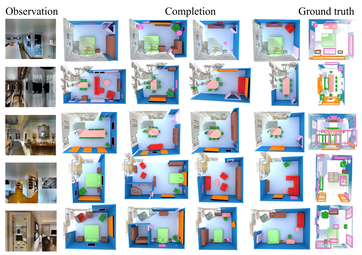


For different input camera configurations, Im2Pano3D provides a unified framework to efficiently fill in missing 3D structure and semantics of the unobserved scene. The observation coverage is shown in parentheses. The data for [RGB-D+motion] comes from NYUv2.

Example Results



Left shows probability distribution of semantics. Right shows semantic labels with the highest predicted class probability for each pixel, and reconstructed 3D point clouds.



Human Completion

First column shows the input observations. Columns 2 to 5 are completion results from different users overlaid on the observations. Sixth column shows ground truth and our prediction.

Code & Data

im2pano3d.cs.princeton.edu