

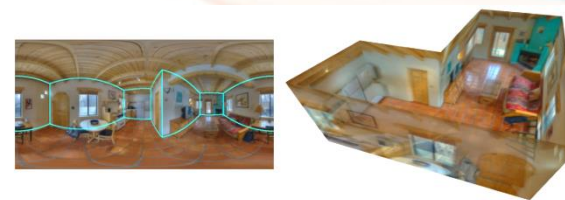
SESSION3: ROOM MODELING

Introduction

- Goal: reconstruction of individual spaces
 - Often identified as rooms
- Input: images associated with the room
 - One or few panoramic images
- Output: 3D room model or pixel-wise information
 - Single scene
 - Walls, ceilings, floor
 - Multi-modal information for specific tasks
 - Editing, VR exploration, etc.
 - Optional: multi-room integration
- Special feature of panoramic images
 - Effective reconstruction even from a single image



MVlayoutNet – Hu ACM MM2022



HorizonNet – Sun CVPR2019



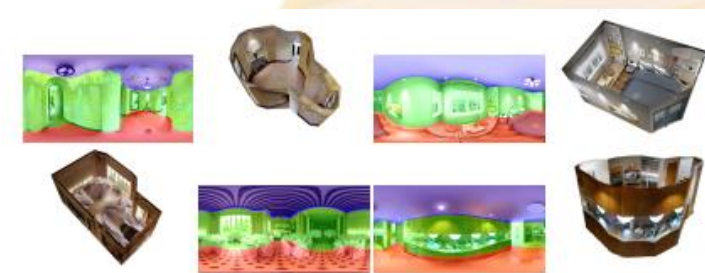
Zhang et al. ECCV 2014

Fundamental tasks

- Pixel-wise information from a single image
 - Depth, semantics, normals
 - 2D – 2D task (at least 3D features of the visible point cloud)
 - Visible scene, dense reconstruction
 - Also needed for model integration (see next Section)
- Underlying structure reconstruction: 3D layout
 - Geometry of the bounding permanent surfaces
 - i.e., ceiling, floor, walls
 - 2D – 3D task (corners, edges, planes, meshes)
 - Dealing with occlusions, sparse approximation



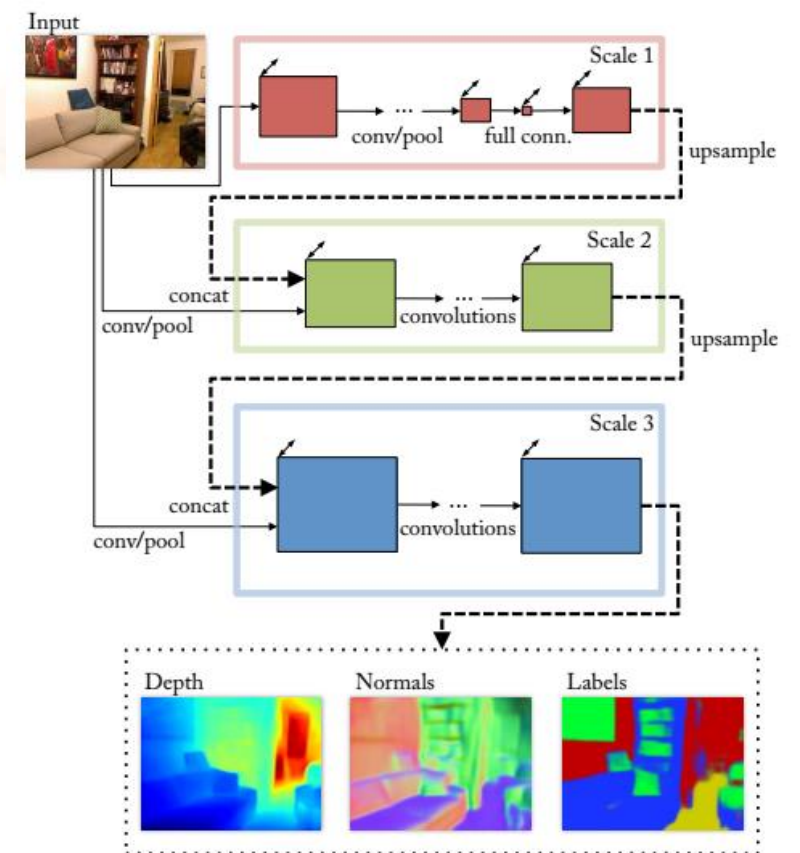
Structured3D ECCV 2020



AtlantaNet ECCV 2020

Pixel-wise reconstruction

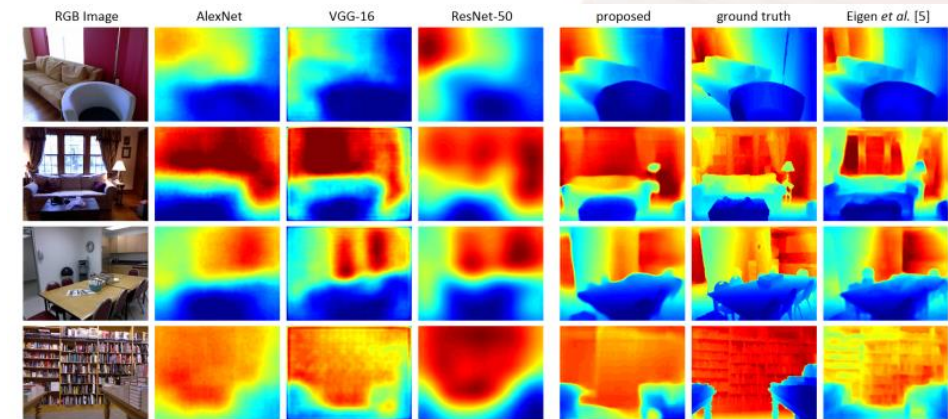
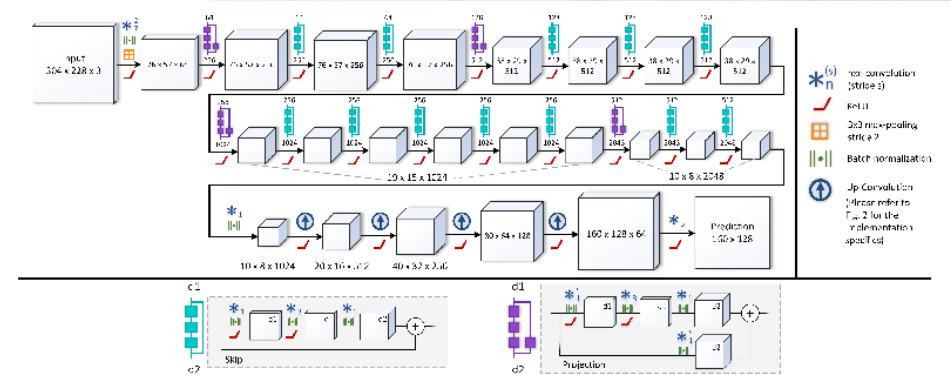
- Most common: depth reconstruction
- Naturally fits deep learning approach
 - Often supervised
- Standard solution: encoder-decoder scheme
 - Changing latest layer and activation function
 - Depth
 - Semantic
 - Normals
 - ...



Eigen et al. ICCV2015

Depth estimation from a single image

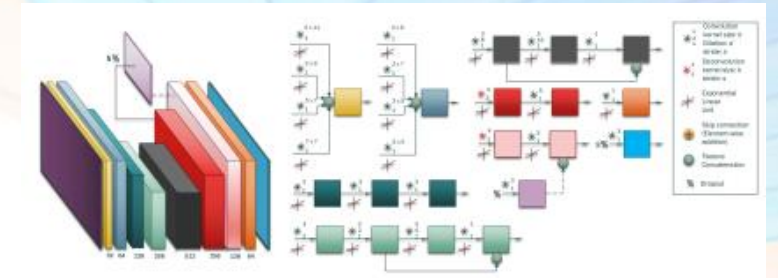
- Early perspective solution: FCRN
 - Baseline of much of the following work
 - Fully convolutional
 - Residual architecture
 - Huber loss
- Generic encoder-decoder solution
 - No indoor priors
 - No spherical assumptions
- Simply adaption leads to poor results for indoor panoramic scenes



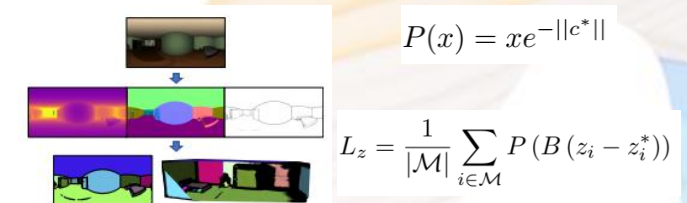
FCRN – Laina 3DV2016

Depth estimation from indoor panorama

- Distortion-aware convolutions: OmniDepth
 - FCRN baseline extended with specialized kernels
 - Targeted for equirectangular images
- Indoor priors in the loss function: panoPopups
 - Planar assumption
 - Principal curvature
 - Depth and normal map prediction



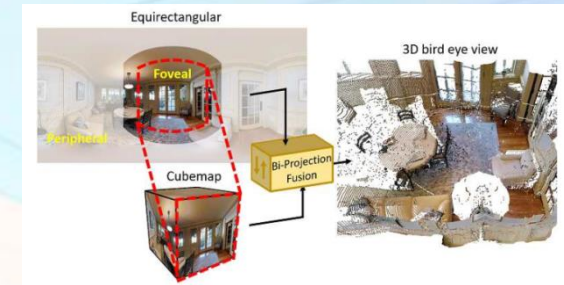
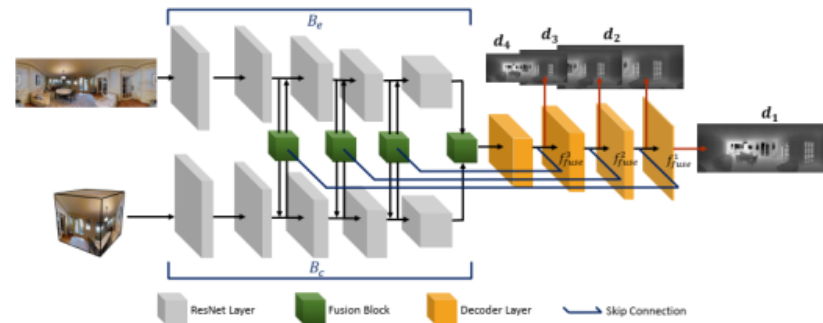
OmniDepth - Zioulis ECCV2018



Pano Popups - Eder CVPR2019

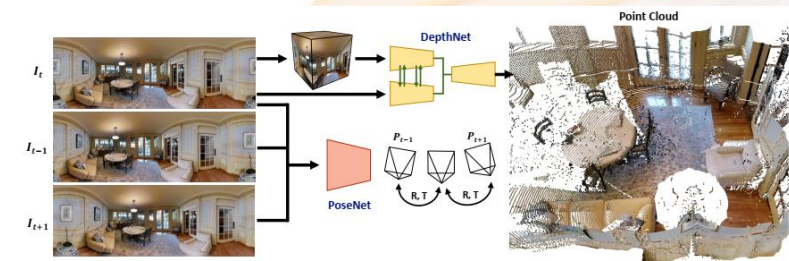
Depth estimation from indoor panorama

- BiFuse: fully-supervised dual branch network
 - Equirectangular branch: capturing wide context
 - Cubemap branch: aiming to minimize distortion
 - Multiple-fusion at feature-level
 - Computationally expensive



BiFuse - Wang CVPR 2020

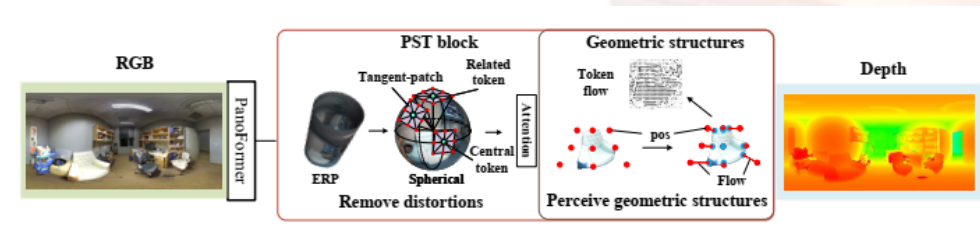
- BiFuse++: self-supervised dual branch network
 - Photo consistency from 3 adjacent panoramas
 - Improves BiFuse performance
- No specific indoor assumption adopted



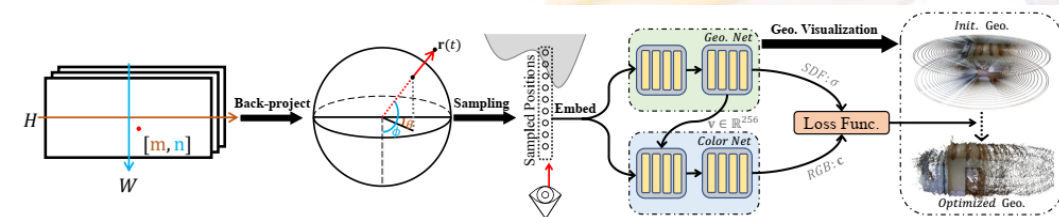
BiFuse++ - Wang TPAMI 2022

Depth estimation from indoor panorama

- Using transformers: PanoFormer
 - Improving small details recovery
 - Combining self-attention and tangent images (Eder CVPR2020)
 - Baselines of latest works (EGFormer ICCV2023, Elite360D CVPR2024)
- Neural scene representation
 - Multi-view: 3 panorama required
 - Better than common SfM
 - But not comparable on large scale datasets
 - NeRF drives self-supervised learning
 - Manhattan World weights initialization



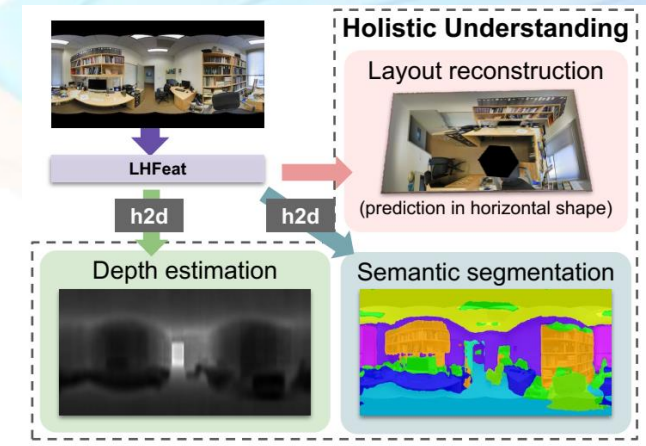
PanoFormer - Shen ECCV2022



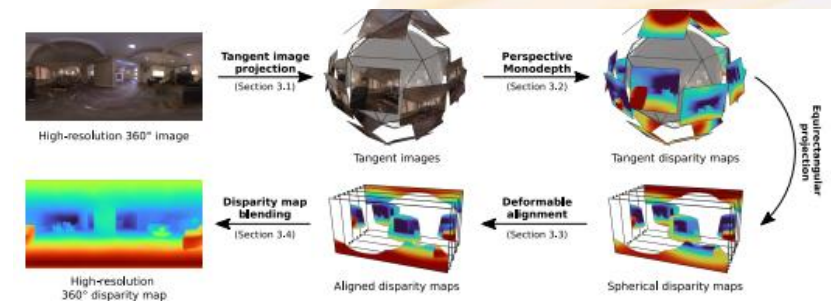
Chan et al. CVPR2023

Pixel-wise reconstruction: summary

- Fundamental task to create high-level models
 - Holistic and semantic understanding
 - Supporting single and multi-view reconstruction
- Open problems
 - Resolution
 - Computational scalability
 - Self-supervised methods limits
 - Less performance for single view
 - Prediction consistency for multi-view
 - Limit to an effective integration



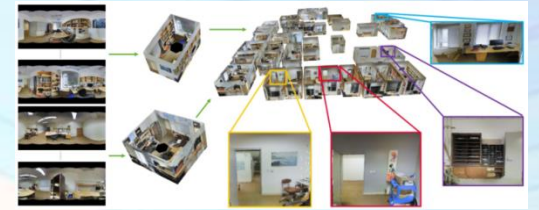
HoHoNet - Sun CVPR2021



360MonoDepth - Rey-Area CVPR2022

3D layout reconstruction

- Geometry of the bounding surfaces
 - Rooms from one or more images
 - combined to form a more complex structured model
- Common in panoramic world: room from a single image
 - Easier to capture (even by stitching with a mobile)
 - Avoid additional image-registration pipeline
 - Multi-view only to compose more complex environments
 - See next Sec.
 - A single 360 image contains enough information



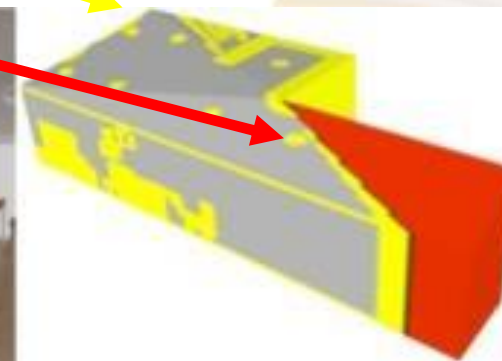
Hu et al. ACM MM2022



Zou et al. IJCV2022

Layout reconstruction: indoor issues

- Differences from depth estimation
 - Occlusions from clutter
 - Self-occlusions
- Ambiguities
 - Texture-poor surfaces
 - ...
- Need for indoor priors!

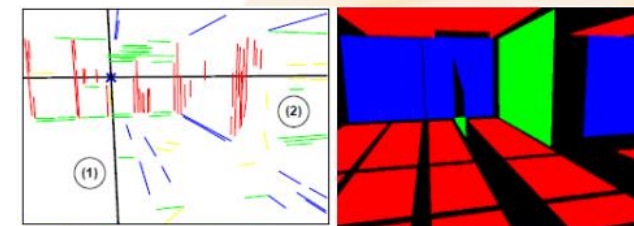


Layout from a single RGB image

- Early approaches (perspective)
 - Fundamental priors (see prev. sessions)
- Geometric context (GC) for indoor scene
 - Cuboid (CB) prior
 - Room box and surface labels jointly estimated
 - floor, ceiling, wall, objects
- Orientation maps (OM) from MW vanishing lines
 - Indoor World Model (IWM) geometric reasoning
 - Manhattan world planes bounding the room



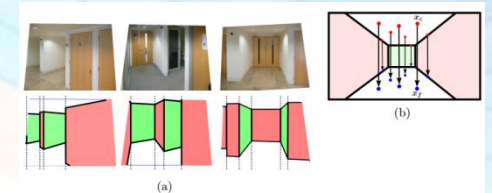
Hedau et al. ICCV 2009



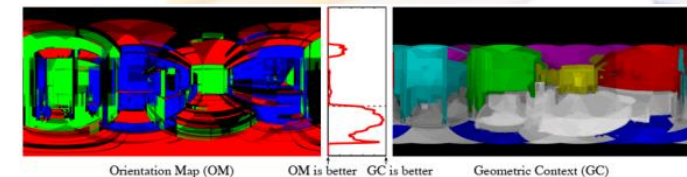
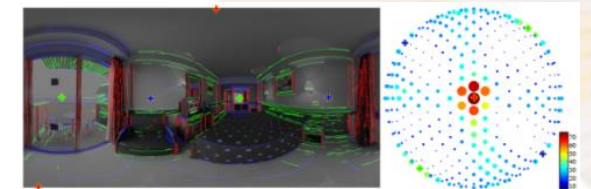
Lee et al. CVPR 2009

Layout from a single RGB image

- Geometric context (GC) and Orientation Map (OM)
 - Basis of indoor geometric reasoning
- Geometric reasoning on the IWM
 - Horizontal floor and ceiling related by a homography
- Geometric reasoning on panorama: PanoContext
 - Panoramic image converted into perspective images
 - e.g, cubemaps
 - GC and OM applied to perspective views
 - Results re-projected on the original panorama



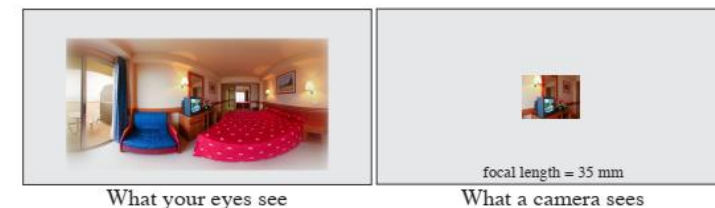
Flint et al. ECCV 2010



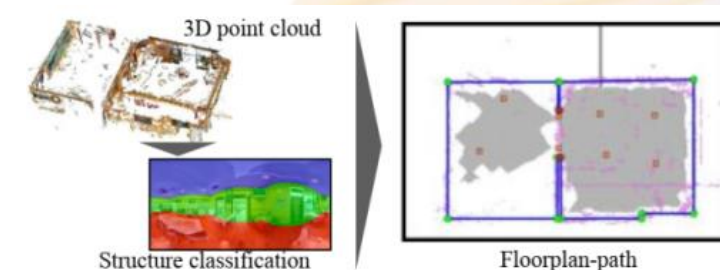
PanoContext - Zhang. ECCV2014

Layout from a single panorama

- PanoContext
 - Demonstrates advantages of panoramic vs. pine-hole view
 - Image preprocessing: GC+OM exploited for aligning to Manhattan World axis
 - Still adopted by many modern pipelines
 - Simplifying reconstruction
 - 3D room box fitting
- Room shape from panorama
 - Improving room shape estimation
 - Exploiting panoramic image to improve reconstruction from a point cloud



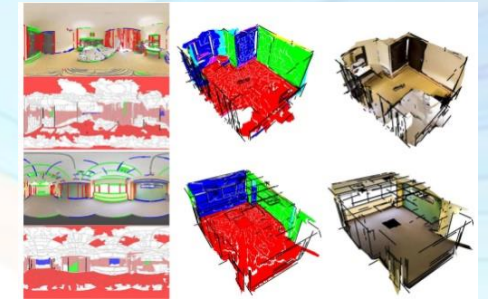
PanoContext - Zhang. ECCV2014



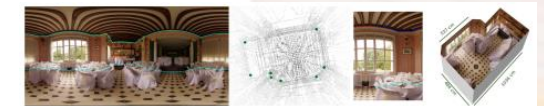
Cabral et al. CVPR2014

Layout from a single panorama

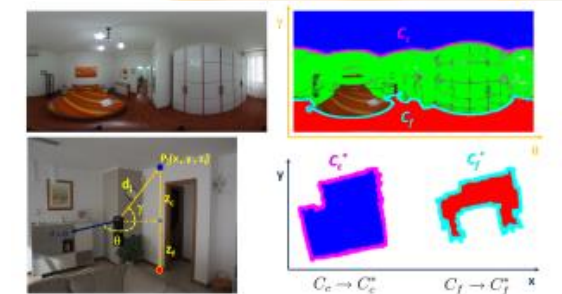
- Fully geometric reasoning approaches
 - Low-level features derived by GC e OM
 - Super-pixels and edges
 - Spatial transforms from indoor priors
 - Super-pixels and edges projected to floorplan
- ..until the rise of deep-learning approaches
 - Data-driven features
 - Indoor priors still hold



Yang et al. CVPR2016



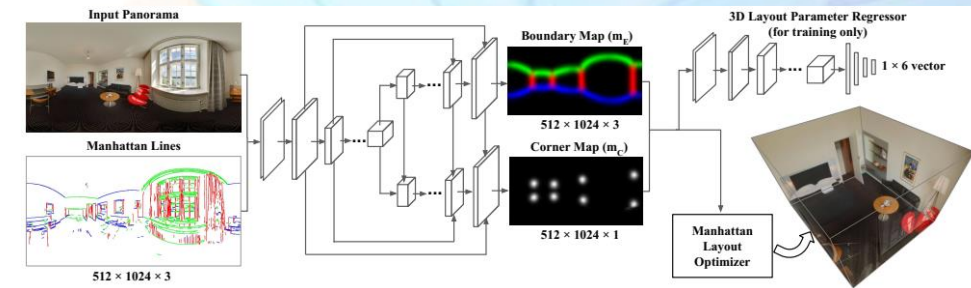
Pintore, Garro, Ganovelli, Gobbetti, Agus. WACV2016



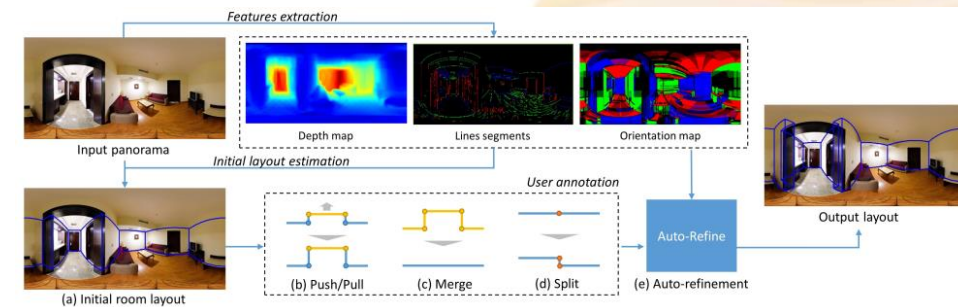
Pintore, Pintus, Ganovelli, Scopigno, Gobbetti C&G 2018

Data-driven layout from panorama

- Impressive results
 - Accuracy and speed
- Typical output
 - Corners and boundaries
- Large annotated datasets needed
 - Synthetic or real
 - Often rectified images for real-world data
 - GC+OM alignment
 - Abstract models
 - User annotation
 - Assisted tools needed



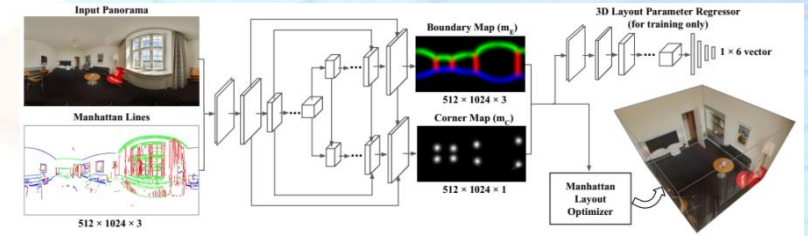
LayoutNet – Zou CVPR2018



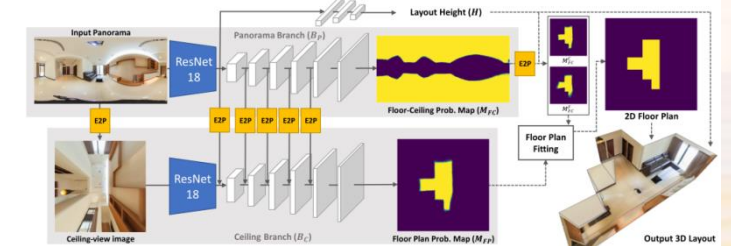
PanoAnnotator – Yang Siggraph Asia post. 2018

Data-driven layout from panorama

- Common pipeline
 - Image pre-processing
 - Reduce 2D-3D error
 - Elements prediction
 - 2D image points: corners, boundaries, floor maps
 - Final model post-processing
 - Priors-guided regularization and 3D model generation



LayoutNet – Zou CVPR2018



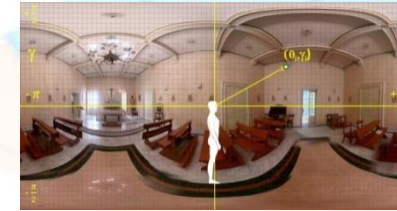
DuLaNet – Yang CVPR2019

Method	Input			Pre-process	Network Architecture				Output			Post-processing	
	RGB in Equi-rectangular View	RGB in Ceiling View	Manhattan Lines Map		Encoder		Decoder		Corner Position	Boundary Position	Floor Map	Equirectangular View	Ceiling View
					SegNet	ResNet	Equirectangular View	Ceiling View					
LayoutNet	●		●	●	●		●		●		●		
DuLa-Net	●	●		●		●		●		●		●	
HorizonNet	●			●		●			●	●		●	

Zou et al. IJCV2021

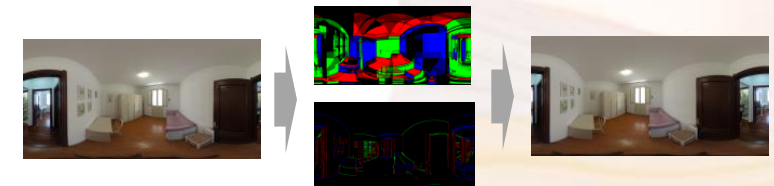
Data-driven layout: pre-processing

- Improve 2D-3D matching
 - Equirectangular to 3D space
 - Undistorted images
 - Warped panorama (usually by GC+OM)
 - Vanishing lines aligned with MW axis
 - LayoutNet, DulaNet, HorizonNet
 - Computational expensive!
 - E2P/A2P projections
 - Highlighting structures
 - DulaNet, AtlantaNet
 - Minimal computational cost but requires previous warping



$$G_h(\theta, \gamma) = \begin{cases} x = h / \tan \gamma * \cos \theta \\ y = h / \tan \gamma * \sin \theta \\ z = h \end{cases} \quad d = \frac{h}{\tan \gamma}$$

Correspondence between angles and 3D points for Atlanta World scenes



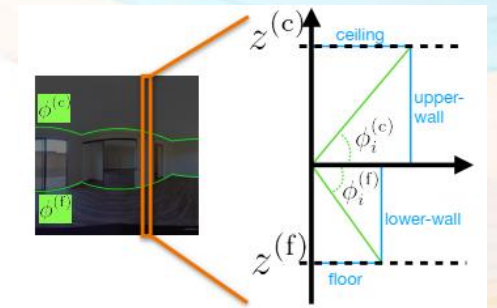
LayoutNet: warping to MW direction



AtlantaNet: A2P transform

Data-driven layout: pre-processing

- Why alignment: indoor equirectangular case
 - Image corners and boundaries match 3D points only if image is aligned
- Basic assumption: image aligned to gravity direction
 - Less restrictive than Manhattan World alignment
 - Avoid full pre-processing
 - Commonly verified in almost all public dataset
 - Often automatically performed by capture devices

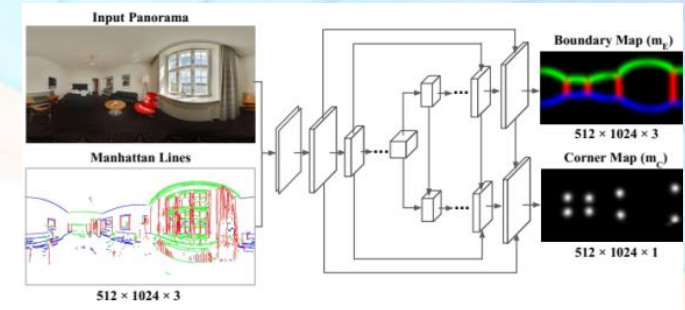


Bifuse++ – Wang TPAMI2022

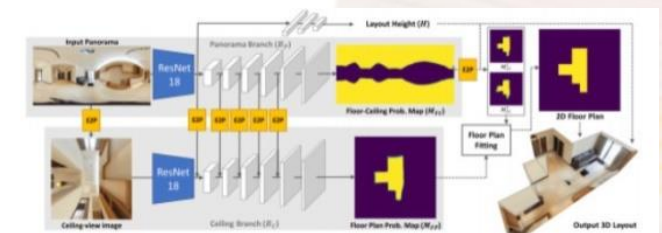


Data-driven layout: prediction

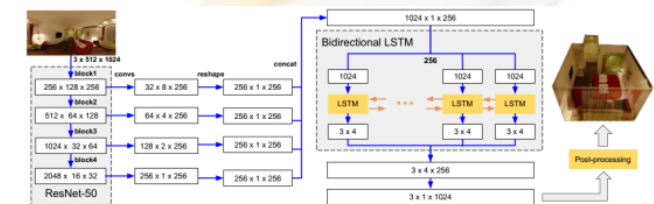
- Input
 - Aligned equirectangular image or its projections
- Encoder-decoder scheme
 - ResNet, SegNet, ...
 - Bottleneck and expansion to original resolution
- Network output/ground truth
 - LayoutNet 2018, HorizonNet 2019
 - Corners position in image space
 - Wall-ceiling and wall-floor boundaries in image space
 - DulaNet 2019, AtlantaNet 2020
 - 2D shape on the floorplan + walls height



LayoutNet: corner and boundary map



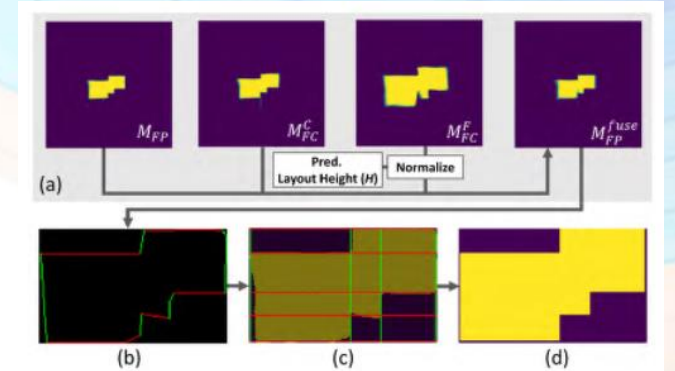
DulaNet: deep learning framework



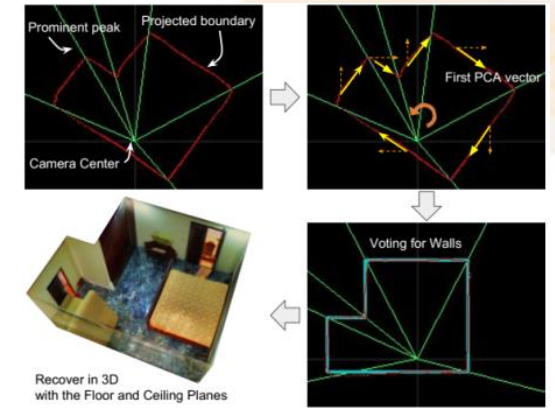
HorizonNet: corners position and W-C W-F edges

Data-driven layout: post-processing

- 2D regularization
 - Noisy or uncomplete shapes
 - Room shape on a 2D floorplan
 - Manhattan World prior
 - Walls are regressed and clustered into horizontal and vertical lines
 - Walls are fitted from corners position and ceiling/floor boundaries
 - NB. Not data-driven! heuristic approach
- 3D extrusion
 - Layout height
 - MW/ Atlanta World model : single height
 - Vertical walls model: multiple heights



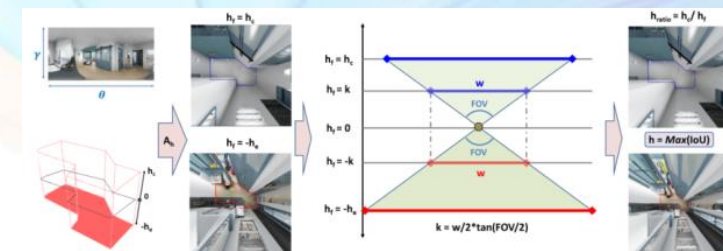
DulaNet: regression to H/V lines



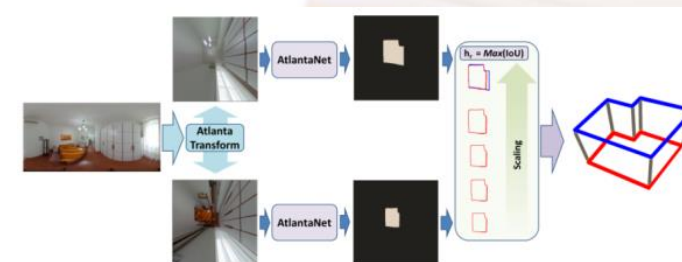
HorizonNet: voting scheme

Layout prediction example: AtlantaNet

- No Manhattan World pre and post-processing
 - Atlanta World model
 - Admit non-right angles, curved walls, etc.
 - Constrains: Horizontal floor and ceiling
- Panoramic image -> two horizontal projections
 - Undistorted 2D room footprint from ceiling map
 - Height layout encoded into floor/ceiling IoU ratio



(a) *Data encoding*



(b) *Layout recovery*

Pintore, Agus, Gobbetti. AtlantaNet. ECCV 2020

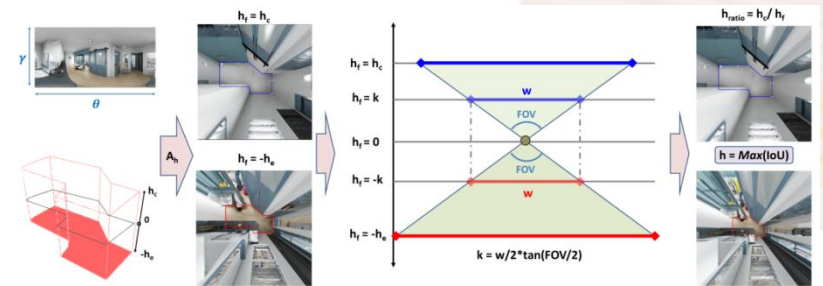
AtlantaNet: data encoding

- Atlanta Transform A_h

- Maps image points in 3D space as if their height was h_f
- Generated two tensors
- h_f can assume only two values
 - h_e (constant): floor plane distance
 - h_c (unknown): ceiling plane distance

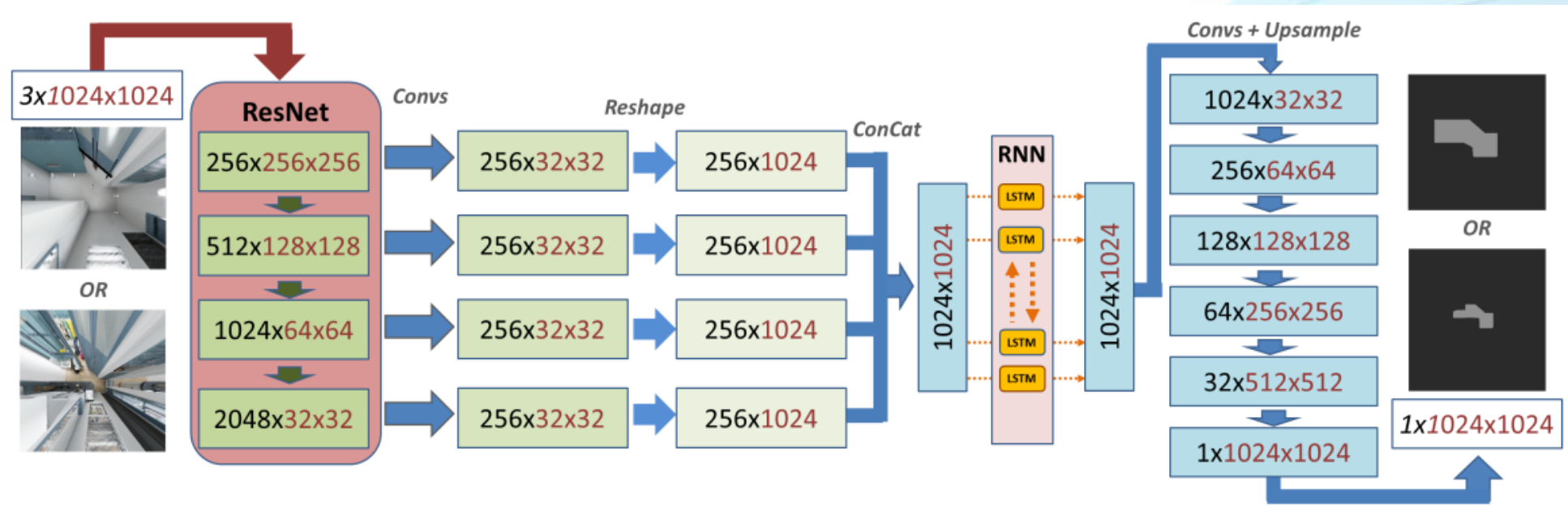
- Room height ($h_c - h_e$) is determined by h_r
 - Ratio between the ceiling and the floor shape
 - Value that matches the floor shape with the ceiling shape

$$A_h(\theta, \gamma, h_f) = \begin{cases} x = h_f / \tan \gamma * \cos \theta \\ y = h_f / \tan \gamma * \sin \theta \\ z = h_f \end{cases}$$



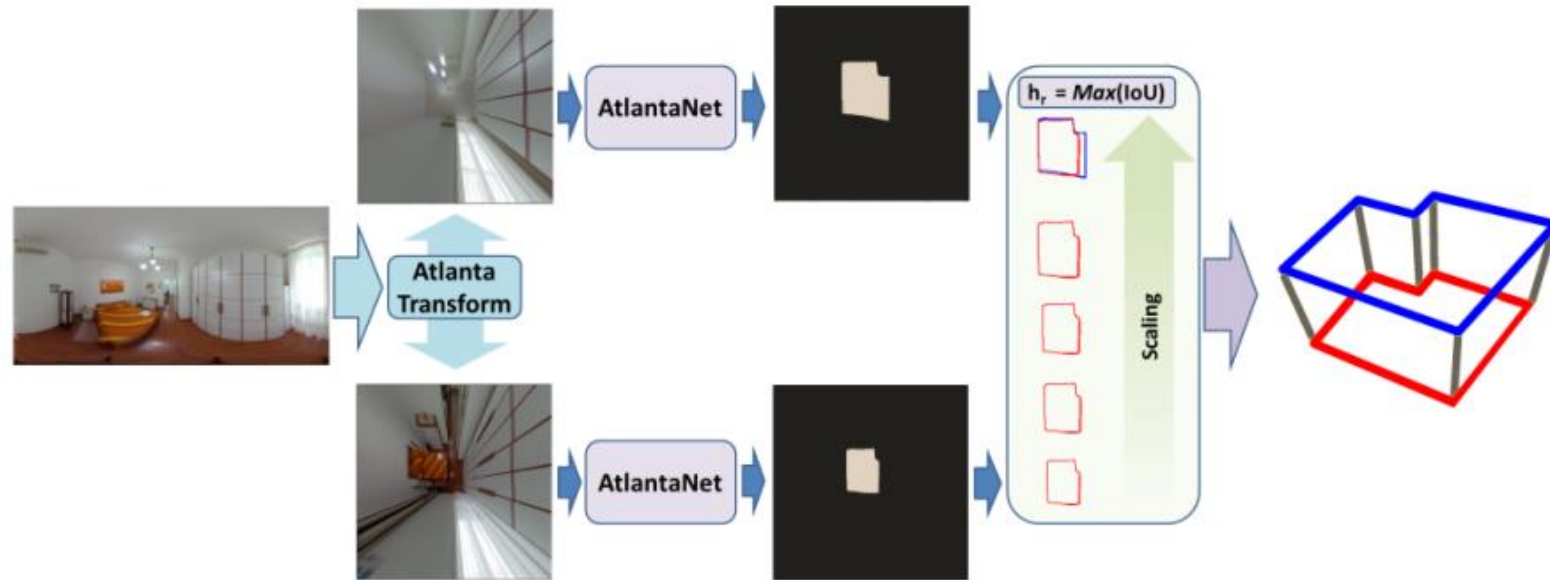
(a) Data encoding

AtlantaNet: Network architecture



- Ceiling and floor inferred separately (joined by training)
- Direct feature fusion from floor and ceiling not possible
 - Requires prior knowledge of the scale (hr)

AtlantaNet: training and inference



- Binary cross entropy on mask and its gradient
- Random feeding of floor and ceiling: augmentation
- Inference: ceiling contour and layout height from floor/ceiling shape ratio

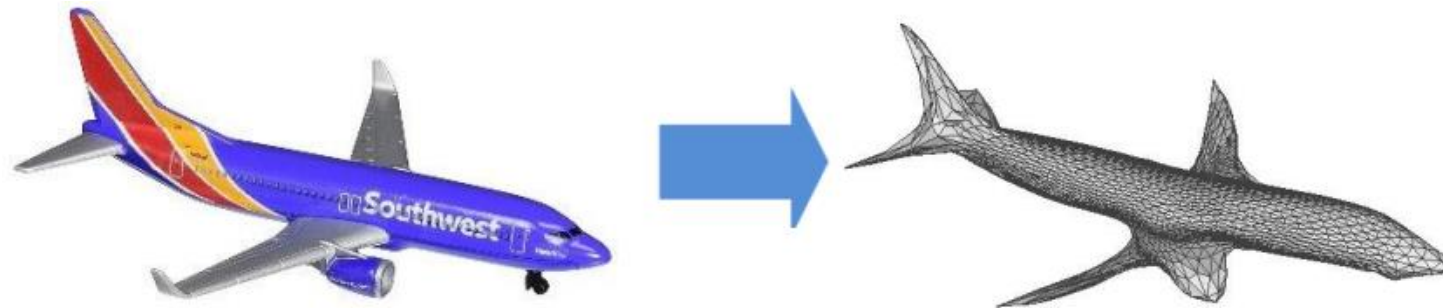
AtlantaNet: example results

- Public Datasets
 - PanoContext
 - Matterport3D
 - Stanford2D3DS
 - Structured3D
 - AtlantaLayout



Generic room shape reconstruction

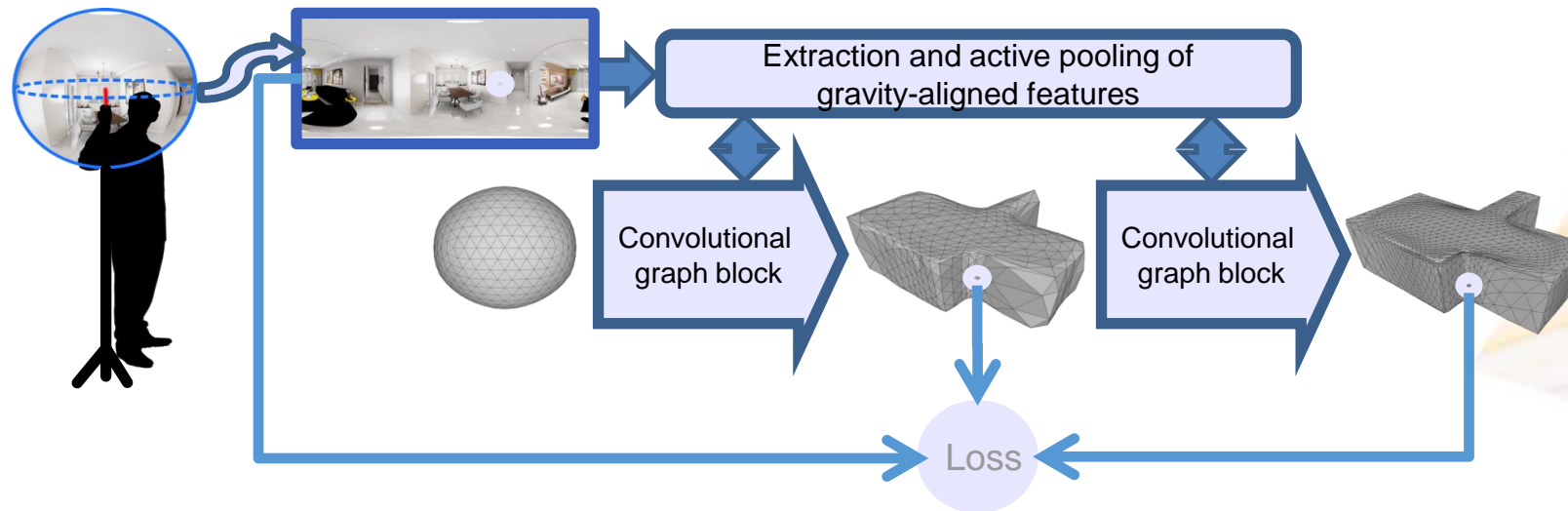
- Data-driven solutions: object surface reconstruction
 - e.g., Wang et al. [2018], Gkioxari et al. [2019], Smith et al. [2019]
 - Progressive deformation of a sphere according to image features



N. Wang et al., "Pixel2Mesh: 3D Mesh Model Generation via Image Guided Deformation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3600-3613, 1 Oct. 2021, doi: 10.1109/TPAMI.2020.2984232.

Room shape as watertight mesh

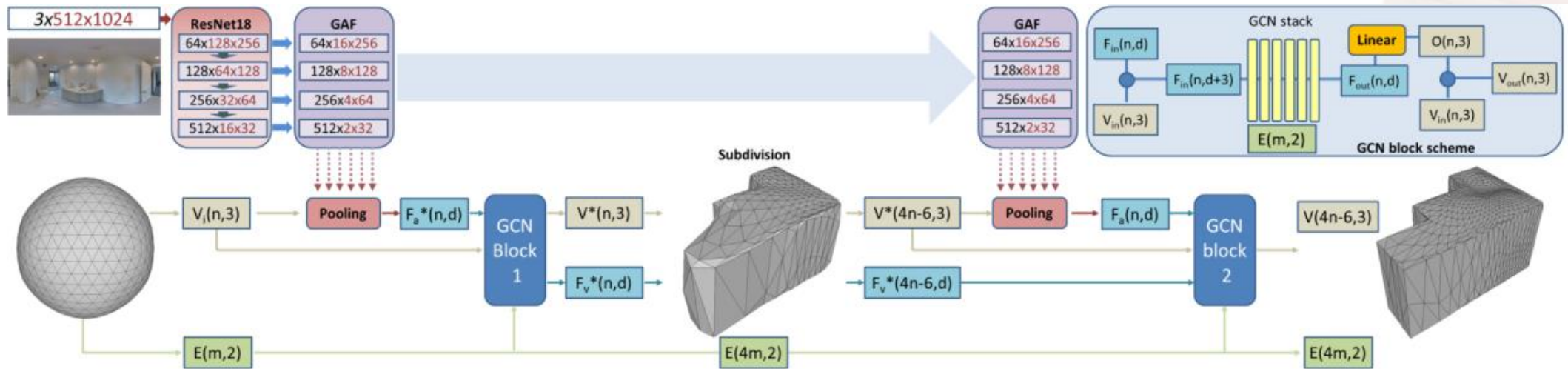
- Indoor reconstruction target: modeling arbitrary shape rooms
- Deep3Dlayout: watertight 3D mesh representing room shape
 - Handling curved walls, sloped ceilings, domes



Deep3Dlayout – Pintore, Almansa, Agus, Gobbetti, ACM TOG (SIGGRAPH Asia) 2021

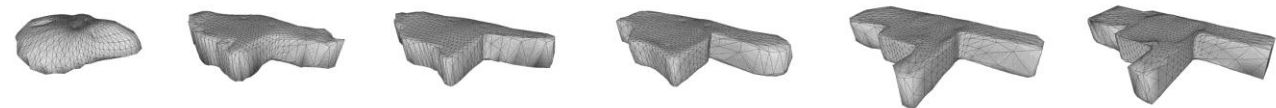
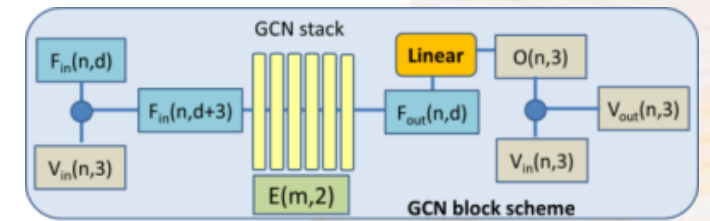
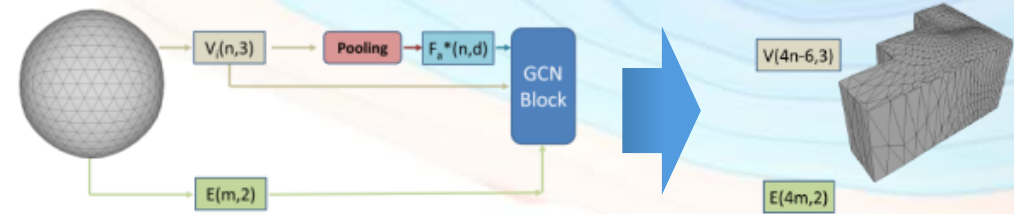
Deep3DLayout: network architecture

- Indoor layout as a 3D graph-encoded object
- Association of indoor panoramic features to 3D vertices
- Domain-specific loss function



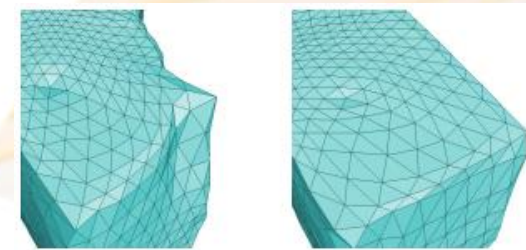
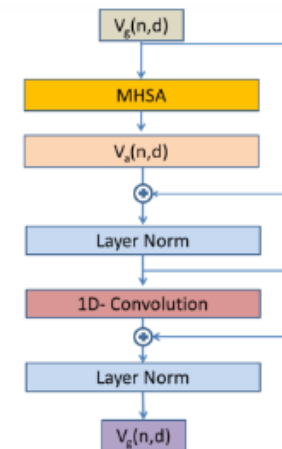
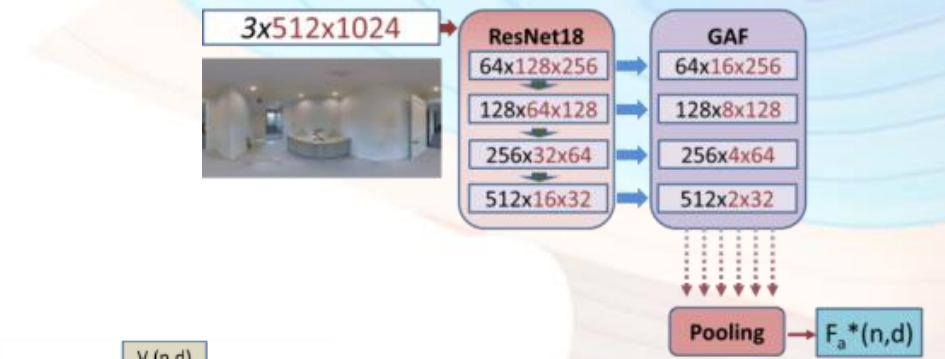
Deep3DLayout: 3D graph-encoded object

- Topological model
 - Closed 3D surface
 - Triangulated mesh
- 3D graph-encoded object
 - Vertices $V(n,3)$ and features $F(n,d)$
 - Connectivity $E(m,2)$
- Layout by mesh deformation
 - Sequence of 2 GCN blocks
 - Driven by associating image features
 - Coarse to fine approach



Deep3DLayout: feature pooling

- Gravity aligned features
 - Anisotropic contractive encoding (see SliceNet)
 - Targeted to indoors
 - Maximize gathered information
 - Minimize interpolation effects
- Spherical pooling
 - Self-attention module (MHSA)
 - Short and long range relationships
 - Cope with major occlusion problems



without MHSA

with MHSA

Deep3DLayout: model and loss functions

- Combination of data and regularization terms

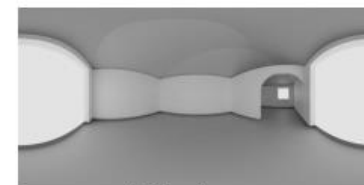
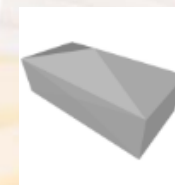
- Plausible reconstruction

$$\mathcal{L}_{data} = \lambda_c \mathcal{L}_{pos} + \lambda_n \mathcal{L}_{norm} + \lambda_{sh} \mathcal{L}_{sharp}$$

$$\mathcal{L}_{reg} = \lambda_e \mathcal{L}_{edge} + \lambda_s \mathcal{L}_{smooth}$$

- Targeted model

- Smooth surfaces joining at sharp edges
- Less restrictive than common indoor priors
 - MWM, IWM, AWM
- Allows other common structures
 - curved walls, vaults, domes



(a) Empty room

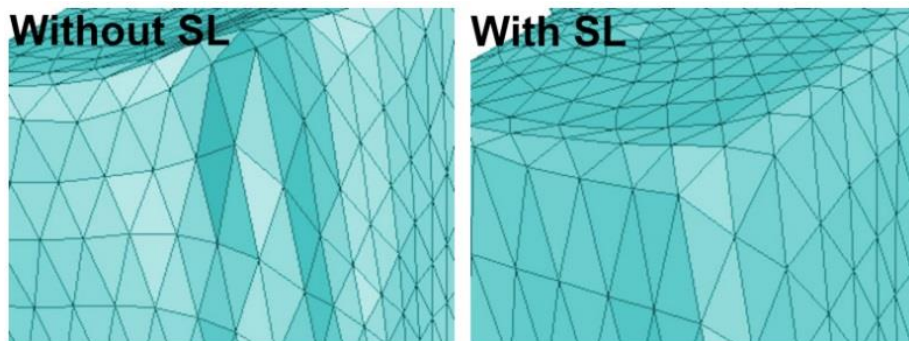


(b) Our representation

Deep3DLayout: model and loss functions

- Data terms
 - Positional and orientation loss
 - Sharpness loss

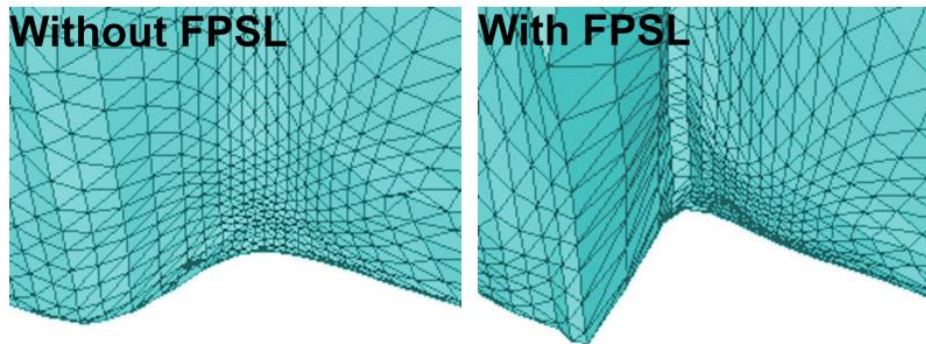
$$\mathcal{L}_{pos} = |P|^{-1} \sum \|p - N(Q, p)\|^2 + |Q|^{-1} \sum \|q - N(P, q)\|^2$$
$$\mathcal{L}_{norm} = -|P|^{-1} \sum_{p \in P} |n_p \cdot n_{N(Q, p)}|^2 - |Q|^{-1} \sum_{q \in Q} |n_q \cdot n_{N(P, q)}|^2$$
$$\mathcal{L}_{sharp} = |S_e|^{-1} \sum_{q \in S_e} \|q - N(P, q)\|^2$$



Difference in using or not the sharpness loss (SL)

Deep3Dlayout: model and loss functions

- Data terms
 - Positional and orientation loss
 - Sharpness loss
- Regularization terms
 - Edge loss
 - Smooth loss



Difference in using or not the feature-preserving smoothness loss (FPSL)

$$\mathcal{L}_{pos} = |P|^{-1} \sum \|p - N(Q, p)\|^2 + |Q|^{-1} \sum \|q - N(P, q)\|^2$$
$$\mathcal{L}_{norm} = -|P|^{-1} \sum_{p \in P} |n_p \cdot n_{N(Q, p)}|^2 - |Q|^{-1} \sum_{q \in Q} |n_q \cdot n_{N(P, q)}|^2$$
$$\mathcal{L}_{sharp} = |S_e|^{-1} \sum_{q \in S_e} \|q - N(P, q)\|^2$$

$$\mathcal{L}_{edge} = |E|^{-1} \sum_{(i, j) \in E} \|v_i - v_j\|^2$$

$$\mathcal{L}_{smooth} = |V|^{-1} \sum_{i \in V} e^{-|K_{H_i}|} |K_{H_i}|$$

Deep3DLayout: example results



(a) MatterportLayout RGB



(g) *AtlantaLayout** RGB



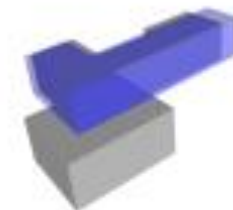
(b) ground truth



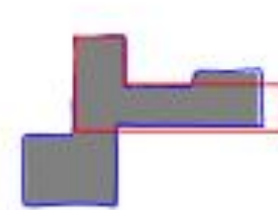
(c) ours



(d) ours vs gt



(e) [Wang et al. 2021]



(f) Floorplan



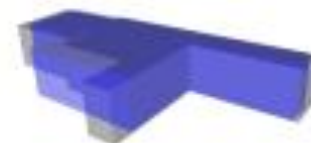
(h) ground truth



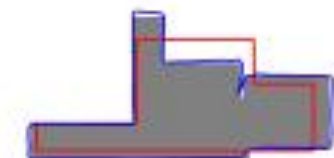
(i) ours



(j) ours vs. gt



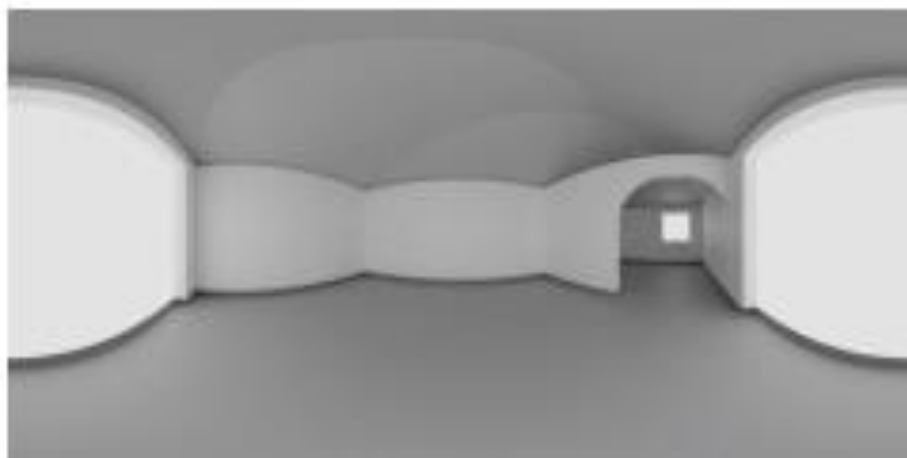
(k) [Pintore et al. 2020a]



(l) Floorplan

Room shape as watertight mesh: limitations

- Promising solution but important limitations
 - Indoor structure often are not closed, watertight mesh
 - Output is not really structured
 - GCN are not stable, transfer-learning limits



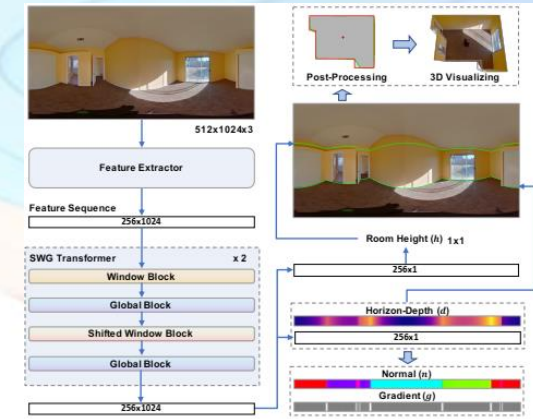
Empty room structure



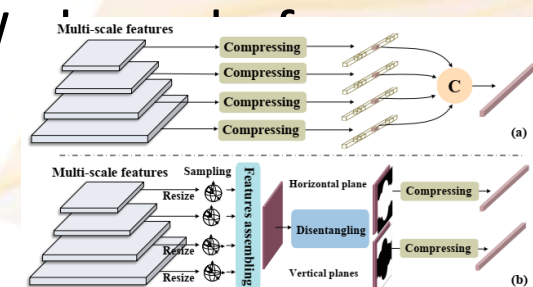
Watertight mesh approximation

Structured layout: latest trends

- Geometry-aware transformers
 - Vertical compression (as SliceNet, HoHoNet, Deep3DLayout)
 - Custom transformer improving LSTM or MHSA
 - Encoding local and global window blocks
 - Horizon depth and planar-aware losses (as LED2-Net)
- Disentangling Orthogonal Planes
 - Multi-scale features disentangled as horizontal and vertical MW compression and gated fusion
 - Cross-scale Distortion Awareness (PanoFormer)
 - MW pre post processing still adopted (same of HorizonNet)



LGT-Net Jiang CVPR2022



DOPNet Jiang CVPR2023

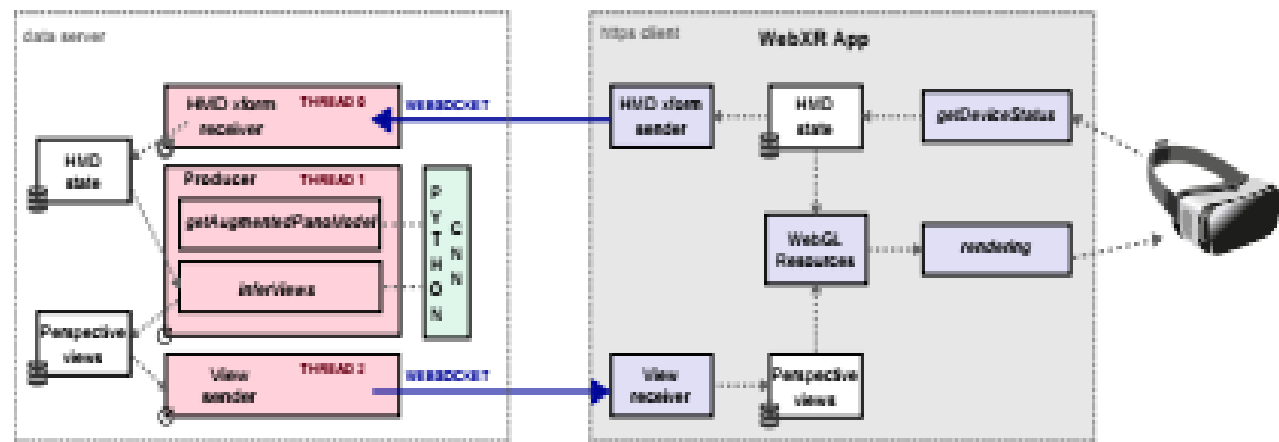
Room layout reconstruction: summary

- Geometry of the bounding permanent surfaces
 - 2D – 3D task (corners, edges, planes, meshes)
 - Common in panoramic world: room from a single 360 image
 - Rooms concept can be extended to larger and complex structures (next)
- Open problems
 - Occlusions: from clutter and from structure itself
 - Manhattan regularization and completion still common
 - Output model: truly representative of the structures and usable (e.g., CAD-like)
 - Corners and continuous edges: very limiting
 - Closed meshes: difficult to structure (walls, floor, ceiling)

HoHoNet - Sun CVPR2021

Example of depth and layout estimation pipeline

- Specific case: simultaneous depth and layout estimation
- Fundamental part of a data-driven VR approach
 - Geometric and structural information from a single indoor panorama
 - See next session *Visual representation generation and exploration*

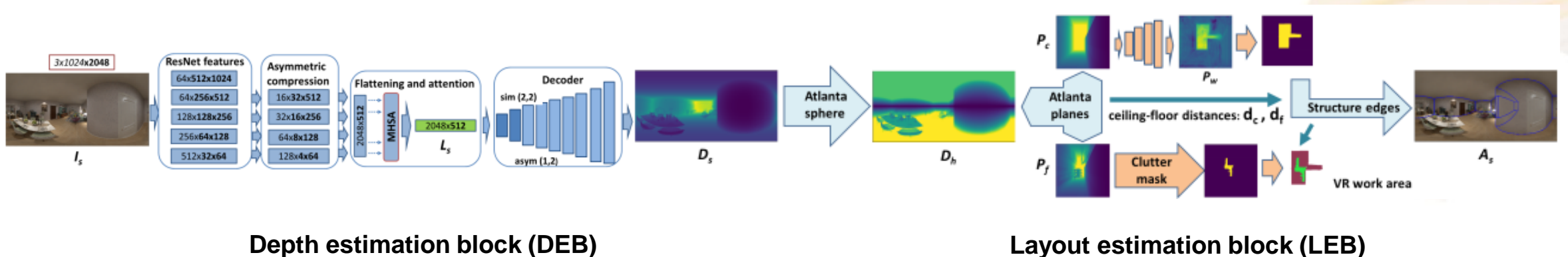


Pintore, Bettio, Agus, Gobbetti. Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image IEEE TVCG 2023

Depth and layout estimation example 1/4

- **Depth and 3D layout prediction**

- DEB: exploits gravity aligned features (GAF)
 - GAF: enables efficient features compression and latent representation
- LEB: exploits Atlanta World prior
 - Efficient layout estimation
- DEB + LEB are integrated in the same end-to-end network

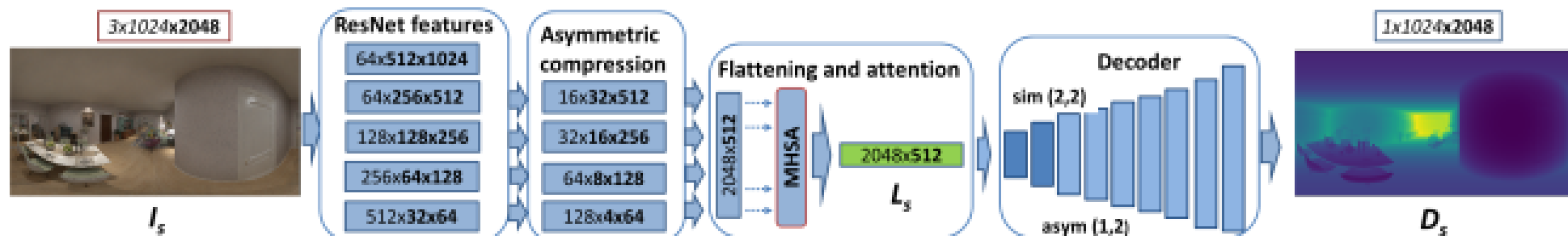


Depth estimation block (DEB)

Layout estimation block (LEB)

Depth and layout estimation example 2/4

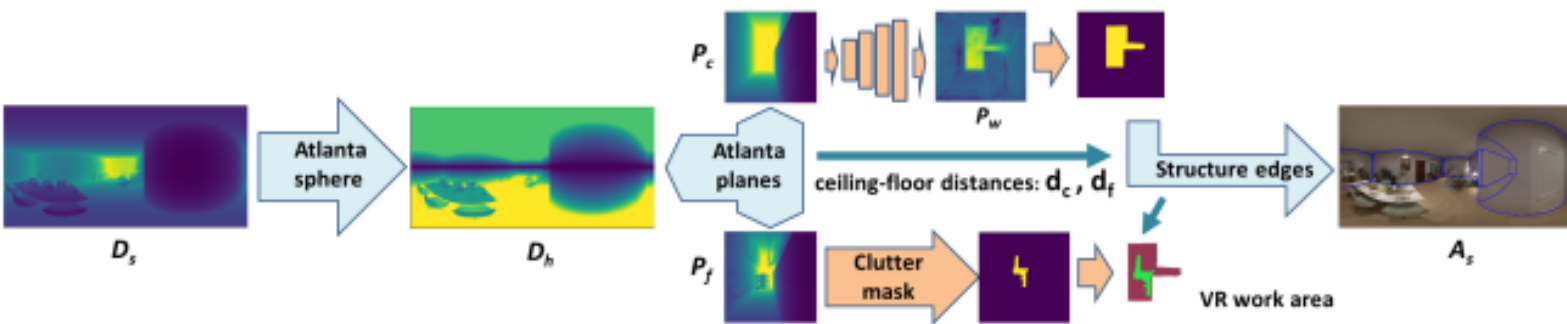
- 5 ResNet layers -> latest 4 processed by GAF encoder
- GAF compression through a (2,1) stride convolution x4
 - Cat and flattening into a 2048-length sequence
- Sequence encoding enables MHSA transformer
 - Avoid arbitrary patch-based encoding
 - Fast and simple decoding: (1,2) stride convolution + upsampling



$$A = \text{softmax} \left(\frac{(LW_q)(LW_k)^T}{\sqrt{l}} \right)$$

Depth and layout estimation example 3/4

- Euclidean depth -> Atlanta depth
- Atlanta depth -> ceiling and floor projection
- Ceiling-floor projection -> room height h
- Ceiling projection (more clutter-free) -> UNet -> 2D room footprint -> 2D polygon
- 2D Polygon + h -> 3D room shape



$$L_{dl} = L_d(D_s) - L_{ss}(D_s) + L_l(P_w) + L_h(h)$$

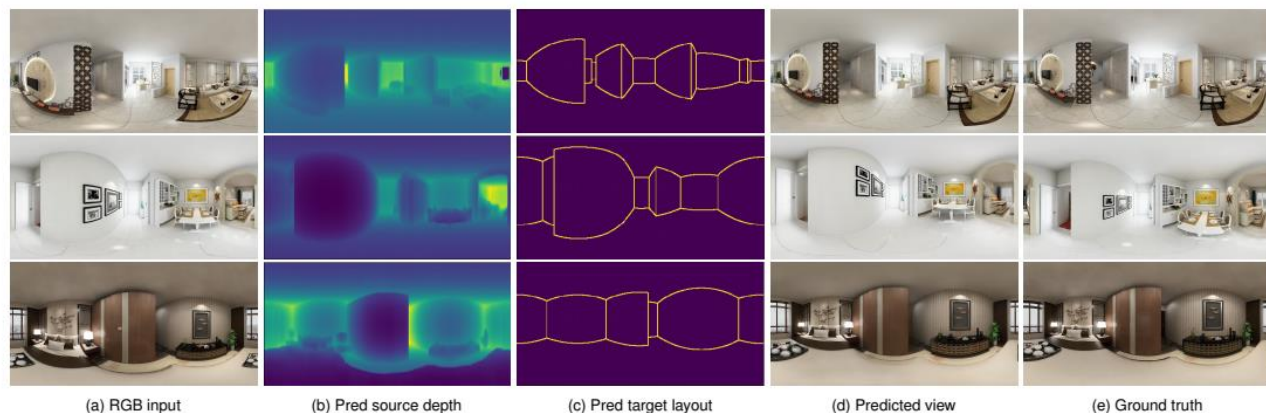
L_d : adaptive reverse Huber loss

L_{ss} : structural similarity index measure (SSIM)

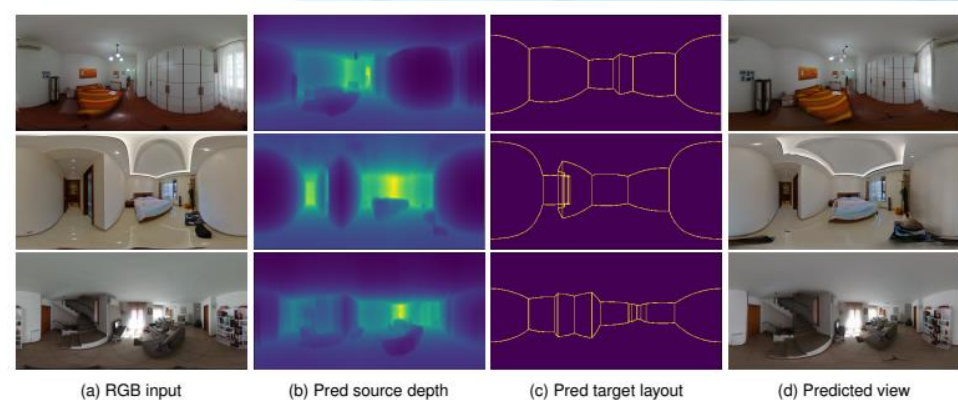
L_l : binary cross entropy with logits loss

L_h : L1 distance error for ceiling-floor distance

Depth and layout estimation example 4/4



Qualitative results on PNVS scenes



Qualitative results on user acquired scenes

Method	Params↓	GFLOPS↓	Output type
Bifuse [55]	253 M	682	only depth
SliceNet [37]	79 M	101	only depth
AtlantaNet [38]	100 M	273	only layout
ADM (our)	29 M	79	depth+layout

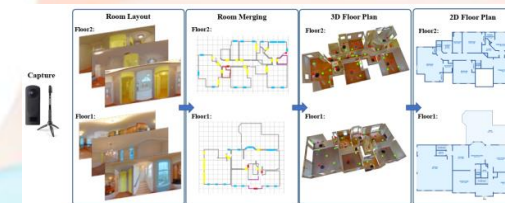
Table 1: **Depth-layout estimation computational performance.** We show our computational performance compared to other specific state-of-the-art works for a 512×1024 image.

Method	mse↓	rmse↑	δ_1 ↑	iou3d↑	iou2d↑
Jin et al. [18]	0.103	0.666	0.91	-	-
SliceNet [37]	0.044	0.174	0.93	-	-
AtlantaNet [38]	-	-	-	82.45	85.78
AVN (Our)	0.008	0.043	0.96	84.56	88.86

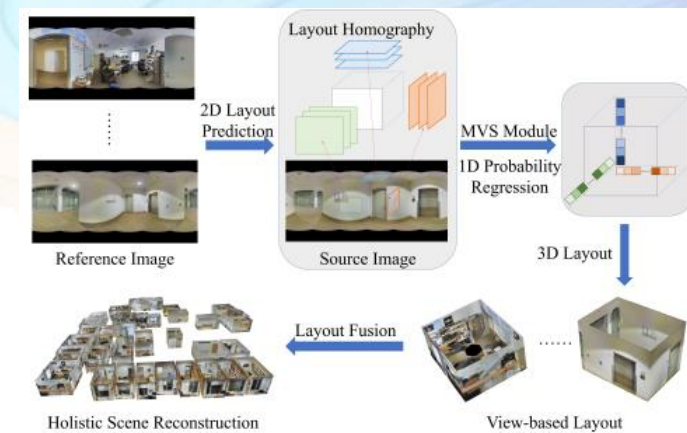
Table 3: **Depth and layout performance.** We show our quantitative performance compared to other representative state-of-the-art works.

Conclusions

- Single-view solutions are very challenging, but useful and effective
 - Minimal capture effort
- Multiple image solutions (next session) have different use cases
 - Cover a multi-room environment with many single-view captures
 - Get 3D info from (multi-view) stereo
 - Reduce noise, missing data, and ambiguities through increased coverage
- Single-view or multi-view model
 - 3D room model and/or pixel-wise information
 - Camera positions and/or multi-view features
 - Joined information from multiple images
 - At least depth maps
 - Pre-requisite: mutual registration of images
- Research topics
 - extreme SfM
 - Extreme reconstruction



ZInD – Cruz CVPR2021



MVlayoutNet – Hu ACM MM2022



Shabani et al. ICCV2021

NEXT SESSION: INTEGRATED INDOOR MODEL