Tutorial: Automatic 3D modeling of indoor structures from panoramic imagery

Giovanni Pintore[1], Marco Agus[2] and Enrico Gobbetti[1]

Visual Computing Group, CRS4, Italy[1]

College of Science and Engineering, HBKU, Qatar[2]

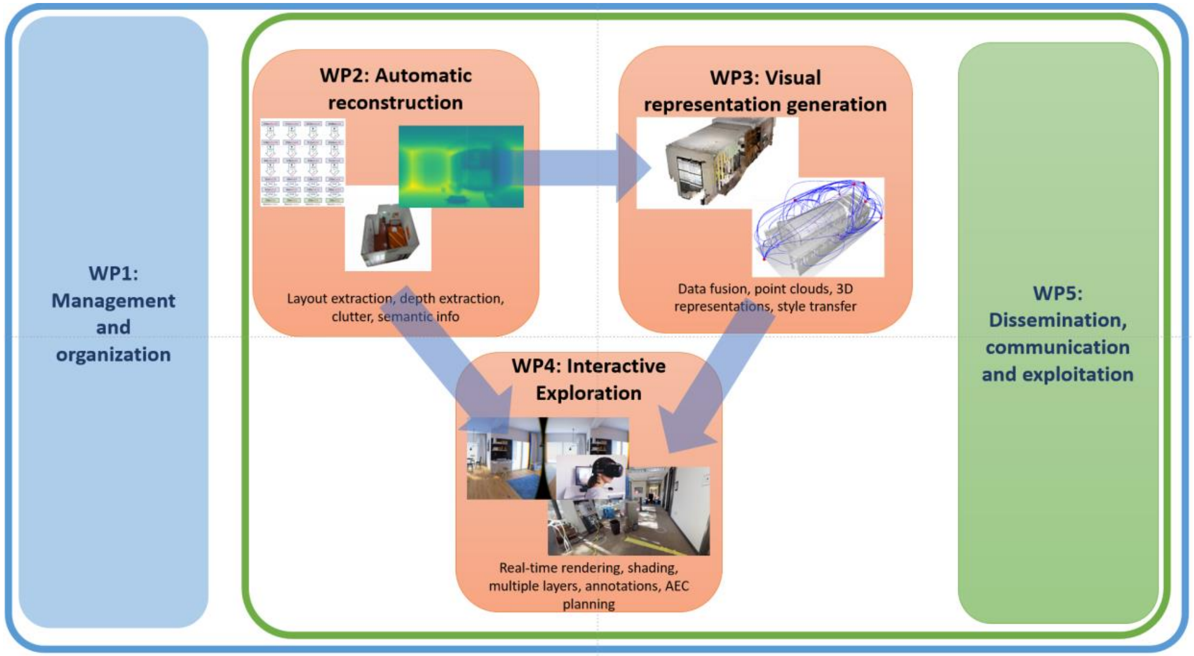
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

SESSION 5: VISUAL REPRESENTATION GENERATION AND EXPLORATION

Speaker: Marco Agus

AIN2: Artificial Intelligence for Indoor Digital Twins

- Qatar National Research Fund: NPRP14S-0403-210132
- Start date: 11/2022, End date: 11/2025
- Partners
  - Hamad Bin Khalifa University
  - CRS4
  - Qatar University
  - GHD Qatar

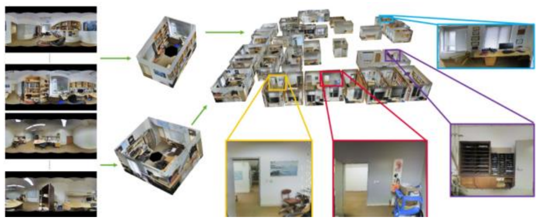


Main objectives:
Data-driven solutions for augmenting panoramic images of indoor Environments, Interactive and immersive solutions for exploring and editing indoor representations

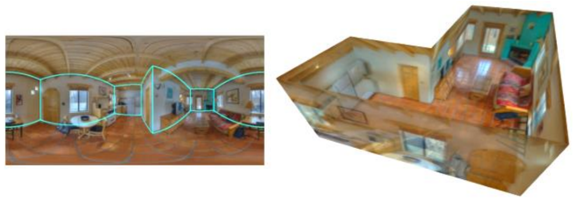
Speaker: Marco Agus

Introduction

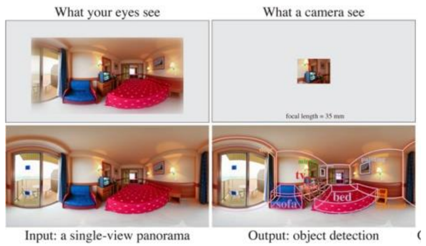
- Input:
  - Images associated with the room
    - Spatially referenced
  - 3D room model or pixel-wise information
    - Single scene
    - Walls, ceilings, floor
    - Multi-modal information for specific tasks
- Output:
  - Editable representations
  - VR exploration, Extended Reality, Editing appearance



MVlayoutNet – Hu ACM MM2022



HorizonNet – Sun CVPR2019



Zhang et al. ECCV 2014

Application context

- **Omnidirectional imagery**
  - Fundamental component for creating immersive content from real-world scenes
- **Virtual tour popular in the real-estate domain**
  - Presentation to virtual visitors
  - Popularized during Covid pandemic
- **Other application domains:**
  - Tourism, architecture, construction

https://matterport.com/industries/real-estate

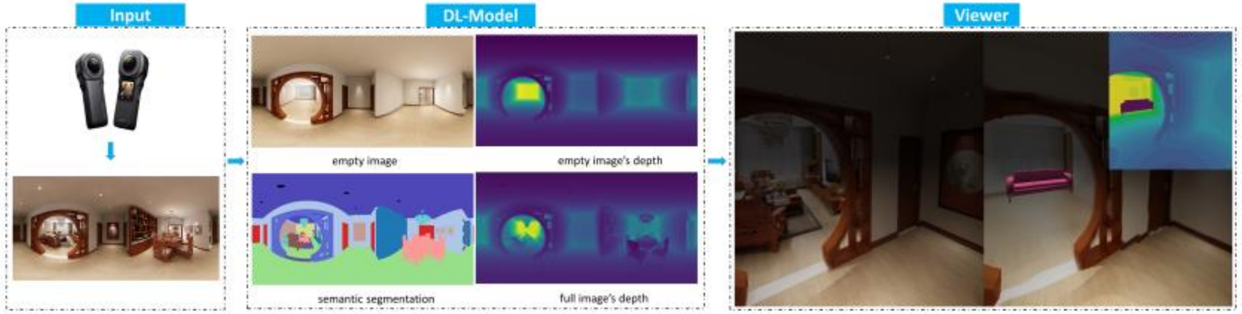
The Met 360° Project | The Metropolitan Museum of Art

Visit

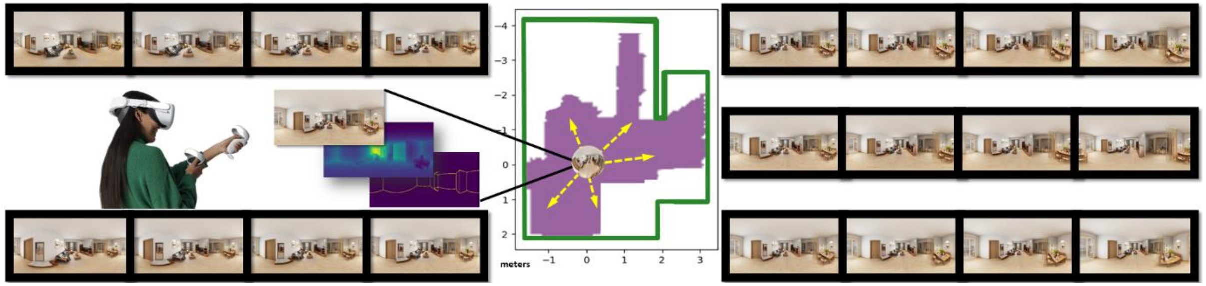
Images may be subject to copyright. **Learn More**

Outline (1/2)

- Overview of SOTA and our recent contributions related to two main tasks related to panoramic indoor scenes
- Interactive and immersive exploration
  - Integration of deep learning models in a rendering framework (Tukur et al., Spider, 2023, Elsevier GMOD)
  - 3-DOF view-synthesis for 6-DOF immersive exploration of indoor AtlantaWorld panoramic scenes (Work in progress)



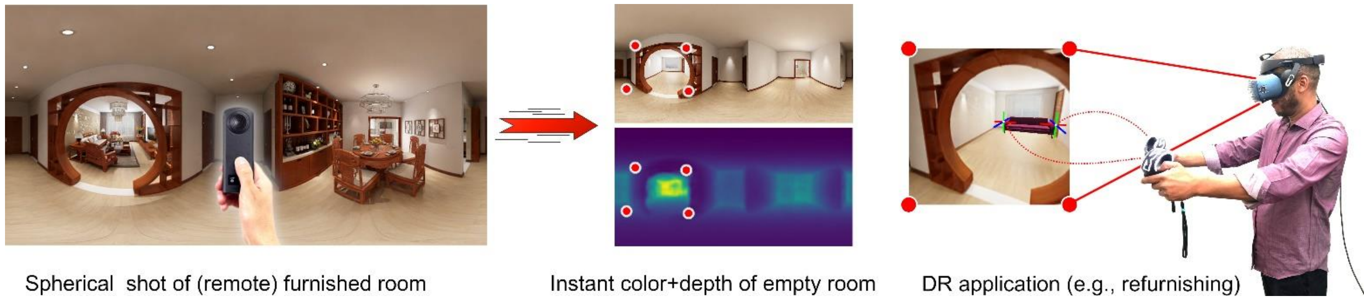
Tukur et al. , GMOD 2023



Work in progress

Outline (2/2)

- Overview of SOTA and our recent contributions related to two main tasks related to panoramic indoor scenes
- Scene modification and editing
  - Instant removal of clutter for diminished reality (Pintore et al., 2022, IEEE TVCG)
  - Photorealistic style transfer between indoor panoramic scenes (Work in progress)



Pintore et al. , IEEE TVCG 2022

Work in progress

IMMERSIVE EXPLORATION

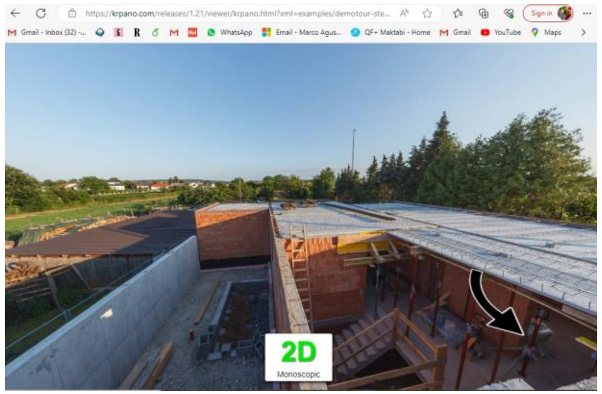
Main tasks (1/2): immersive exploration

- Support interaction and immersivity
- Desktop, mobile, XR setups
- Pano or Sphere Viewers
- 3D geometric representations
  - Textured domes, cubemaps, point clouds, tessellated meshes
- Enriched image representations for view synthesis
  - Multi-planar images (MPI)
  - Neural Radiance Fields (NERF)

Pano, Omni, Sphere viewers

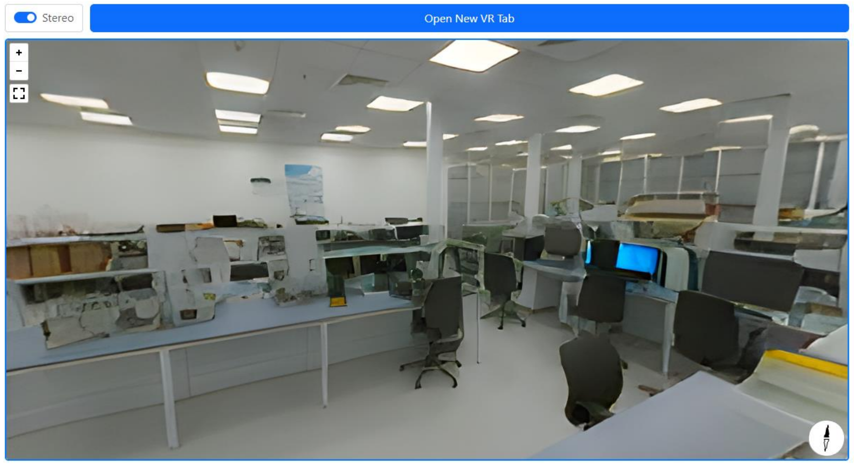
- Available online and using various representations
- Integration with WebVR and WebXR for direct usage with VR devices
- Cubemaps (krpano)
- Stereo panoramic couples (sphere stereo viewer)
  - A-frame

Pannellum.org



krpano

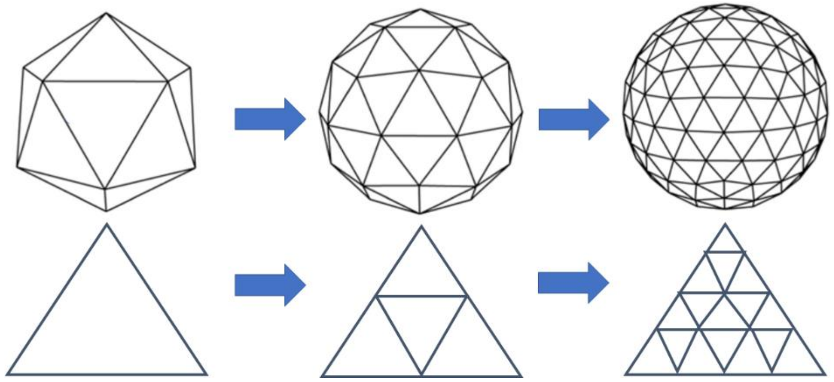
CSE-LAB-ICT



360° | CSE-LAB-ICT (renderstuff.com)

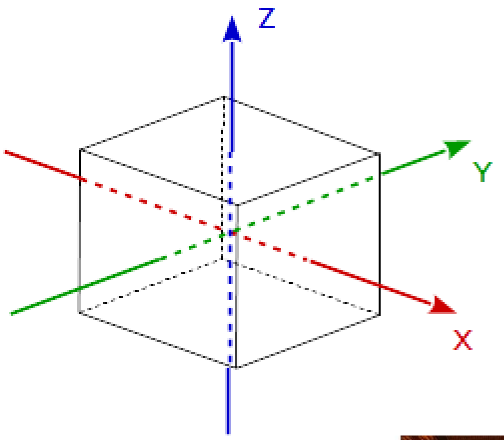
Geometric representation: mesh-based rendering

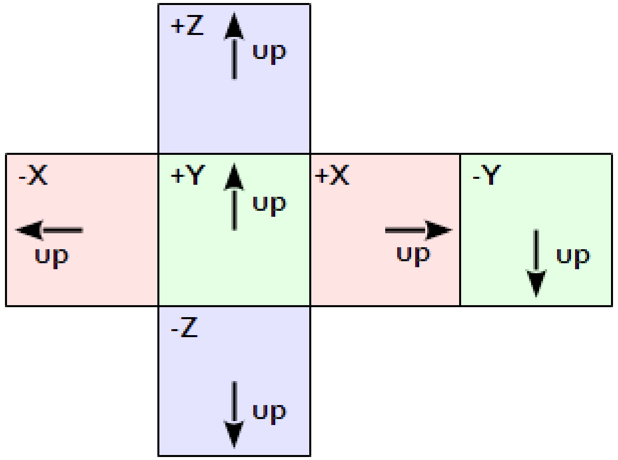
- Spherical dome tessellation
  - Iterative subdivision from icosahedron
  - Subdivision level 8 leads to ~1.3M verts and ~1.3 M triangles
- Basic rendering mode for original images and head rotation movements (viewer in the camera position)



Geometric representation: cubemaps

- From equirectangular image to six textures to be mapped to the faces of a cube
- Graphics hardware accelerates texture fetching in shaders (GL_TEXTURE_CUBEMAP)
  - Popular for environment maps in games
- Used in popular panoramic image viewer like krPano





Courtesy: paulbourke.net

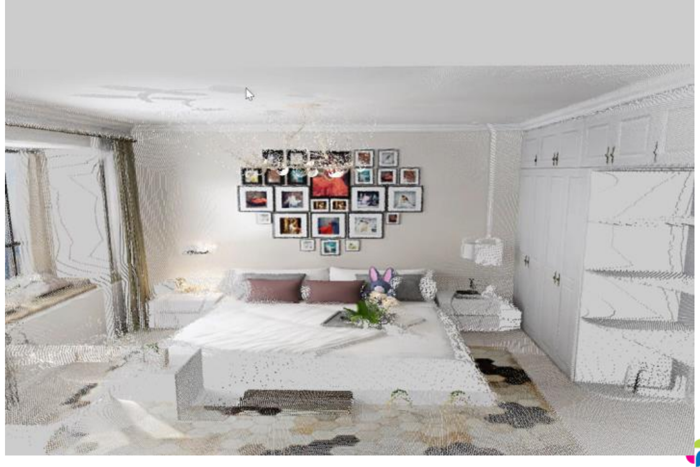
Geometric representation: depth integration

- Possible signal integration: depth, normal maps, semantic labelling
- Depth: 16-bit resolution mm scale, distance range from 0 to ~65.5m
- Geometry shader: fetch depth from texture and move dome vertices
  - Polygon rendering or Point Clouds

Without depth

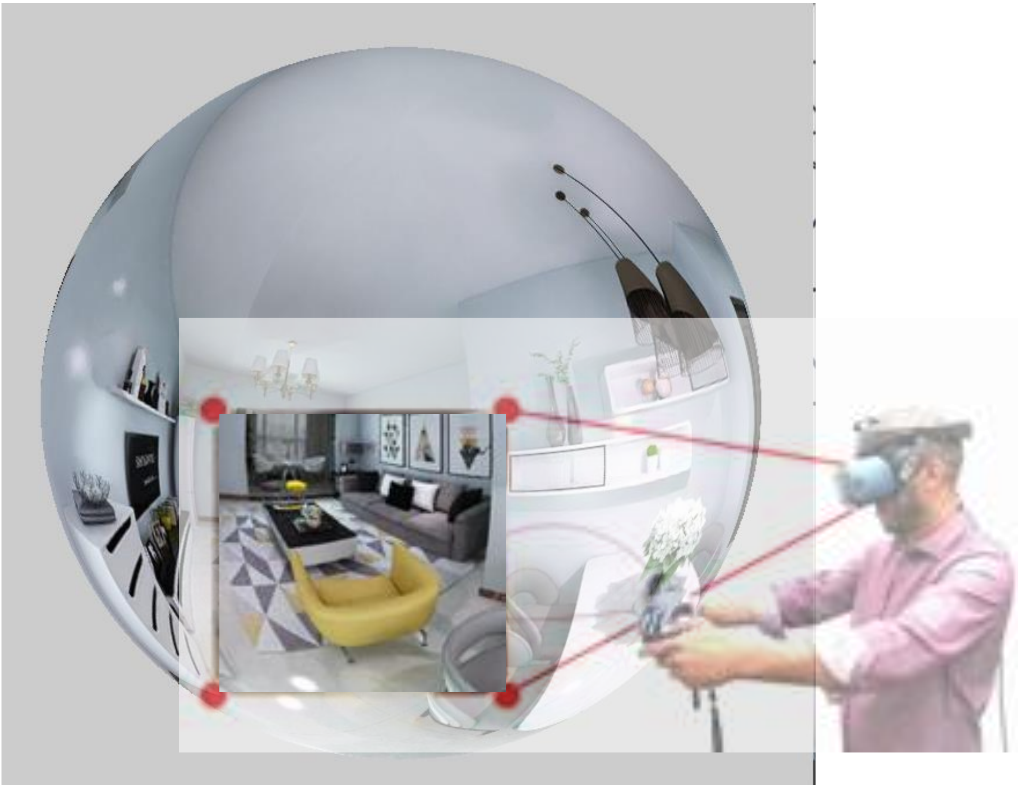
With depth

Point cloud



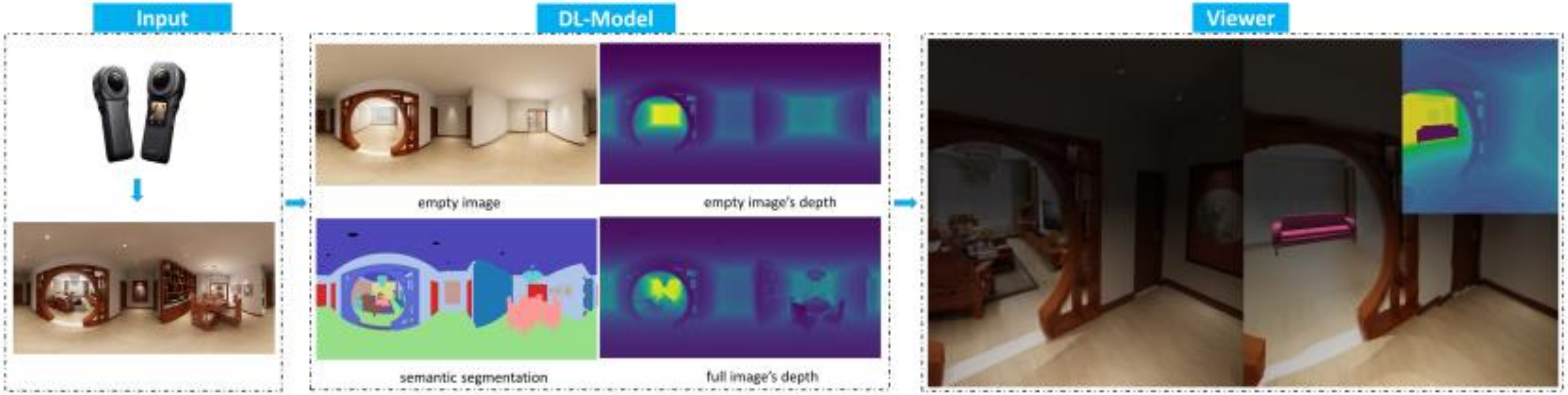
Fast rendering: single-pass ray casting

- Draw a quad in screen coordinates
- Fragment shader:
  - Pass view and perspective parameters: fov, distance of view plane
  - For each fragment:
    - cast a ray from eye position to intersect the spherical scene
    - fetch the corresponding texels from equirectangular images through inverse spherical mapping



Our contribution: AI-integrated rendering

- An interactive editing and rendering system for indoor DR/XR applications from a single panoramic image

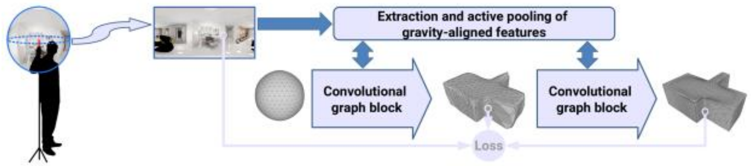


Tukur et al. Spider, Elsevier GMOD 2023

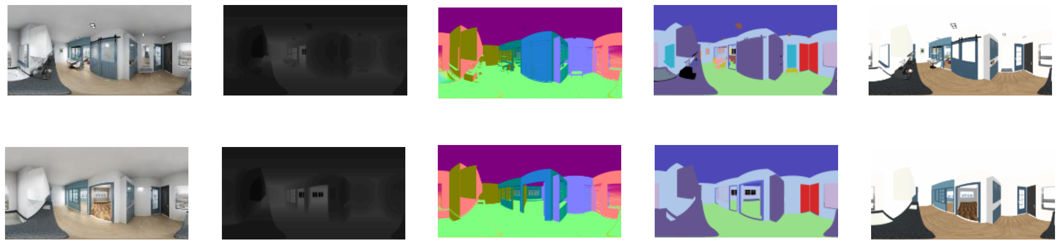
Speaker: Marco Agus

Full model creation

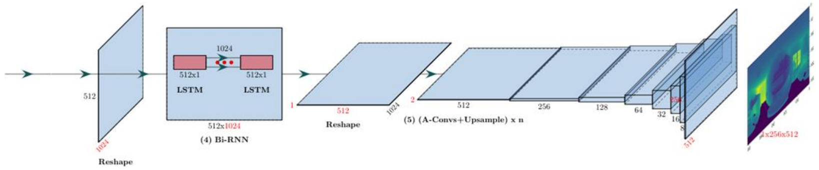
- **Inspired by SoA baselines**
- **Geometric structure**
  - Spherical deformation
    - TOG 2021
- **Pixel-wise signals**
  - Large scale synthetic data
    - ECCV 2020
  - Spherical features compression
    - CVPR 2021



Pintore et al. Siggraph Asia 2021



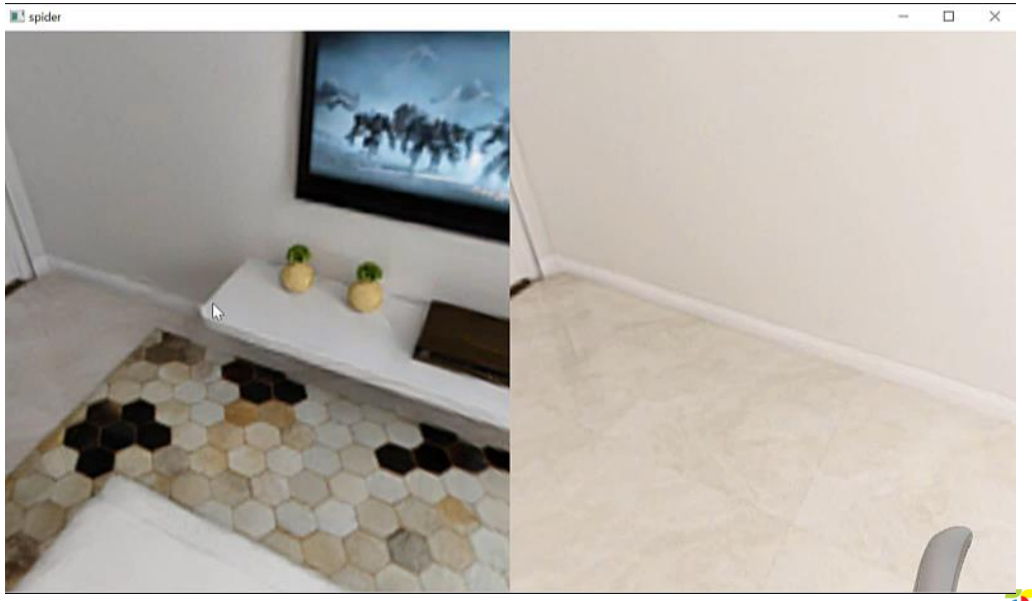
Strctured3D ECCV 2020



Pintore et al. CVPR 2021

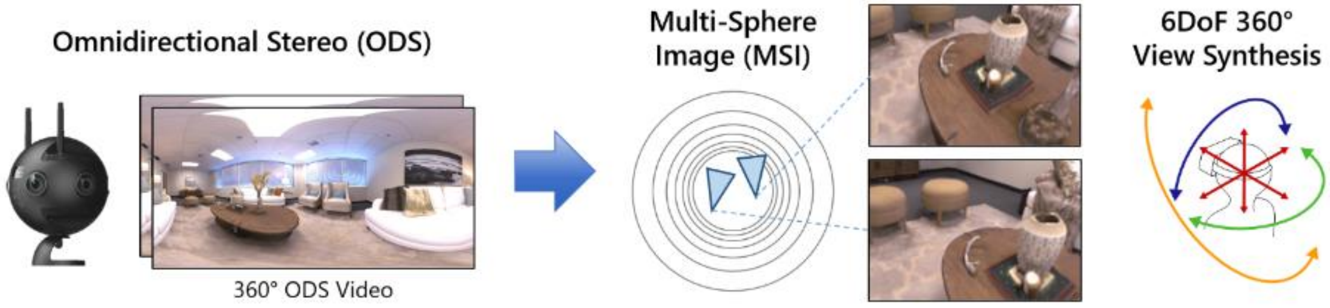
Applications

- Basic operations for Virtual Staging
  - Placement of synthetic objects
  - Transfer of semantic content from cluttered scene to empty scene

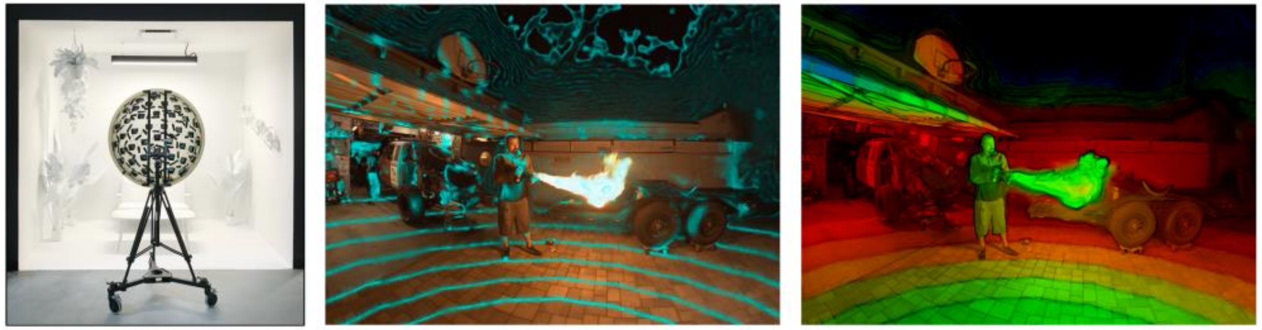


View synthesis: image-based methods

- Multi-spherical images
  - Extension of Multi-planar images for spherical shells (Attal et al, 2020)
  - Conversion to layered mesh representation (Broxton et al., 2020)



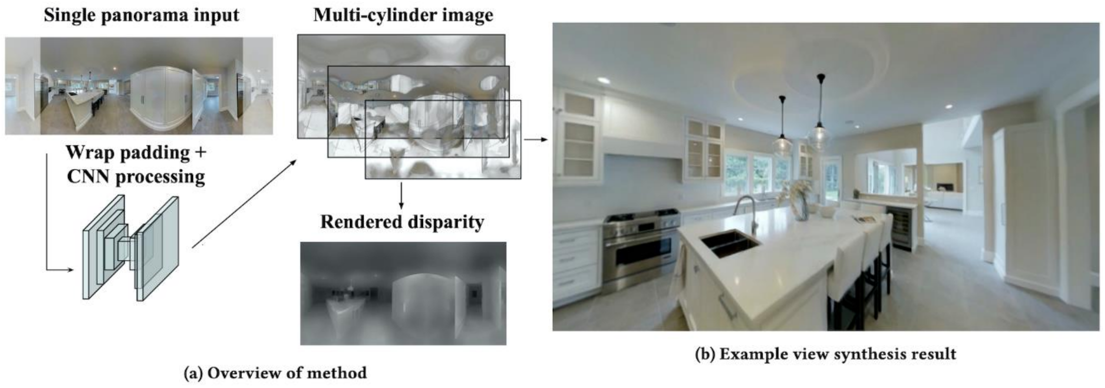
Attal et al. , Matryodska, ECCV 2020


(a) Capture Rig
(b) Multi-Sphere Image
(c) Layered Mesh Representation

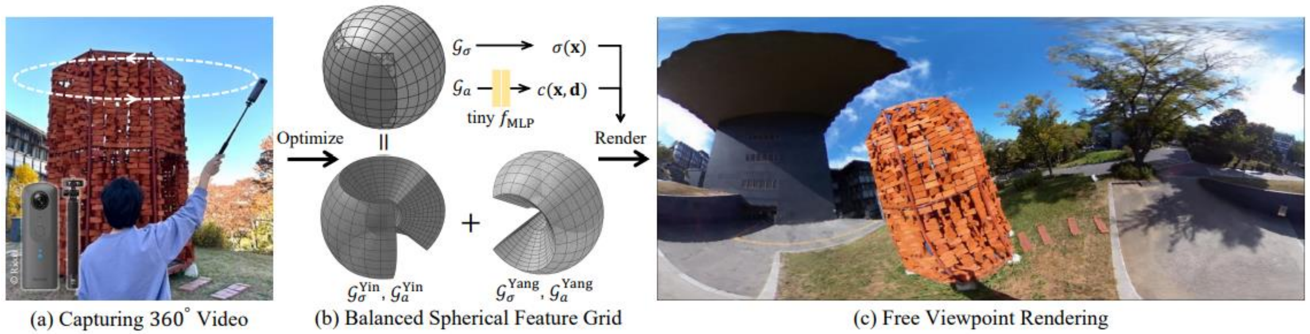
Broxton et al. ACM TOG 2020

View synthesis: image-based methods

- Multi-cylinder image
  - Representation of multi-plane-image on cylindrical proxy
  - PanoSynthVr (Waldhofer et al, 2022)
- Neural Radiance Field
  - Extension to spherical grids with conversion to spherical coordinates
  - EgoNerf (Choi et al, CVPR 2023)



Waidhofer et al. , PanoSynthVr, ISMAR 2022



Choi et al. , EgoNerf, CVPR 2023

Work in progress: GAN-based view synthesis

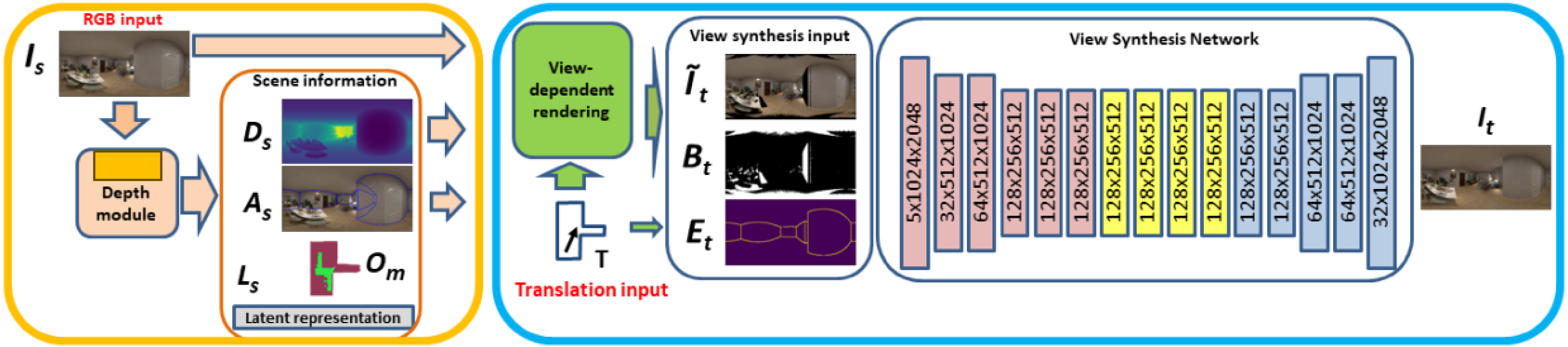
- State of the art systems need complicated setup for acquisition or videos with coherent information
- Explicit or implicit geometry estimation, to perform occlusion-aware reprojection and synthesize the disoccluded content
  - Complex training and inference
- Low-latency extraction of novel poses to extract perspective images in real-time responding to both translation and rotation

Key ideas

- Client-server architecture
  - Thin WebGL client manages head motion
  - Server computes images for head translation
  - 70 Hz refresh, 10 fps panorama updates, workspace ~30 cm
- Novel views synthesis respecting Atlanta World model constraints
  - Model exploits Gravity Aligned Features and LSTM for managing spatial relationships
  - Depth and layout prediction for constraining view synthesis

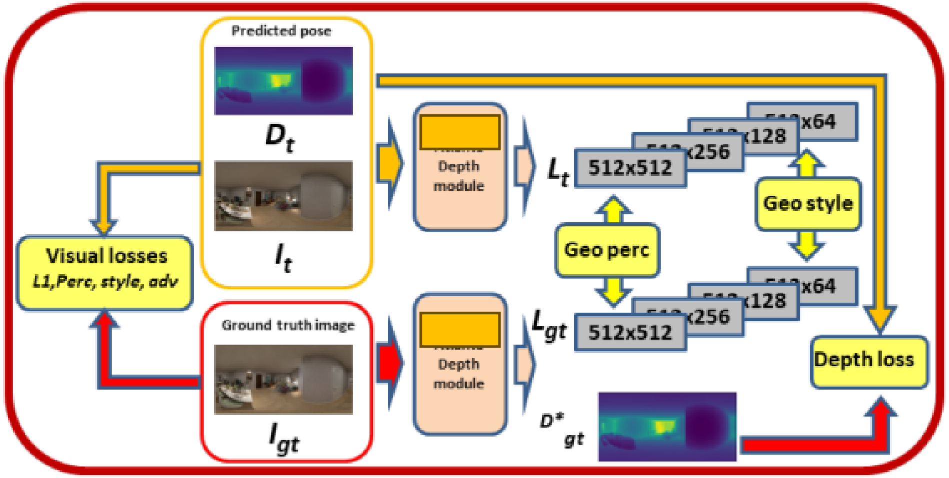
Forward pipeline

- **Signal extraction module:** concurrent estimation of scene depth, scene latent representation, 3D room shape and floor occupancy map
- **View Synthesis module:** lightweight approach to generate novel panoramic views
  - Limited number of layers, combining gated and dilated convolutions



Training stage

- Objective functions for indoor structural consistency
  - Design of losses based on direct estimation and latent-space features
  - Geometric perceptual and geometric style loss

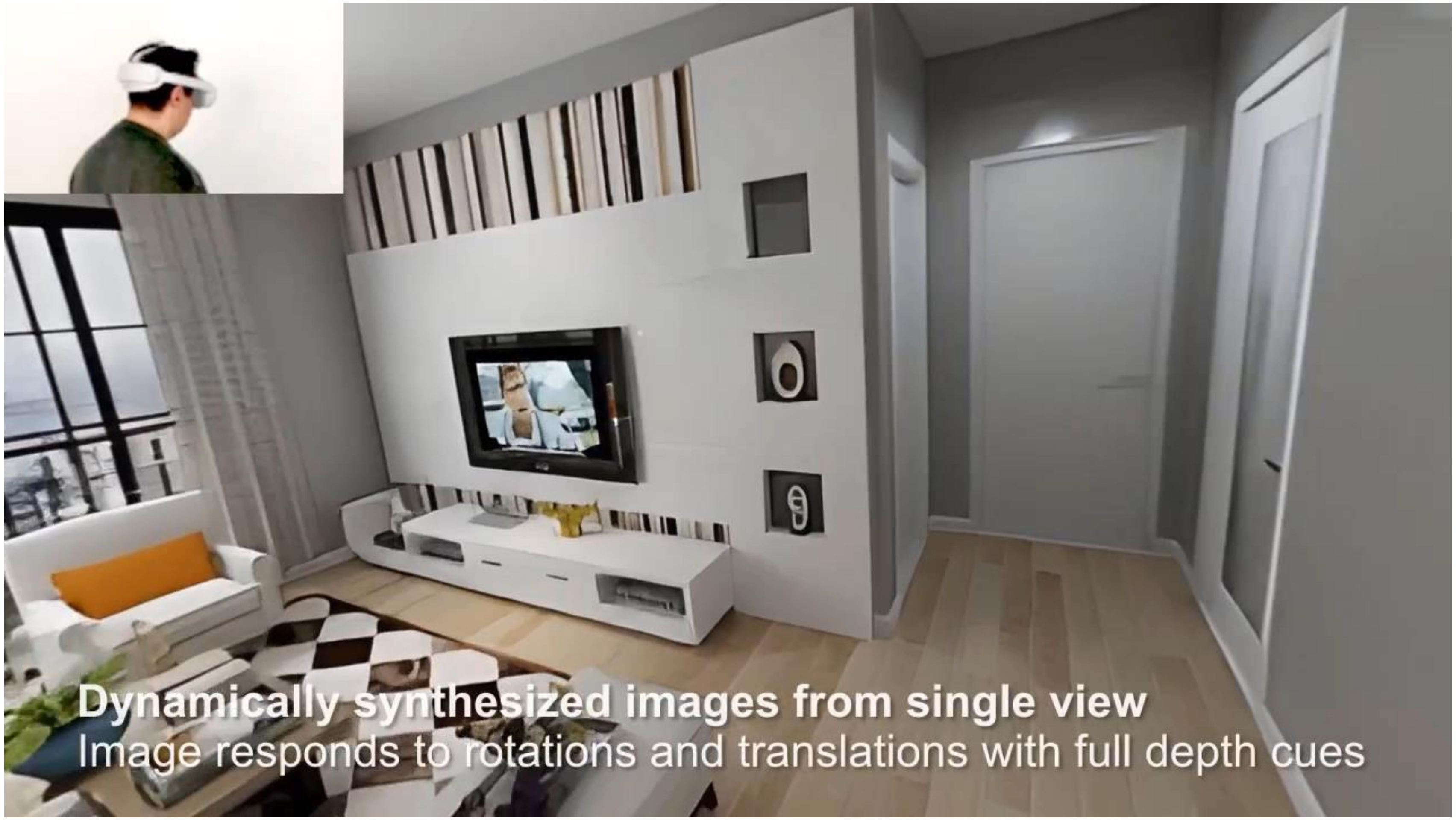


$$\mathcal{L}_{adm} = \lambda_d \mathcal{L}_d - \lambda_{ss} \mathcal{L}_{ss} + \lambda_l \mathcal{L}_l + \lambda_h \mathcal{L}_h$$

$$\mathcal{L}_{geocont} = \sum_{n}^{4} \left\| L_n(I_t) - L_n(I_{gt}) \right\|_1$$

$$\mathcal{L}_{geostyle} = \sum_{n}^{4} \left\| K_n(L_n(I_t)^T L_n(I_t)) - L_n(I_{gt})^T L_n(I_{gt}) \right\|_1$$

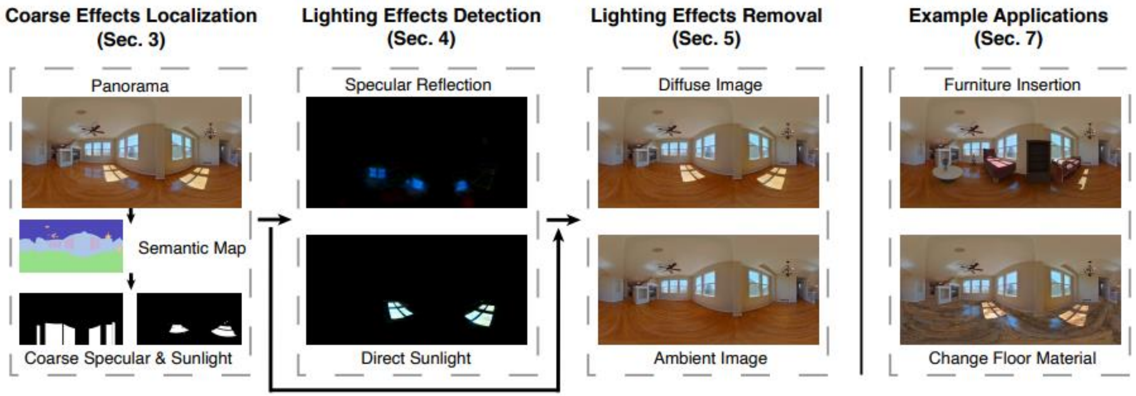
Preliminary results



SCENE MODIFICATION

Main tasks (2/2): scene modification

- Support editing and modifications
  - Adding/removing clutter/objects
  - Place POI/annotations
  - 3D multimedia hyperlinks
- Appearance modification
  - Lighting (Zhi et al, ACM TOG 2022)
  - Material (Work in progress)
- Virtual staging as emerging field



Zhi et al. , ACM TOG 2022

Our contribution: Diminished Reality

- Instant photorealistic view and depth of a panoramic indoor scene emptied of furniture and clutter
- Enables compelling and immersive XR applications, such as re-furnishing or planning of interior spaces

Pintore et al. IEEE TVCG 2022

Our contribution: diminished reality

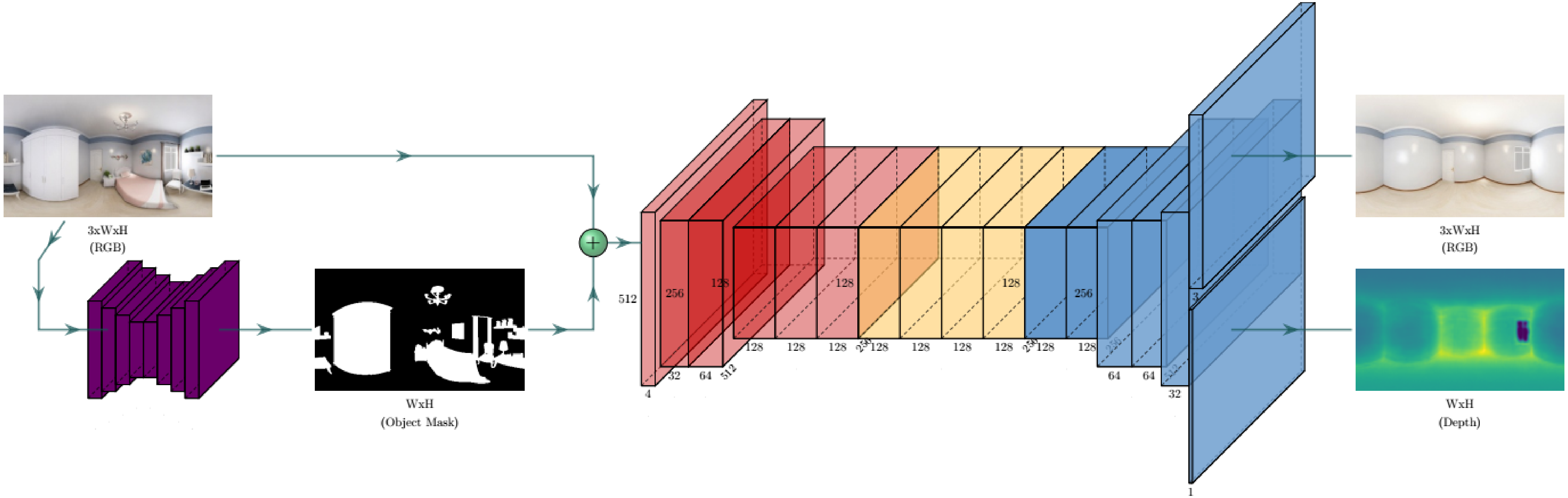
- **Light-weight end-to-end deep network**
  - Input: 360 image of a furnished indoor space
  - Output: 360 photorealistic view and architecturally plausible depth of the same scene emptied
    - Very low latency
  - NB. Learning on synthetic dataset transferred to real-world cases

Pintore et al. IEEE TVCG 2022

Key contributions

- **End-to-end network providing, at interactive rate, a panoramic indoor scene emptied automatically without user intervention**
  - Linear fashion and depth-separable gating
  - Visual and geometric constraints are applied only at training time
- **Geometric representation of the scene as additional output**
  - Basis for further processing in XR application
  - Enables robust and effective pixel-wise geometric priors
- **Loss function that combines photorealistic and geometric terms**
  - Virtual normals to recover the salient characteristics of indoor structures
  - Flatness and smoothness, less restrictive than Manhattan World, etc.

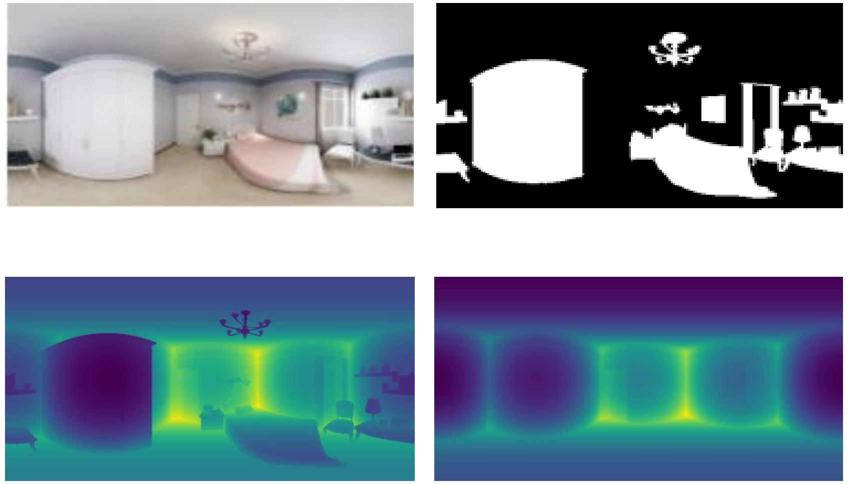
Model architecture



Pintore et al. IEEE TVCG 2022

Methods

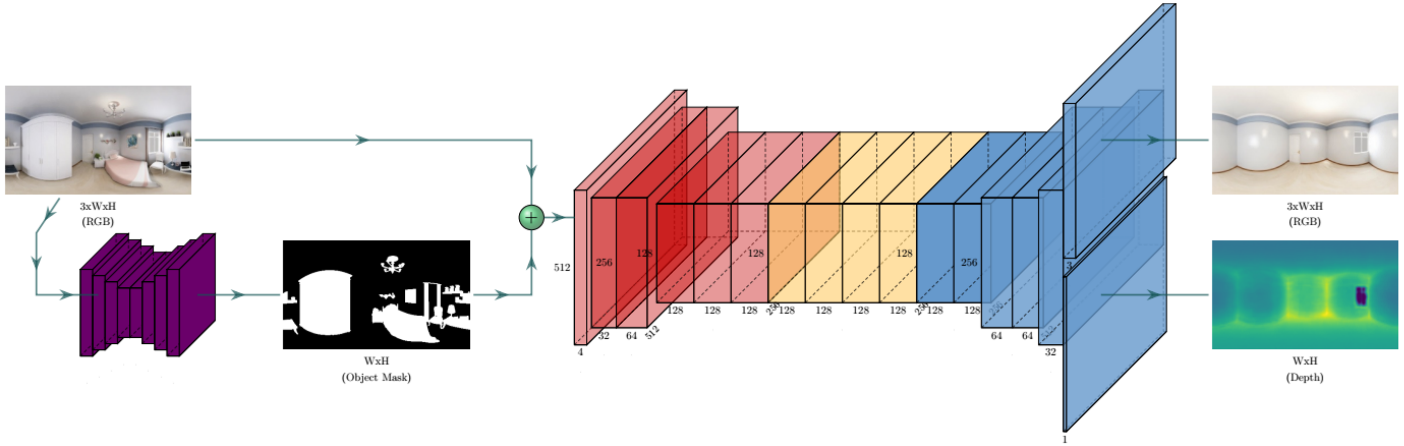
- **Clutter identification**
  - Automatic binary mask
  - Geometric mask obtained by comparing the ground-truth depths
  - Very lightweight encoder-decoder network
  - Binary cross-entropy loss



$$-\frac{1}{n}\sum_{p\in D_m{}^c}(\hat{p}\log p+(1-\hat{p})\log(1-p))$$

Methods

- **Empty scene synthesis**
  - Image inpainting baseline
    - Learnable gating
  - Light Weight Gated Convolutions (LWGC)
    - simplify training
    - low latency at inference time
  - Repeated dilations used for the bottleneck
    - Aggregates multi-scale contextual information without losing resolution
    - Avoid increasing number of weights



$$
\begin{aligned}
G &= conv(W_g, I) \\
F &= conv(W_f, I) \\
O &= \sigma(G) \odot \psi(F)
\end{aligned}
$$

$$D_{y,x} = \sigma(b + \sum_{i=-k'_h}^{k'_h} \sum_{i=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+\eta i,x+\eta j})$$

Methods

- **Training and losses**
  - Combination of a visual term and a geometric term
  - Visual term
    - L1 with data-driven perceptual and style losses
  - Geometric term
    - combination of low- and high-order 3D constraints
    - High-order based on virtual normal consistency

$$\mathcal{L}_{vis} = \lambda_{px}\mathcal{L}_{px} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style}$$

$$\mathcal{L}_{geom} = \lambda_d\mathcal{L}_d + \lambda_n\mathcal{L}_n$$

$$n_i = \frac{\overrightarrow{P_aP_b} \times \overrightarrow{P_aP_c}}{\left\|\overrightarrow{P_aP_b} \times \overrightarrow{P_aP_c}\right\|} \qquad \mathcal{L}_n = \frac{1}{N}\sum_{i=1}^{N}\left\|n_i{}^{pred} - n_i{}^{gt}\right\|$$

$$C = \{\alpha \geq \angle(\overrightarrow{P_aP_b}, \overrightarrow{P_aP_c} \leq \beta, \alpha \geq \angle(\overrightarrow{P_bP_c}, \overrightarrow{P_bP_a} \leq \beta\}$$

$$\mathcal{L}_{perc} = \sum_{n}^{N-1}\left\|\psi_n(I_{out}) - \psi_n(I_{gt})\right\|_1$$

$$\mathcal{L}_{style} = \sum_{n}^{N-1}\left\|K_n(\psi_n(I_{out})^T\psi_n(I_{out})) - \psi_n(I_{gt})^T\psi_n(I_{gt})\right\|_1$$

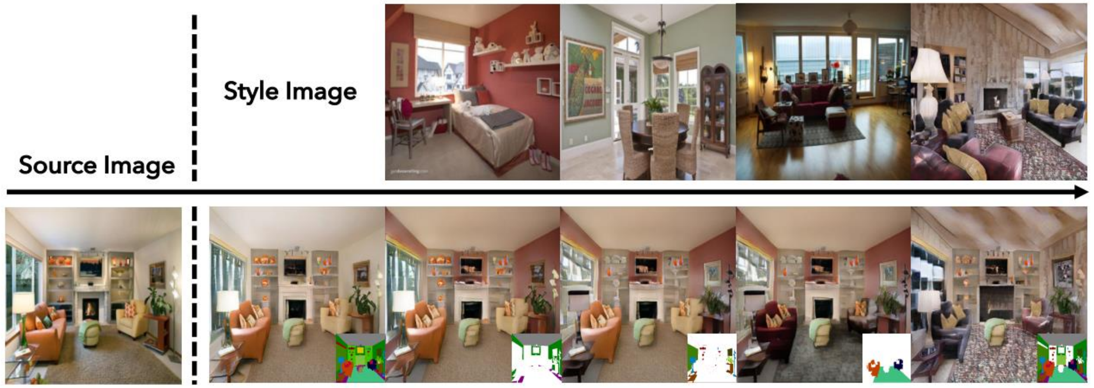
Some results

Work in progress: editing indoor panoramas

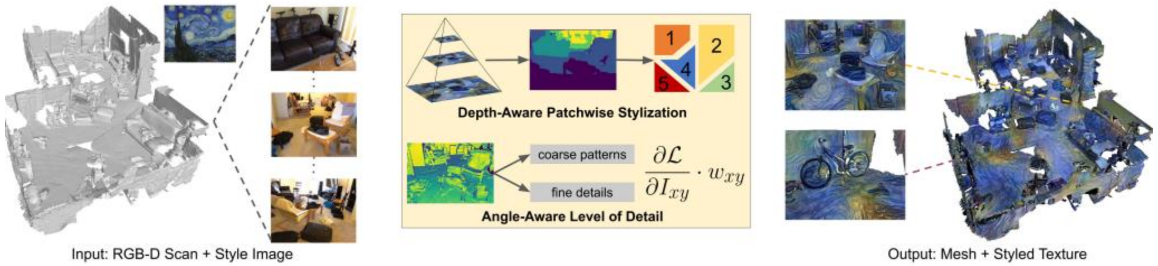
- GAN-based photorealistic style transfer

Motivation

- Current style transfer methods are not adequate for indoor panoramic images
    - lack of content preservation (SEAN, CVPR 2020)
    - Need of multiple poses and not photorealistic (StyleMesh, CVPR 2022)
- Specific complexity of indoor equirectangular images
    - High resolution requirements
    - Complex illumination patterns
    - Preservation of geometric characteristics
    - Equirectangular geometric distortion



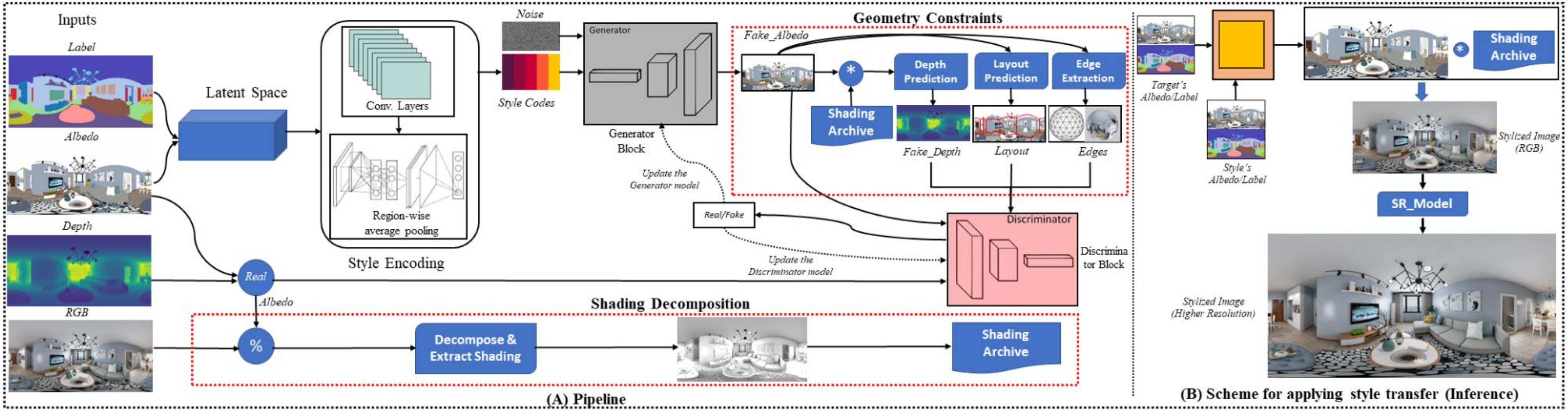
Zhu et al. SEAN, CVPR 2020



Hollein et al. StyleMesh, CVPR 2022

GAN-based photorealistic style transfer

- Two main additions on top of a classical GAN-based style transfer architecture:
  - Shading decomposition
  - Geometry constraints



Intrinsic shading decomposition

- Normalized shading signal for removing secondary effects

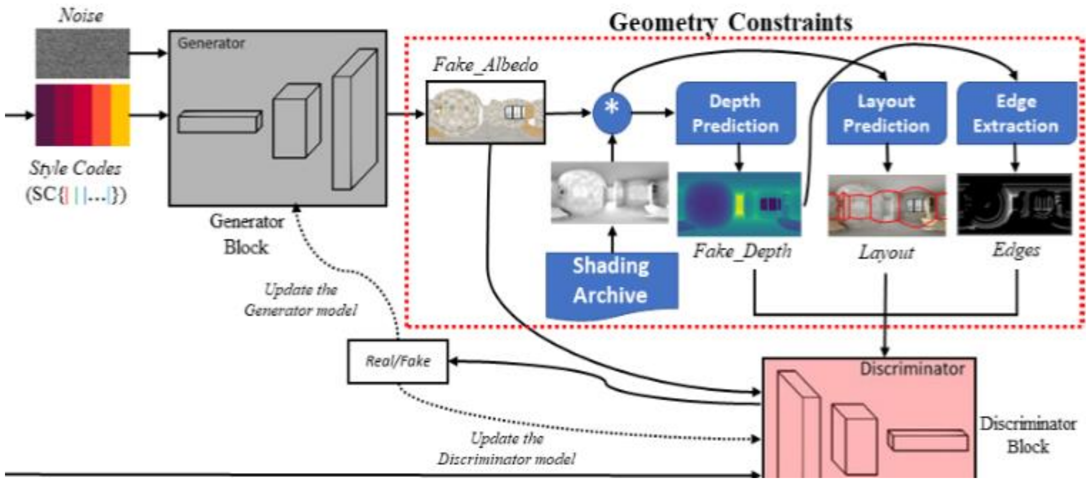
$$I_{\text{shad}} := \max\left(\left\|I_{\text{rgb}} \oslash I_{\text{alb}}\right\|_2, 1\right) \quad \Rightarrow \quad \hat{I}_{\text{rgb}} = I_{\text{shad}} \cdot I_{\text{alb}}$$

- Style codes computed on albedo and shading a-posteriori

Original RGB | Euclidean image difference SSIM/W-PSNR/PSNR 0.97/32.40/30.10 | Approximated RGB | Shading | Albedo

Geometry constraints

- Enforce scene depth, layout and edge consistency with additional geometry losses
  - For depth prediction, SliceNet [Pintore et al, 2021]
  - For layout prediction, HorizonNet [Sun et al., 2019]



$$\mathcal{L}_{\text{depth}}^{\text{geo}} = \sum_{ij} w_{ij} \left\| D_{ij}^{\mathrm{G}} - D_{ij}^{\mathrm{R}} \right\|_1$$

$$\mathcal{L}_{\text{depth}}^{\text{glob}} = \sum \left\| F_n(D^{\mathrm{G}}) - F_n(D^{\mathrm{R}}) \right\|_1$$

$$\mathcal{L}_{\text{depth}}^{\text{loc}} = \sum_{n} \left\| K_n \left(F_n\left(D^{\mathrm{G}}\right)^T F_n\left(D^{\mathrm{G}}\right) - F_n\left(D^{\mathrm{R}}\right)^T F_n\left(D^{\mathrm{R}}\right) \right) \right\|_1$$

$$\mathcal{L}_{\text{layout}}^{\text{geo}} = \left\| L^{\mathrm{G}} - L^{\mathrm{R}} \right\|_1,$$

$$\mathcal{L}_{\text{layout}}^{\text{glob}} = \sum_{n} \left\| H_n\left(L^{\mathrm{G}}\right) - H_n\left(L^{\mathrm{R}}\right) \right\|_1,$$

$$\mathcal{L}_{\text{layout}}^{\text{loc}} = \sum_{n} \left\| K_n \left(H_n\left(L^{\mathrm{G}}\right)^T H_n\left(L^{\mathrm{G}}\right) - H_n\left(L^{\mathrm{R}}\right)^T H_n\left(L^{\mathrm{R}}\right) \right) \right\|_1$$

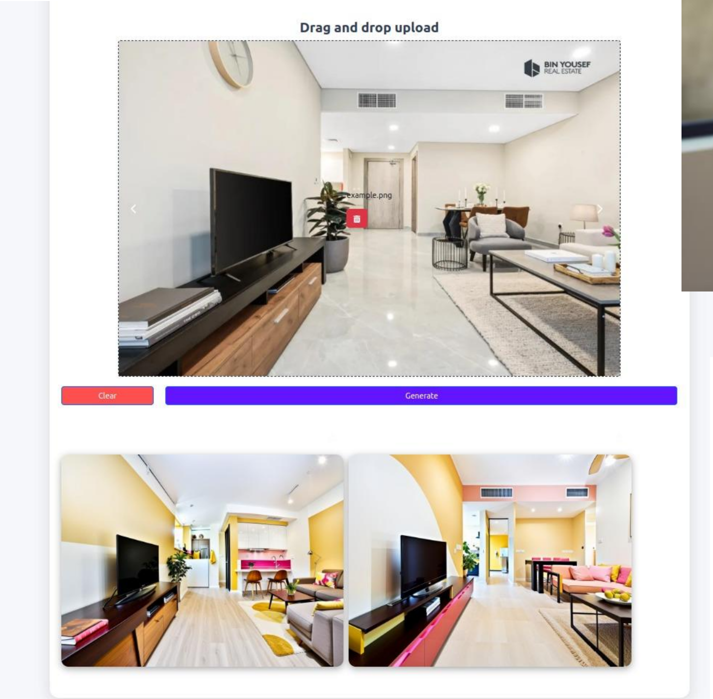
Preliminary results

Recap

- AI-based technologies for performing immersive exploration of scenes obtained through spherical imaging
- AI-based technologies for performing automatic modification of indoor environments
- Limitations:
  - Data hungry methods (rely on high-quality time-consuming data acquisition campaigns and processing)
    - We still mostly rely on synthetic datasets, like Structured3D
  - Resolution (most methods still work on 1024x512)
    - Partial workaround (usage of superresolution methods, like ESRGan or LAUNet)

Take-home messages

- The field is developing very fast
  - Thanks also to academic efforts
- Many challenges to address
  - Generalization to real-world scenarios
  - Increasing resolution
- Tech companies are investing huge resources
  - New solutions for XR
  - Automatic solutions for virtual staging



From HomeGPT.app, 2023

Apple VisionPro, 2023

Meta, Project ARIA

Hamad Bin Khalifa University

- Founded in 2010 (member of Qatar Foundation)
- College of Science and Engineering (founded in 2015)
- Mostly focused on graduate programs
- Focus on Qatar National Thematic Research

HBKU at a Glance

Number of Programs

36

Graduate Programs

35

Undergraduate Programs

1

Females Enrolled

55%

Males Enrolled

45%

Qatari Students

34%

Non-Qatari Students

66%

Nationalities

60+

Alumni

900+

Total Number of Employees

670+

Faculty

75+

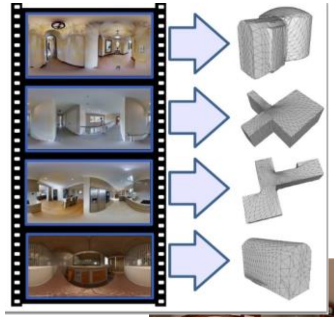
Researchers

350+

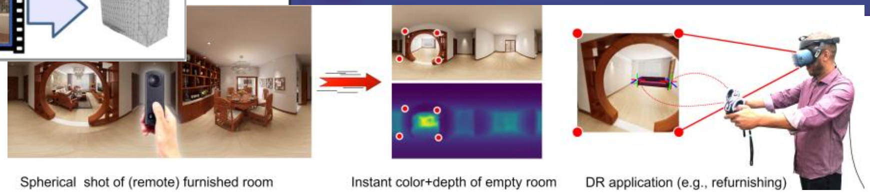
IDEALab - Interaction, Data Exploration, Accessibility

- Four faculties, 2 PostDoc, 8 Ph.D Students, ? Master Students
- Various research interests:
  - Interactive Visualization of complex data
  - Machine Learning applied to 2D/3D problems
  - Applications: medicine, biology, architecture, food computing, cultural heritage
  - Etc, etc.
  - We look for PostDocs and Ph.D. students

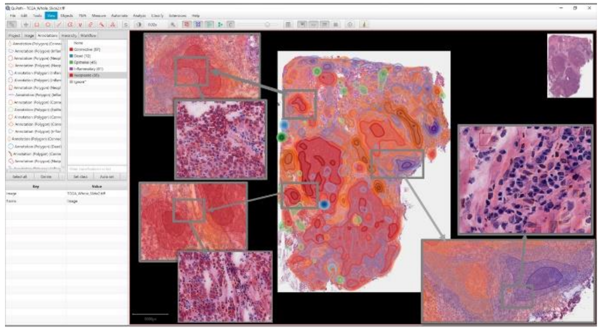
Deep3DLayout, ACM TOG 2021



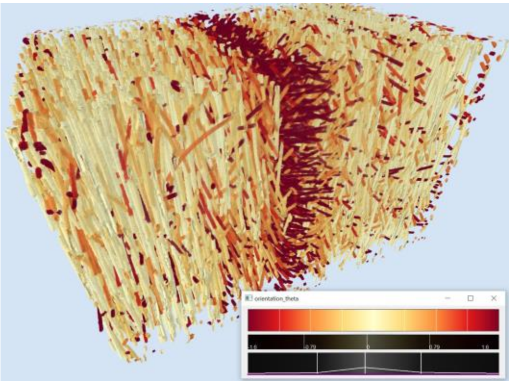
SliceNet, IEEE CVPR 2021



DR-EmptyRoom, IEEE TVCG 2022



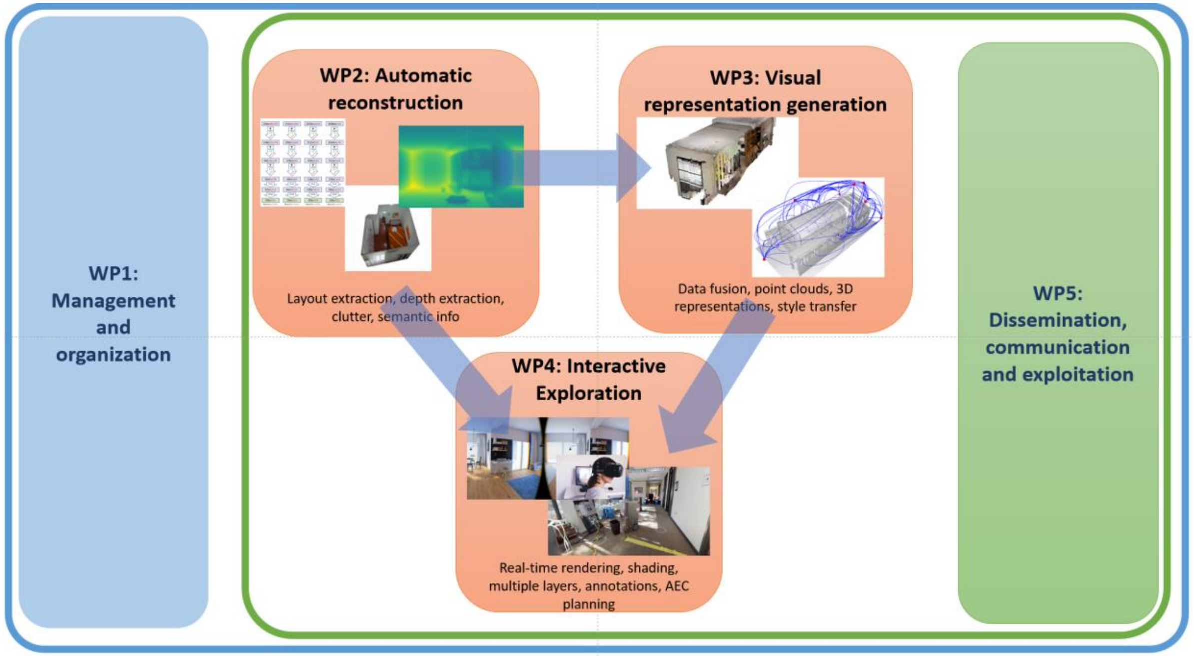
HistoContours, EG VCBM 2022, Best full paper



Mixture Graph, IEEE TVCG 2021
Volume Puzzle, IEEE VIS 2022 SP

AIN2: Artificial Intelligence for Indoor Digital Twins

- Qatar National Research Fund: NPRP14S-0403-210132
- Start date: 11/2022, End date: 11/2025
- Partners
  - Hamad Bin Khalifa University
  - CRS4
  - Qatar University
  - GHD Qatar



Main objectives:
Data-driven solutions for augmenting panoramic images of indoor Environments, Interactive and immersive solutions for exploring and editing indoor representations

Speaker: Marco Agus

SESSION6: CLOSING

Speaker: Enrico Gobbetti