

HouseLayout3D: A Benchmark and Training-Free Baseline for 3D Layout Estimation in the Wild

Valentin Bieri¹ Marie-Julie Rakotosaona² Keisuke Tateno²

Francis Engelmann³ Leonidas Guibas³

¹ETH Zurich ²Google ³Stanford University

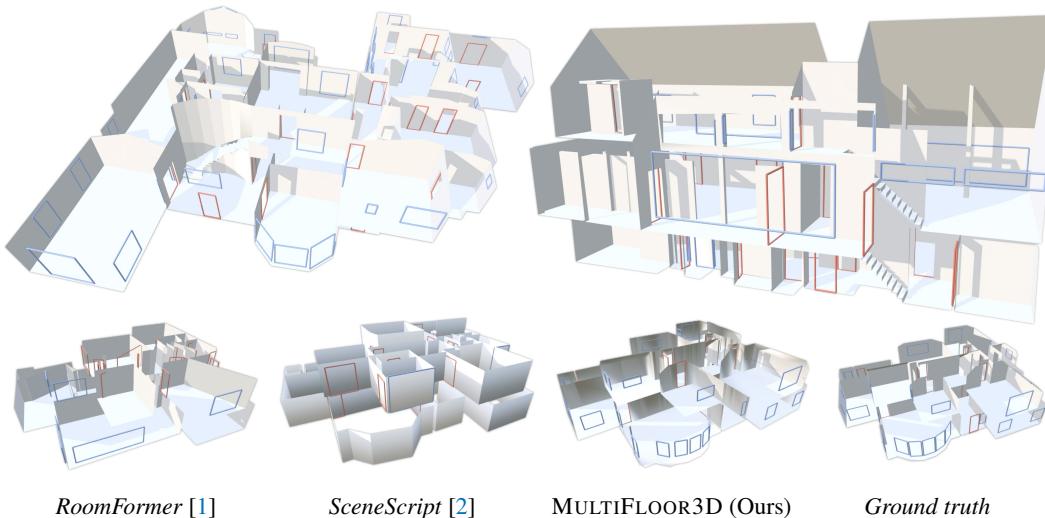


Figure 1: *Top*: We present HOUSELAYOUT3D, a benchmark for 3D house layout estimation with greater diversity than existing datasets, including multi floor buildings and detailed annotations for doors, windows, and staircases. *Bottom*: We introduce MULTILOOR3D, a training free approach for 3D layout estimation that improving over existing methods on both our benchmark and prior datasets.

Abstract

Current 3D layout estimation models are primarily trained on synthetic datasets containing simple single room or single floor environments. As a consequence, they cannot natively handle large multi floor buildings and require scenes to be split into individual floors before processing, which removes global spatial context that is essential for reasoning about structures such as staircases that connect multiple levels. In this work, we introduce HOUSELAYOUT3D, a real world benchmark designed to support progress toward full building scale layout estimation, including multiple floors and architecturally intricate spaces. We also present MULTILOOR3D, a simple training free baseline that leverages recent scene understanding methods and already outperforms existing 3D layout estimation models on both our benchmark and prior datasets, highlighting the need for further research in this direction. Data and code are available at: <https://houselayout3d.github.io>.

1 Introduction

Having spatial awareness of the surrounding 3D layout is a key requirement for many perception algorithms [1, 3–5] and robotic systems [6, 7]. The goal is to derive a concise vectorized layout by abstracting a 3D scene into a set of polygons that represent structural elements such as walls, floors, ceilings, staircases, doors, and windows, while filtering out furniture and other occluders commonly found in indoor environments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Recent state-of-the-art models for layout prediction [1, 2, 8] are feed-forward deep-learning models trained on large-scale synthetic datasets [2, 9] and demonstrate impressive results even on real-world scenes. A key limitation of current models is their reliance on synthetic training data, which predominantly contains single rooms or small apartments. Such data is appealing because it can be generated automatically at scale [10] or created in controlled settings [9], yet it lacks the complexity of large real buildings. As a result, models trained on these datasets struggle to generalize to buildings with many rooms and cannot handle multi-floor layouts. One workaround is to partition a building into individual floors or rooms, process each part independently, and then merge the results. However, this removes global context that is important for local reasoning, for example when identifying staircases that connect multiple floors, and it requires post hoc integration to support building-level tasks such as localization [11, 12] or scene-level reasoning [13].

To drive progress in 3D layout prediction for large-scale multi-floor buildings, we introduce HOUSE-LAYOUT3D, a benchmark built on real-world scans from Matterport3D [14]. The dataset contains architecturally complex buildings spanning as many as five floors and up to forty rooms per floor, including diverse room types and partially open spaces that remain difficult for current room-centric methods. We provide detailed manual annotations of structural elements such as walls, floors, ceilings, staircases, windows, and doors, including the opening direction of each door.

Inspired by recent progress in reconstruction and segmentation, we propose MULTIFLOOR3D, a training free approach for large scale 3D layout estimation. By combining modern 3D scene reconstruction with a layout fitting strategy, we show that a simple method can outperform existing approaches on the more challenging task of layout estimation in multi floor buildings.

Our experiments on HOUSELAYOUT3D reveal clear limitations of current state of the art methods in complex multi floor buildings. In contrast, our approach produces more accurate and coherent layouts, particularly in challenging multi level structures. We hope that these findings, together with our benchmark, will encourage further research in scalable multi floor 3D layout estimation.

In summary, our contributions are:

- We introduce HOUSELAYOUT3D, the first benchmark for 3D layout estimation in large scale multi floor buildings.
- We present MULTIFLOOR3D, a training free baseline leveraging modern reconstruction and segmentation models, achieving improved performance over existing deep learning approaches.
- Through extensive experiments we expose the challenges faced by current layout estimation methods, motivating progress in this problem setting.

2 Related Work

Manhattan Scene Layout. Early approaches to layout estimation commonly assume a Manhattan world and solve a constrained optimization problem using detected walls, as in Scan2Bim [15], or using detected corners, as in DuLaNet [16], LayoutNet [17], and FloorNet [18]. A notable exception by Ochmann *et al.* [19] relaxes the Manhattan constraint by subdividing the 3D space into cells and formulating layout estimation as an integer linear program, allowing angled walls.

2D Scene Layout. Early methods such as [20] infer 2D floorplans by computing shortest paths around free space, while Floor-SP [21] extends this idea with a room segmentation network. RoomFormer [1] estimates semantic floorplans with a transformer. HovSG [22] combines BEV point-density maps with 2D object detection to construct a scene graph of floors, rooms, and objects, but without recovering their 3D geometry. This class of approaches remains fundamentally limited to 2D predictions.

3D Scene Layout. End-to-end learning has driven recent progress in 3D layout estimation: SceneCAD [10] predicts layouts and object boxes using a graph network, and SceneScript [2] introduces a structured scene language for joint prediction of walls, openings, and objects. Compact scene abstraction has also emerged as a complementary direction, for example SuperDec [23], which decomposes indoor environments into superquadric primitives and highlights the value of structured, geometry-efficient representations. However, most available training data consists of single rooms [10] or simple one-floor layouts [2, 9, 24], often synthetic [2, 9], unfurnished [24], or restricted to Manhattan geometries [25]. Other datasets only capture partial scenes [26, 27]. This lack of diverse building-scale data limits generalization and leaves current models ill suited for complex multi-floor environments.

Dataset	Real-world	Multi-room	Multi-floor	Full Scenes	Windows, Doors	Objects	Depth	3D Layouts
SceneCAD [10]	✓	(✓)	✗	✓	✓	✓	✓	✓
ASE [2]	✗	✓	✗	✓	✓	✓	✓	✓
Stru3D [9]	✗	✓	(✓)	✓	✓	✓	✓	✓
Zillow Indoor [24]	✓	✓	(✓)	✓	✗	✗	✗	✗
MP3D-Layout [27]	✓	✗	✗	✗	✓	✓	✓	✓
Zou et al [25]	✓	✗	✗	✗	✗	✗	✓	✓
CADeState [26]	✓	✓	✗	✗	(✓)	✗	✗	✓
FloorNet [18]	✓	✓	✓	✓	(✓)	✗	✗	✗
HOUSELAYOUT3D (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Dataset Comparisons** of existing dataset benchmarks for evaluating 3D layouts estimation.

3 The HOUSELAYOUT3D Dataset

We introduce a new dataset of hand-annotated CAD layouts derived from the Matterport3D [14] (MP3D) dataset. Example annotations are shown in Fig. 2. Unlike prior works [2, 10], this is the first real-world dataset to provide CAD annotations for large-scale, multi-floor houses, encompassing numerous rooms, staircases, windows, and doors. Each structural element is annotated as a polygon in 3D space. Since our dataset is annotated on 3D meshes from MP3D [14], it inherits their per-vertex room ids and object instances.

Dataset Statistics. The dataset includes 16 buildings, 33 distinct levels, and 317 rooms, captured across more than 26,000 RGB-D frames. Its scale is comparable to the validation split of ScanNet [28]. In total, we annotated 292 doors, 379 windows, and 34 staircases. The lower number of doors compared to rooms is due to many connected by open passages or staircases rather than actual doors. Each building comprises between 1 and 5 levels and contains between 4 and 40 rooms. The annotation time varies depending on the building’s size and the number of rooms, typically ranging from 4 to 10 hours per building. All annotations undergo visual verification by separate expert annotators. Tab. 1 compares and summarizes properties across different datasets.

Annotation Tool and Labeling Details. To annotate the 3D scans, we use a free academic license of Scasca’s PinPoint [29], a specialized software for building modeling from point clouds. It enables precise 3D geometry extraction even in occluded or incomplete areas through intuitive tools that automatically snap to edges and corners, streamlining the annotation process. In the 3D scans, doors are typically open, so we annotate both the current open position and the expected closed position, along with the opening direction. For doors that appear closed in the scans, we infer the opening direction from the door hinge locations in the RGB images. For window annotations, we utilize the existing annotations from MP3D [14], projecting them onto the nearest annotated wall plane and fitting axis-aligned rectangles.

4 Method

Given N input RGB images of a scene, our goal is to recover a simple 3D layout represented by polygons. Each polygon is assigned a label from a fixed set of classes: wall, floor, ceiling, stairs, door, or window. The resulting layout is organized into a scene graph whose nodes correspond to rooms and whose edges correspond to doors or stairs, with each polygon associated either with a room node or with an edge in this graph.

Fig. 3 shows an overview of our four-stage approach. First, we reconstruct a 3D mesh of the scene. Second, we extract the main structural elements (floors, walls, ceilings) to form a *skeleton* of the layout. Third, we fit a layout *prototype* to this skeleton using geometric and semantic cues. Finally, we parse the prototype into a scene graph from which we derive the final layout.

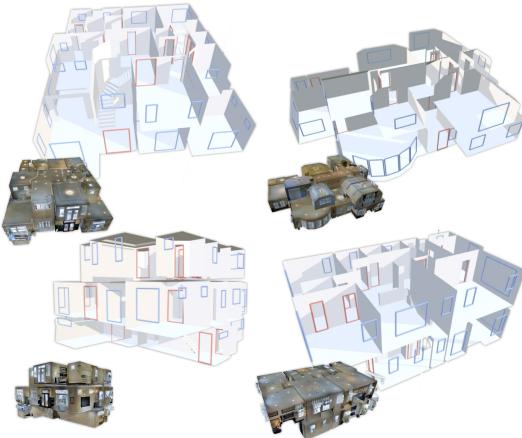


Figure 2: **Examples of our HOUSELAYOUT3D.** Our dataset includes multi-floor houses with annotations for walls, floors, ceilings and stairs, as well as windows (blue) and doors (red). We also show the corresponding 3D meshes from MP3D [14].

spaces, such as hallways and dining areas, being connected by open passages or staircases rather than actual doors. Each building comprises between 1 and 5 levels and contains between 4 and 40 rooms. The annotation time varies depending on the building’s size and the number of rooms, typically ranging from 4 to 10 hours per building.

All annotations undergo visual verification by separate expert annotators. Tab. 1 compares and summarizes properties across different datasets.

spaces, such as hallways and dining areas, being connected by open passages or staircases rather than actual doors. Each building comprises between 1 and 5 levels and contains between 4 and 40 rooms. The annotation time varies depending on the building’s size and the number of rooms, typically ranging from 4 to 10 hours per building.

All annotations undergo visual verification by separate expert annotators. Tab. 1 compares and summarizes properties across different datasets.

spaces, such as hallways and dining areas, being connected by open passages or staircases rather than actual doors. Each building comprises between 1 and 5 levels and contains between 4 and 40 rooms. The annotation time varies depending on the building’s size and the number of rooms, typically ranging from 4 to 10 hours per building.

All annotations undergo visual verification by separate expert annotators. Tab. 1 compares and summarizes properties across different datasets.

spaces, such as hallways and dining areas, being connected by open passages or staircases rather than actual doors. Each building comprises between 1 and 5 levels and contains between 4 and 40 rooms. The annotation time varies depending on the building’s size and the number of rooms, typically ranging from 4 to 10 hours per building.

All annotations undergo visual verification by separate expert annotators. Tab. 1 compares and summarizes properties across different datasets.

spaces, such as hallways and dining areas, being connected by open passages or staircases rather than actual doors. Each building comprises between 1 and 5 levels and contains between 4 and 40 rooms. The annotation time varies depending on the building’s size and the number of rooms, typically ranging from 4 to 10 hours per building.

All annotations undergo visual verification by separate expert annotators. Tab. 1 compares and summarizes properties across different datasets.

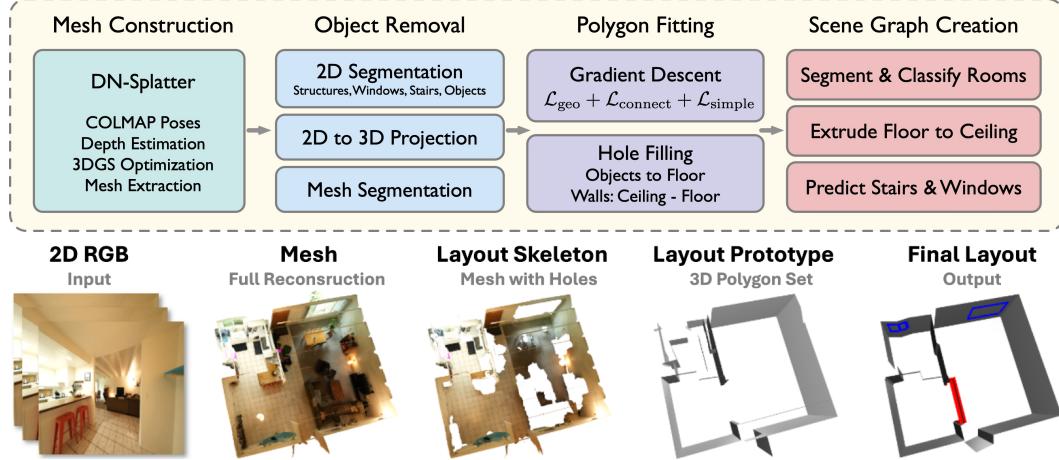


Figure 3: Illustration of the MULTIFLOOR3D model for 3D layout estimation.

4.1 Mesh Reconstruction from RGB Images

Given a set of unposed 2D images, we follow DN-Splatter [30] to obtain a triangle mesh and depth maps for each frame. DN-Splatter uses COLMAP [31] poses together with a 2D depth model to train a 3D Gaussian Splatting [32] reconstruction, and then produces a Poisson surface [33] by sampling from the depths rendered by 3DGs. In our implementation, we use the Metric3D [34] depth model.

4.2 Layout Skeleton Extraction from Mesh

Once the mesh is obtained, we extract a minimal and reliable geometry that serves as the layout *skeleton* using a pre-trained 2D segmentation model. The skeleton should contain only the geometry intended for the final layout. To separate such geometry, we distinguish four semantic categories:

- **Structural Components** (walls, ceilings, floors, and large furniture such as closets): these form the core of the layout skeleton and provide accurate geometry that should be preserved.
- **Geometrically Inaccurate Surfaces** (windows, mirrors): these often suffer from poor depth estimates and are therefore excluded from the skeleton.
- **Objects** (small furniture and household items): these are removed from the skeleton but later help infer missing layout regions.
- **Stairs**: due to their complexity and diversity, they are detected and processed separately.

To construct the skeleton, we first segment the 3D mesh into the four semantic categories. We run the OneFormer model [35] on the input images and map its output classes [36] to our categories. To transfer these labels to the mesh, we back-project $M = 5000$ randomly sampled pixels per image along with their predicted class into 3D, assigning each back-projected point to the nearest mesh vertex and accumulating class votes. We then refine the segmentation by clustering the mesh into *superpoints* following the preprocessing of [37] and assigning each vertex the majority label within its cluster. This yields a mesh labeled with our semantic classes. From this mesh, we extract the *layout skeleton* by selecting the structural components, and isolate the *object* and *stair* subsets for later processing. See Fig. 3 (bottom) for an illustration of a layout skeleton.

4.3 Fitting a Layout Prototype to the Skeleton

We observe significant artifacts in the layout skeletons, including holes and unobserved regions. For example, areas hidden behind furniture, or areas corresponding to windows are missing. In this stage, we use geometric and semantic information to correct the artifacts and infer a more complete *layout prototype*. To this end, we run an optimization that aims to improve the completeness of the obtained skeleton: We first initialize a collection of planar 3D polygons \mathcal{P} from the layout skeleton. In particular, each segmented superpoint (see Sec. 4.2) of the skeleton is fitted to one or more planes. We then optimize the *vertex positions* and *plane equations* of the polygons using three main objectives:

- \mathcal{L}_{geo} reconstructs an accurate scene geometry
- $\mathcal{L}_{\text{connect}}$ produces a continuous and connected geometry
- $\mathcal{L}_{\text{simple}}$ produces a mesh with low vertex count

During the optimization, we constrain the vertices of each polygon to be coplanar. We also allow and encourage polygons to share vertices. The initialization and the implementation of vertex constraints and shared vertices is detailed in the supplementary material.

Definitions. Given a polygon P consisting of edges E and a point $p \in \mathbb{R}^3$ we define the point-to-polygon distance $D_{pp}(P, p)$ as the minimal distance between p and any point on the surface of P . For $e \in E$ we define the point-to-edge distance $D_{pe}(p, e)$ as the minimal distance between p and any point on the line segment representing e .

Losses. We fit the polygon set \mathcal{P} using gradient descent and three losses. The first loss \mathcal{L}_{geo} encourages the polygons to reconstruct the original geometry and respect the *observed empty space*:

$$\mathcal{L}_{\text{geom}} = \mathcal{L}_{\text{prox}} + \mathcal{L}_{\text{empty}} \quad (1)$$

$\mathcal{L}_{\text{prox}}$ penalizes the distance of each vertex $v \in V_{\text{skeleton}}$ of the Layout Skeleton to the closest polygon surface:

$$\mathcal{L}_{\text{prox}} = \sum_{v \in V_{\text{skeleton}}} \min_{P \in \mathcal{P}} D_{pp}(v, P) \quad (2)$$

To prevent occluding the observed empty space (*i.e.*, the space we believe to be empty based on the depth maps), we sample a set L of line segments using the input camera poses and computed depth maps. Each line segment extends from the camera pose to the back-projected depth. We then penalize line segment-polygon intersections as follows: If a line segment l intersects a polygon, the nearest polygon edge e^* should be moved closer to the intersection point p_{inter} .

$$\mathcal{L}_{\text{empty}} = \sum_{l \in L} \sum_{\substack{P \in \mathcal{P} \\ l \cap P \neq \emptyset \\ D_{pe}(p_{\text{inter}}, e^*) \leq \tau_{\text{inter}}}} D_{pe}(p_{\text{inter}}, e^*) \quad (3)$$

$$\text{where } p_{\text{inter}} = l \cap P \quad \text{and} \quad e^* = \operatorname{argmin}_{e' \in \text{edges}(P)} D_{pe}(v, e').$$

Note that we ignore intersections with $D_{pe}(p_{\text{inter}}, e^*)$ greater than the threshold τ_{inter} to avoid noise from intersections far from the polygon boundary.

The second loss $\mathcal{L}_{\text{connect}}$ prevents small gaps and encourages shared boundaries by making polygons attract vertices. Concretely, $\mathcal{L}_{\text{connect}}$ penalizes the distance from each polygon vertex to the closest surface of another polygon:

$$\mathcal{L}_{\text{connect}} = \sum_{P \in \mathcal{P}} \sum_{v \in \text{vertices}(P)} \min_{P' \in \mathcal{P}, P' \neq P} D_{pp}(v, P') \quad (4)$$

As for $\mathcal{L}_{\text{empty}}$, we ignore points with $D_{pp}(v, P')$ greater than a threshold.

The third loss encourages simplicity and smooth polygon boundaries. $\mathcal{L}_{\text{simple}}$ penalizes the length of all edges that are not shared by at least two polygons. (*i.e.*, not all edge vertices are shared). Intuitively, $\mathcal{L}_{\text{simple}}$ promotes shared edges (for instance, an edge between two walls) to represent the scene while edges that are not shared are shrunk until they are eliminated.

$$\mathcal{L}_{\text{simple}} = \sum_{P \in \mathcal{P}} \sum_{e \in \text{edges}(P)} \mathbf{1}_{[\# P' \in \mathcal{P} \setminus \{P\}: e \subset P']} \|e\|_2 \quad (5)$$

Our final loss is given by $\mathcal{L} = \mathcal{L}_{\text{geom}} + \mathcal{L}_{\text{connect}} + \mathcal{L}_{\text{simple}}$.

Vertex Merging $\mathcal{L}_{\text{simple}}$ itself does not reduce the number of vertices or polygons in the polygon set. Instead, we periodically manually simplify P by (1) merging vertex pairs with distance below τ_{merge} , (2) applying the RDP [38] algorithm with tolerance τ_{merge} to the polygons individually, and (3) merging close polygons with similar normal. (Close in terms of minimal D_{pp} distance among the vertices.) For (3) we additionally verify that the merged polygon does not increase $\mathcal{L}_{\text{prox}}$ too strongly. Note that step (1) is the source of shared vertices between polygons.

Closing Holes in the Floor. We observe that there is a floor under most objects in a room. We exploit this information by projecting objects to the floor, *i.e.*, we project each triangle of the *object* mesh extracted in Sec. 4.2 to the plane equation of the nearest floor-classified polygon whose centroid lies below the triangle. We recompute the floor polygon from the union of the original floor polygon and the projected triangle surfaces.

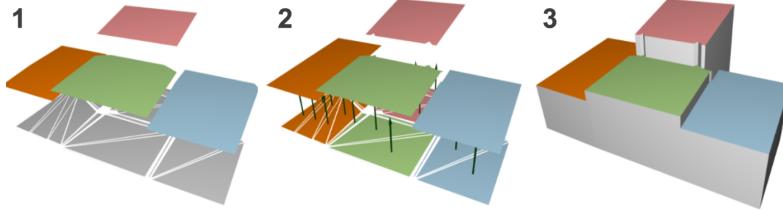


Figure 4: **Our proposed floor extrusion algorithm.** 1) Floor triangulation. 2) Triangles assigned to ceilings using midpoints. 3) Triangles extruded to ceiling planes.

Closing Wall Holes. We extend walls to span from ceiling to floor. Specifically, we identify polygon edges of wall-classified polygons whose normals face downward, and count how many line segments in L (representing observed empty space) intersect the region between each edge and the floor. If the number of intersections per cm^2 falls below τ_{extend} , we extend that edge to the floor. We apply the same procedure to ceilings and to wall edges whose normals face upward. The output of this stage is the *layout prototype*.

4.4 Scene Graphs from a Layout Prototype

Next, we convert the prototype (a set of semantically labeled polygons) into the final layout. The layout is represented as a scene graph whose nodes correspond to rooms and whose edges correspond to doors and stairs. Each room node contains one floor along with its associated wall, ceiling, and window polygons. To obtain this structure, we first generate 2D floorplans and then extrude them into 3D. The indirection through 2D is motivated by the fact that the 3D layout prototype does not provide a clear indoor–outdoor separation and does not guarantee that the layout is closed or connected.

Creation of a Scene Graph of 2D Floorplans In this step we use the layout prototype and its semantics to (1) identify the different levels (floors) of the building, (2) create a 2D layout (floorplan) of each level, and (3) segment each level into rooms, extracting a per-level 2D scene graph from each floor and (4) detect stairs to connect the individual levels. In the following, we provide an outline of the applied algorithms, which are detailed in the appendix.

- To identify building floors, we use the floor-classified polygons of the layout prototype, merging close levels with similar heights.
- To create a 2D floorplan of each level, we merge each level’s floor polygon(s) with suitable ceiling polygons — since ceilings are rarely occluded by objects and thus are more robustly represented in the layout prototype.
- To segment each level into rooms, we apply Hov-SG [22]’s room segmentation algorithm on each 2D floorplan (and the walls of the layout prototype). The segmentation outputs a scene graph with rooms as nodes, and *openings* as edges. We consider an opening edge a *door* if its width is below 1.5 m. Otherwise, we retain its edge but label it as *opening*. Furthermore, each room is associated with a room type (kitchen, office, …).
- To identify stairs, we cluster connected components of the stair mesh extracted in Sec. 4.2. For each component, we add an edge to the scene graph between the rooms and floors it connects.

Back to 3D: Room Extrusion Sec. 4.4 describes how we use the layout prototype to generate a scene graph of rooms. In this section, we propose a simple algorithm inspired by layout annotation tools [29], that extrudes each node’s 2D floorplan to the ceiling. For a single room, the extrusion algorithm creates a closed room shell using a 2D floorplan and a set of potential 3D ceiling polygons that at least partially cover the floorplan. Fig. 4 visualizes the extrusion process. Its core idea is to (1) triangulate the 2D floorplan, (2) assign each triangle to a ceiling polygon and (3) extrude each triangle to its ceiling. Specifically, we triangulate the room’s 2D floorplan using a 2D Constrained Delaunay Triangulation [39] built from the boundary of the floorplan, the ceiling candidates’ edges, and the projections of the pairwise intersection lines of the ceiling candidates’ planes. For each triangle center, we cast a ray upward. If the ray hits a ceiling candidate, we assign the triangle to that ceiling’s plane. Intuitively, this assignment partitions the floorplan by ‘rendering’ ceiling polygons on the floor. Triangles that do not hit a ceiling are assigned to the lowest ceiling plane reachable in the graph of unassigned triangles. (Lowest in terms of the triangle midpoint’s projected z-coordinate on the target plane.) Lastly, we extrude each floor triangle to its assigned ceiling plane. That is, we produce ceiling and floor triangles on the ceiling and floor planes respectively, and add axis-aligned wall rectangles for triangle edges coinciding with a wall in the 2D floorplan. To ensure a closed room

Method	Structures		Doors		Windows		Stairs		Depth		
	F1@0.5	Avg F1	Δ_5	Δ_{10}	#Vertices						
RoomFormer [1] (per floor)	0.24 \pm 0.06	0.22 \pm 0.06	0.23 \pm 0.10	0.20 \pm 0.09	0.07 \pm 0.06	0.07 \pm 0.04	—	—	24.9 \pm 11.5	32.9 \pm 14.9	764.9
RoomFormer [1] (per room)	0.18 \pm 0.14	0.16 \pm 0.12	0.18 \pm 0.14	0.16 \pm 0.12	0.08 \pm 0.08	0.09 \pm 0.07	—	—	37.3 \pm 10.4	44.8 \pm 10.7	1134.5
SceneScript [2] (per floor)	0.28 \pm 0.11	0.26 \pm 0.08	0.23 \pm 0.26	0.20 \pm 0.23	0.16 \pm 0.18	0.15 \pm 0.17	—	—	22.5 \pm 8.6	33.8 \pm 11.7	677.1
SceneScript [2] (per room)	0.23 \pm 0.12	0.21 \pm 0.11	0.31 \pm 0.26	0.28 \pm 0.23	0.11 \pm 0.11	0.10 \pm 0.09	—	—	23.5 \pm 7.2	32.9 \pm 6.7	1333.6
MULTIFLOOR3D (Ours)	0.40 \pm 0.10	0.38 \pm 0.10	0.55 \pm 0.16	0.44 \pm 0.15	0.43 \pm 0.29	0.38 \pm 0.22	0.42 \pm 0.48	0.41 \pm 0.44	61.1 \pm 9.2	76.3 \pm 7.9	1957.0

Table 2: **Scores on HOUSELAYOUT3D.** Performance comparison with state-of-the-art layout estimation methods in terms of average and standard deviation across scenes. Structures include wall, floor and ceilings. MULTIFLOOR3D is the only method predicting stairs.

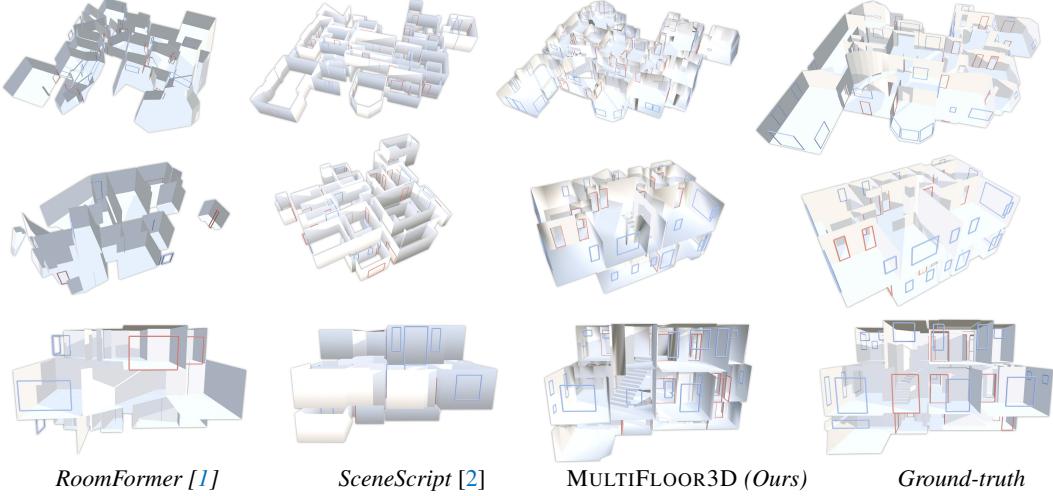


Figure 5: **Qualitative Results on HOUSELAYOUT3D.** We present layout estimation samples from our model alongside state-of-the-art methods. To enhance visualization, we apply back-face culling to the layout meshes, allowing a clear view inside the buildings. Since SceneScript represents walls as boxes, back-face culling is ineffective; instead, we remove the added floors and ceilings for better visibility.

shell, we further add vertical rectangles along potential discontinuous edges in the extruded ceiling surface. To limit complexity, we only consider the 30 largest ceilings per room. Details on how we add doors and stairs after extruding are provided in the appendix.

Window Detection To detect windows, we back-project the 2D window segmentation of the input images obtained in Sec. 4.2 onto layout walls and cluster the result. Concretely, we create rays for window-classified pixels and intersect them with the walls of our 3D layout. We then filter outliers [40], split the points by wall instance, and run DBSCAN [41] for each wall to identify window clusters. To every cluster with at least $k = 10$ vertices, we fit an axis-aligned bounding rectangle. Finally, we predict a window for every rectangle with height and width greater than 30 cm.

5 Experiments

In this section, we introduce the metrics used to evaluate 3D room layout estimation and compare our approach to recent state of the art methods on HOUSELAYOUT3D (Sec.5) and on ScanNet++[42] (Sec.5). We then analyze the contribution of individual pipeline components (Sec.5), and conclude with qualitative results and potential applications (Sec. 5).

Methods in Comparison. We compare our approach to two recent scene layout estimation methods: RoomFormer [1] and SceneScript [2]. Training these baselines on our multi floor dataset is non trivial: RoomFormer targets 2D floorplan prediction, and SceneScript is limited to four corner primitives, so we evaluate them using their publicly available weights on the full HOUSELAYOUT3D dataset. Both baselines are trained on large synthetic datasets (about 100k samples), whereas our method is training free. Following [2], we extrude RoomFormer’s 2D layouts into 3D. As neither baseline predicts floors or ceilings, we append floor and ceiling polygons to each predicted room to enable a fair depth evaluation.

Layout Metrics. To assess the accuracy of the estimated layouts, we adopt the F1 score based on the *entity distance* d_E , following SceneScript [2]. This metric measures the alignment between

Method	#Vertices	Δ_5	Δ_{10}
DN-Splatter Mesh [30]	354k	84.1	92.6
RoomFormer [1]	32.5	36.8	48.9
SceneScript [2]	41.2	55.1	68.5
MULTIFLOOR3D (Ours)	83.1	67.8	84.7

Table 3: **Scores on ScanNet++ [42]**. Metrics evaluate depth accuracy as an approximation of layout estimation error. Scores are averaged over validation scenes.

ground truth entities E and predicted entities E' . For rectangular entities (*e.g.*, doors and windows), d_E is computed as the maximum distance between corresponding corners of two rectangles of the same class: $d_E(E, E') = \max\left\{\|c_i - c'_{\pi(i)}\| : i = 1, \dots, 4\right\}$ where $\pi(i)$ denotes the optimal corner permutation obtained via Hungarian matching. The F1 score $@\tau$ is then computed by applying a threshold τ to d_E as in [2]. For non-rectangular entities, we introduce a generalized entity distance d_H which allows comparison between entities with different numbers of corners. We define d_H as the Hausdorff distance between two polygon surfaces (*i.e.* entities) P, P' and their vertices V, V' :

$$d_H(P, P') = \max\left\{\max_{v \in V} D_{pp}(v, P'), \max_{v' \in V'} D_{pp}(v', P)\right\} \quad (6)$$

for the point-to-polygon distance D_{pp} defined in Sec 4.3. We then use d_H analogously to d_E to compute the F1 score for walls, floors, and ceilings.

Depth Metrics. Following [25], we use input camera poses to render depth maps for the ground truth geometry D_{GT} and predict layouts D_{pred} . When explicit layout annotations are unavailable (*e.g.*, ScanNet++ [42]), depth consistency serves as a proxy for evaluating layout accuracy. Specifically, we compute the percentage of predicted pixel depths that fall within a threshold τ cm of the GT depth:

$$\Delta_\tau = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{[|D_{pred}(i) - D_{GT}(i)| \leq \tau]} \quad (7)$$

This Δ_τ metric was introduced [43] and is commonly used in monocular depth estimation [34].

Results on HOUSELAYOUT3D Tab. 2 shows scores for the F1-based metrics across semantic classes, and depth metrics on our HOUSELAYOUT3D dataset. For this experiment, we use the camera poses, RGB-D images and mesh of MP3D [14]. As neither baseline is designed for multi-floor layout prediction, we apply them separately per floor (or per room) and then merge the per-floor (or per-room) predictions. We use the ground-truth MP3D floor and room segmentation and report scores per-floor and per-room. Our MULTIFLOOR3D does not have access to this privileged information.

MULTIFLOOR3D significantly outperforms state-of-the-art layout estimation methods, despite not using ground-truth floor or room segmentation. While both baselines perform better on individual rooms than full floors, this gap is smaller for SceneScript, which favors compactness (*i.e.*, fewer vertices) at the cost of geometric accuracy.

Results on Scannet++ Tab. 3 shows additional results on the Scannet++ [42] *DSLR* validation split, consisting of 50 scenes captured with a monocular hand-held camera and COLMAP-generated image poses. As ScanNet++ does not provide ground truth layout annotation, we only report depth metrics as an approximation of the layout error. Since ScanNet++ scenes are populated with objects, we use ground truth semantic annotations to ignore those points during the evaluation, as well as points on windows which are typically not well reconstructed in the ground truth laser scan. As input for all methods, use the mesh provided by the Gaussian Splatting approach DN-Splatter [30] in the first stage of our method (Sec. 4.1). The results indicate that MULTIFLOOR3D outperforms the baselines at the cost of compactness (larger number of vertices).

Method	Avg F1	#Vertices	Sem.
Input Mesh + QSLIM [44]	0.109	2000.0	✗
Layout Skeleton + QSLIM [44]	0.223	2000.0	✗
Layout Prototype	0.373	2553.0	✗
MULTIFLOOR3D (Ours)	0.381	1957.1	✓
(w/o prototype fitting)	0.214	2269.8	✓
(w/o room segmentation)	0.359	2442.2	(✓)

Table 4: **Ablation Study on HOUSELAYOUT3D**.

Analysis Experiments. Tab. 4 shows the contributions in terms of F1 score of each stage in our approach. Note that the outputs of the first and second stages (*mesh* from Sec. 4.1 and *layout skeleton* from Sec. 4.2) are triangle meshes, which we convert to polygon sets by first applying mesh simplification (QSLIM [44]), and then greedily merging adjacent triangles whose normals differ by less than 20° . Scores drop significantly when either layout fitting or room segmentation is removed.

Qualitative Results. We show qualitative results of our approach in Fig. 5 and compare to RoomFormer [1] and SceneScript [2]. Both baselines methods struggle with large areas consisting of multiple rooms, RoomFormer even more than SceneScript. The baselines are also inherently limited to predicting rectangular primitives and cannot represent more complex shapes such as sloped ceilings (*top example*). In Fig. 6 we visualize qualitative results when removing loss objectives from the mesh fitting stage introduced in Sec. 4.3.

Applications. Next, we demonstrate a potential application of full-building 3D layouts. First, we obtain the 3D scene graph where nodes represent rooms, and edges are connections between rooms (doors, stairs, etc.) Then, we feed the scene graph in JSON format to an LLM, together with a user-prompt asking for directions. The LLM responds with turn-by-turn directions on how to reach the desired location. This concept is illustrated in Fig. 7. Beyond navigation and planning, rich structural layouts also form a natural foundation for generative scene reasoning, as demonstrated by video-conditioned synthesis approaches such as VIPScene [45], suggesting a future where large-scale layouts and video-perception models become tightly coupled.

Limitations. MULTIFLOOR3D has a longer runtime than the feed-forward baselines, taking one to two hours per HOUSELAYOUT3D scene on an NVIDIA GeForce RTX 4090, compared to one to two minutes for SceneScript [2] and RoomFormer [1]. Furthermore, MULTIFLOOR3D occasionally struggles to remove outdoor elements perceived through large windows, which can introduce artifacts.

6 Conclusion

We introduced HOUSELAYOUT3D, the first benchmark for evaluating 3D layout estimation in large-scale multi-floor buildings. Existing methods remain restricted to single-floor settings, and our experiments show that they struggle to parse complex buildings with multiple floors and rooms, whereas our learning-free baseline already outperforms these approaches. Looking forward, there is a clear need for learning-based models that reason across entire buildings rather than relying on heuristics that, while effective, are substantially slower than feed-forward networks. Beyond layout estimation, the large-scale structural annotations in HOUSELAYOUT3D also provide a foundation for 3D scene synthesis [3, 45, 46], enabling the generation of coherent multi-floor environments that can in turn supply additional training data for a wide range of 3D perception models [47–49] as in [50, 51]. In summary, we hope this benchmark will drive progress both in robust multi-floor layout estimation and in generative models that support the broader development of 3D perception methods.

Acknowledgments. Francis Engelmann is supported by an SNSF PostDoc Mobility Fellowship.

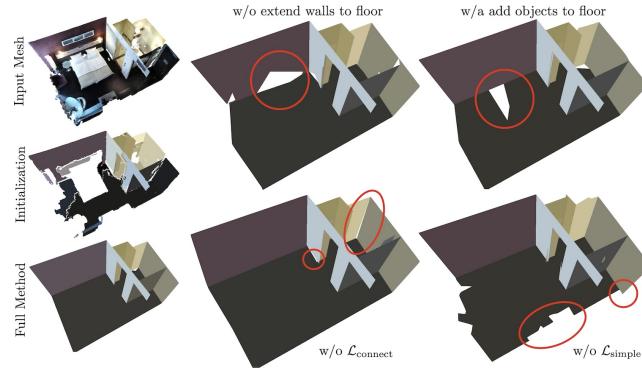


Figure 6: **Effect of Loss Terms.** Left: input and output of our approach. Right: result when ablating losses and components. Omitting object projection (top center) or wall extension (top right) produces holes in the layout. Without $\mathcal{L}_{\text{simple}}$, the polygon boundaries show dents. Without $\mathcal{L}_{\text{connect}}$, we observe gaps between polygons that otherwise share edges.

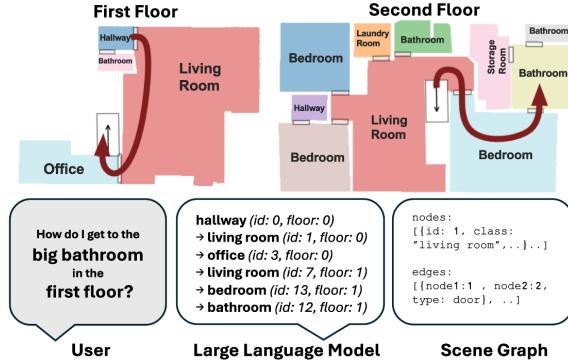


Figure 7: Navigation application based on 3D layouts and LLMs.

References

- [1] Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [2] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. SceneScript: Reconstructing Scenes with an Autoregressive Structured Language Model. *European Conference on Computer Vision (ECCV)*, 2024.
- [3] Ata Çelen, Marc Pollefeys, Daniel Barath, and Iro Armeni. HouseTour: A Virtual Real Estate a (I) Gent. In *International Conference on Computer Vision (ICCV)*, 2025.
- [4] Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni. WildGS-SLAM: Monocular Gaussian Splatting SLAM in Dynamic Environments. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [5] Shengze Jin, Iro Armeni, Marc Pollefeys, and Daniel Barath. Multiway Point Cloud Mosaicking with Diffusion and Global Optimization. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [6] Oliver Lemke, Zuria Bauer, René Zurbrügg, Marc Pollefeys, Francis Engelmann, and Hermann Blum. Spot-Compose: A Framework for Open-Vocabulary Object Retrieval and Drawer Manipulation in Point Clouds. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024.
- [7] René Zurbrügg, Yifan Liu, Francis Engelmann, Suryansh Kumar, Marco Hutter, Vaishakh Patil, and Fisher Yu. ICGNet: A Unified Approach for Instance-Centric Grasping. In *International Conference on Robotics and Automation (ICRA)*, 2024.
- [8] Jiacheng Chen, Ruizhi Deng, and Yasutaka Furukawa. Polydiffuse: Polygonal Shape Reconstruction via Guided Set Diffusion Models. *International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [9] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A Large Photo-Realistic Dataset for Structured 3D Modeling. In *European Conference on Computer Vision (ECCV)*, 2020.
- [10] Armen Avetisyan, Tatiana Khanova, Christopher Choy, Denver Dash, Angela Dai, and Matthias Nießner. SceneCAD: Predicting Object Alignments and Layouts in RGB-D Scans. In *European Conference on Computer Vision (ECCV)*, 2020.
- [11] Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-Modal Coarse Visual Localization on 3d Scene Graphs. In *European Conference on Computer Vision (ECCV)*, 2024.
- [12] Matthias Wüest, Francis Engelmann, Ondrej Mikšík, Marc Pollefeys, and Daniel Barath. UnLoc: Leveraging Depth Uncertainties for Floorplan Localization. *arXiv preprint arXiv:2509.11301*, 2025.
- [13] Chenyangguang Zhang, Alexandros Delitzas, Fangjinhua Wang, Ruida Zhang, Xiangyang Ji, Marc Pollefeys, and Francis Engelmann. Open-Vocabulary Functional 3d Scene Graphs for Real-World Indoor Spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
- [14] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3d Vision (3dV)*, 2017.
- [15] Srivathsan Murali, Pablo Speciale, Martin R. Oswald, and Marc Pollefeys. Indoor Scan2BIM: Building Information Models of House Interiors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [16] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. DuLa-Net: A Dual-Projection Network for Estimating Room Layouts from a Single RGB Panorama. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [17] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

- [18] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. FloorNet: A Unified Framework for Floorplan Reconstruction from 3D Scans. In *European Conference on Computer Vision (ECCV)*, 2018.
- [19] Sebastian Ochmann, Richard Vock, and Reinhard Klein. Automatic Reconstruction of Fully Volumetric 3D Building Models from Oriented Point Clouds. *Journal of Photogrammetry and Remote Sensing (JPRS)*, 2019.
- [20] Ricardo Cabral and Yasutaka Furukawa. Piecewise Planar and Compact Floorplan Reconstruction from Images. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [21] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-SP: Inverse CAD for Floorplans by Sequential Room-Wise Shortest Path. In *International Conference on Computer Vision (ICCV)*, 2019.
- [22] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. *Robotics: Science and Systems (RSS)*, 2024.
- [23] Elisabetta Fedele, Boyang Sun, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. SuperDec: 3D Scene Decomposition with Superquadric Primitives. In *International Conference on Computer Vision (ICCV)*, 2025.
- [24] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow Indoor Dataset: Annotated Floor Plans with 360° Panoramas and 3D Room Layouts. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [25] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan Room Layout Reconstruction from a Single 360 Image: A Comparative Study of State-Of-The-Art Methods, 2020.
- [26] Denys Rozumnyi, Stefan Popov, Kevins-Kokitsi Maninis, Matthias Nießner, and Vittorio Ferrari. Estimating Generic 3D Room Structures from 2D Annotations. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [27] VSIS Lab. Matterport3D-Layout, 2020.
- [28] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-Annotated 3d Reconstructions of Indoor Scenes. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [29] SCASA. PinPoint.
- [30] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.
- [31] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions On Graphics (TOG)*, 2023.
- [33] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson Surface Reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, 2006.
- [34] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards Zero-Shot Metric 3D Prediction from a Single Image. In *International Conference on Computer Vision (ICCV)*, 2023.
- [35] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [36] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014.
- [37] Damien Robert, Hugo Raguet, and Loic Landrieu. Scalable 3D Panoptic Segmentation as Superpoint Graph Clustering. *International Conference on 3d Vision (3dV)*, 2024.

- [38] D. H. Douglas and T. K. Peucker. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature. *The Canadian Cartographer*, 1973.
- [39] CGAL Team. *CGAL::Constrained_Delaunay_triangulation_2*, 2025.
- [40] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000.
- [41] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.
- [42] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *International Conference on Computer Vision (ICCV)*, 2023.
- [43] Lubor Ladický, Jianbo Shi, and Marc Pollefeys. Pulling Things Out of Perspective. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [44] Michael Garland and Paul S. Heckbert. Surface Simplification Using Quadric Error Metrics. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 1997.
- [45] Rui Huang, Guangyao Zhai, Zuria Bauer, Marc Pollefeys, Federico Tombari, Leonidas Guibas, Gao Huang, and Francis Engelmann. Video Perception Models for 3D Scene Synthesis. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- [46] Martin JJ Bucher and Iro Armeni. ReSpace: Text-Driven 3D Scene Synthesis and Editing with Preference Alignment. *arXiv preprint arXiv:2506.02459*, 2025.
- [47] Ayca Takmaz, Alexandros Delitzas, Robert W. Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3D: Hierarchical Open-Vocabulary 3D Segmentation. *IEEE Robotics and Automation Letters (RA-L)*, 2025.
- [48] Mathias Vogel, Keisuke Tateno, Marc Pollefeys, Federico Tombari, Marie-Julie Rakotosaona, and Francis Engelmann. P2p-bridge: Diffusion Bridges for 3d Point Cloud Denoising. In *European Conference on Computer Vision (ECCV)*, 2024.
- [49] Laszlo Szilagyi, Francis Engelmann, and Jeannette Bohg. SLAG: Scalable Language-Augmented Gaussian Splatting. *IEEE Robotics and Automation Letters (RA-L)*, 2025.
- [50] Silvan Weder, Hermann Blum, Francis Engelmann, and Marc Pollefeys. LabelMaker: Automatic Semantic Label Generation from RGB-D Trajectories. In *International Conference on 3d Vision (3dV)*, 2024.
- [51] Guangda Ji, Silvan Weder, Francis Engelmann, Marc Pollefeys, and Hermann Blum. Arkit Labelmaker: A New Scale for Indoor 3d Scene Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.