MVSDet: Multi-View Indoor 3D Object Detection via Efficient Plane Sweeps

Yating Xu[1], Chen Li[2], Gim Hee Lee[1]

National University of Singapore[1], A*STAR[2]

Multi-View Indoor 3D Object Detection

- **Task Definition:** It predicts 3D bounding box of objects in the scene and their corresponding classes from ***N posed images***.
- **Challenge:** How to estimate geometry information from 2D images alone?
- **Existing work:**

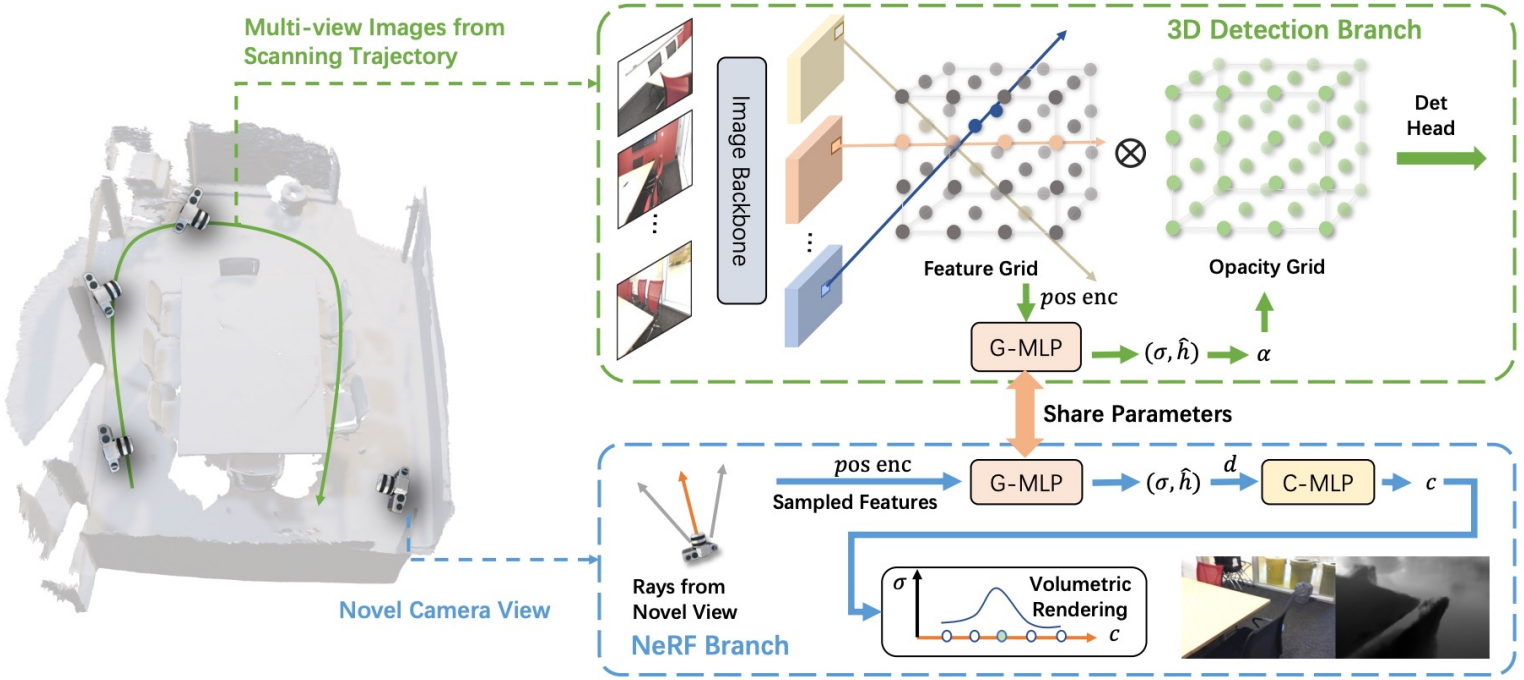
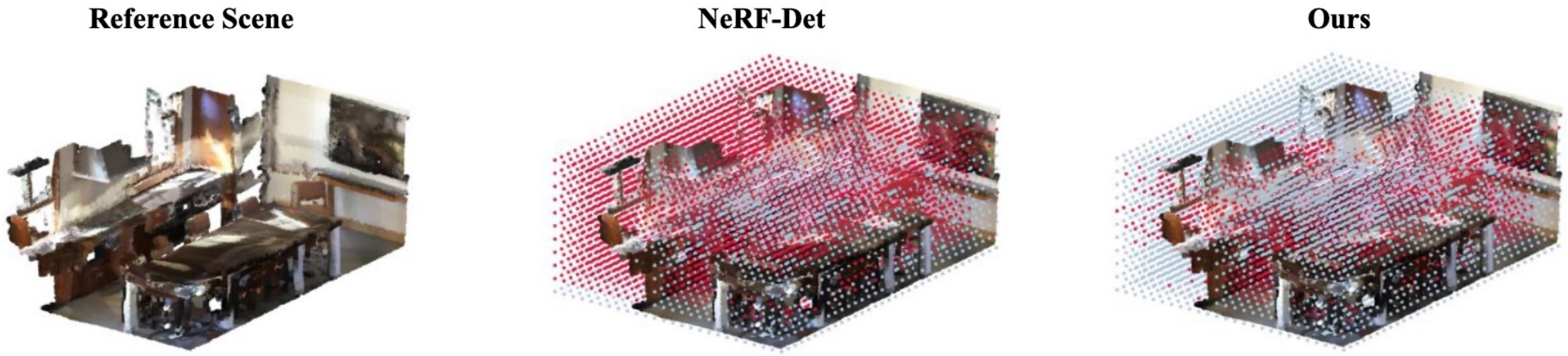


Image source: Xu et al, "NeRF-Det: Learning Geometry-Aware Volumetric Representation for Multi-View 3D Object Detection", ICCV 2023.

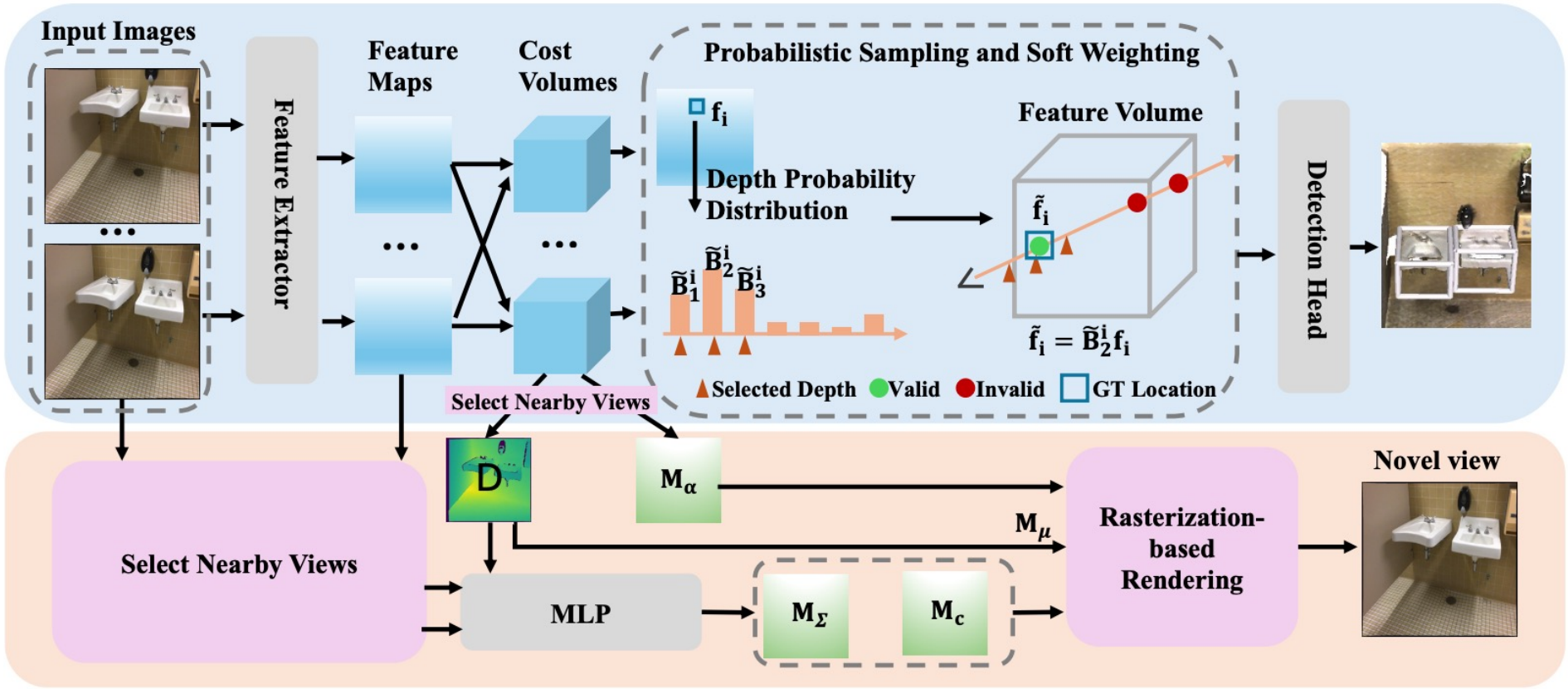
Multi-View Indoor 3D Object Detection

- **Limitation of existing work:** inaccurate geometry estimation



- **Our solution: MVSDet**

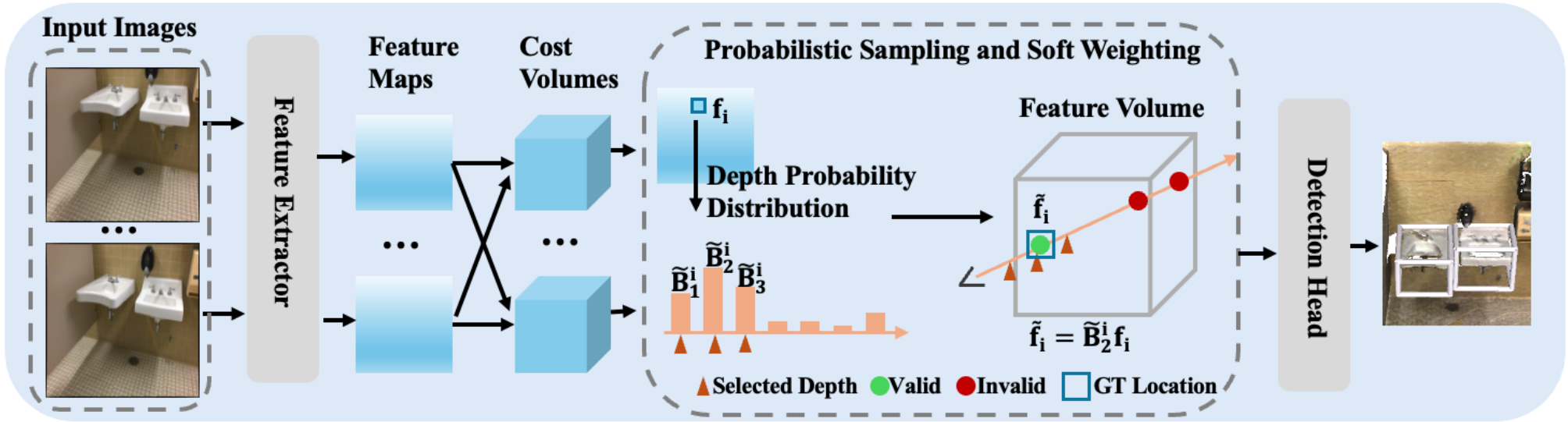
Method Overview



- **Contribution 1:** Probabilistic Sampling and Soft Weighting as *efficient* plane sweep

- **Contribution 2:** pixel-aligned Gaussian Splatting as a *light* depth regularizer

Probabilistic sampling and soft weighting

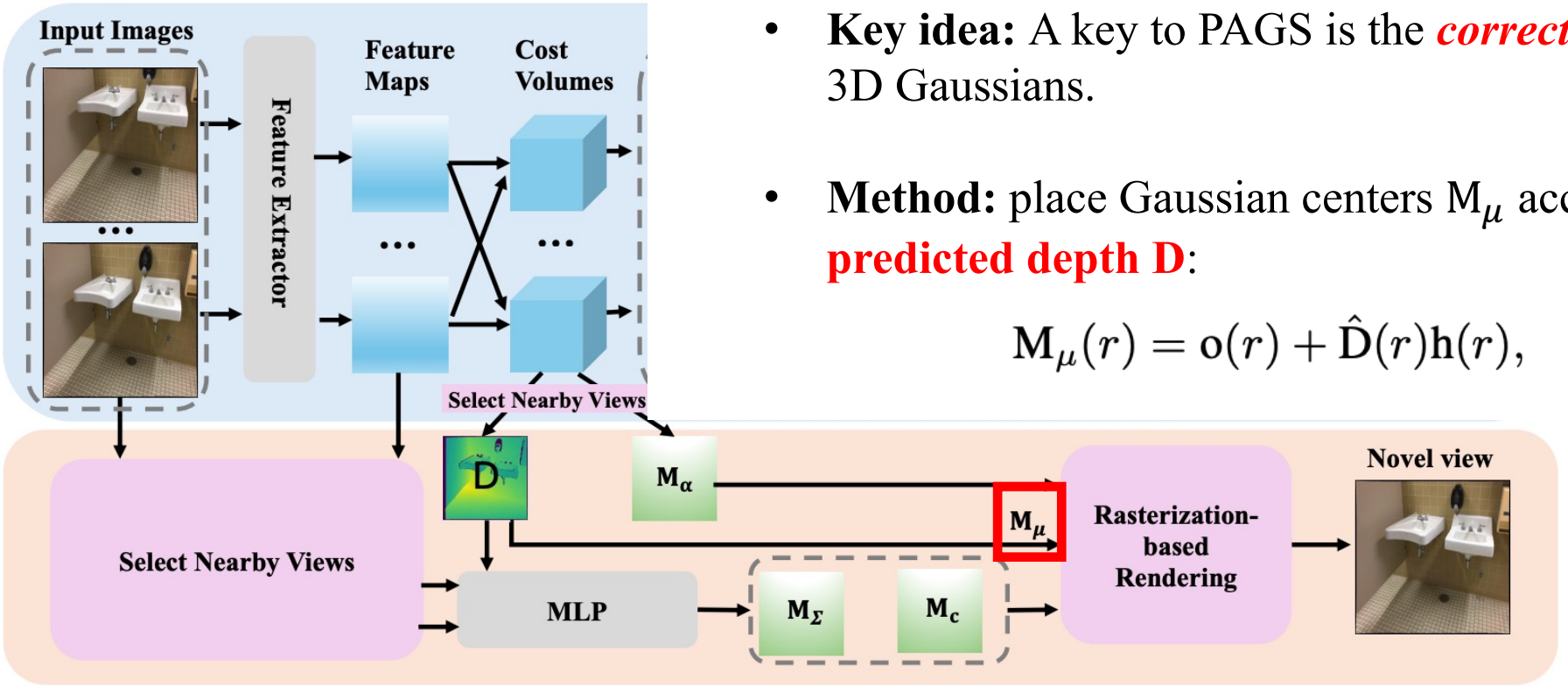


- **Goal:** efficiently learn geometry *without* sampling many depth planes.
- **Method:**
  - Sample **top-k** depth proposals $\{d_{idx_1}, \ldots, d_{idx_k}\}$ with normalized probability score $\{\tilde{B}_{idx_1}, \ldots, \tilde{B}_{idx_k}\}$
  - Feature back-projection to 3D voxel center p from *i*-th image:

$$\tilde{f}_i = \begin{cases} \tilde{B}^i_{\phi(p)} f_i & \text{if } d(p) \subset \{d_{idx_1}, \ldots, d_{idx_k}\} \\ 0 & \text{otherwise} \end{cases}$$

Pixel-aligned Gaussian Splatting (PAGS)

- **Goal:** enhance depth prediction *without* much computation overhead.



- **Key idea:** A key to PAGS is the ***correct positioning*** of the 3D Gaussians.
- **Method:** place Gaussian centers M_μ according to the **predicted depth D**:

$$M_\mu(r) = o(r) + \hat{D}(r)h(r), \quad D = BG$$

- **Optimization**

$$\mathcal{L}_{render} = ||\hat{C}_{color} - C_{color}||^2$$

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{render}$$

Experiments

We use $M=12$ depth planes by default.

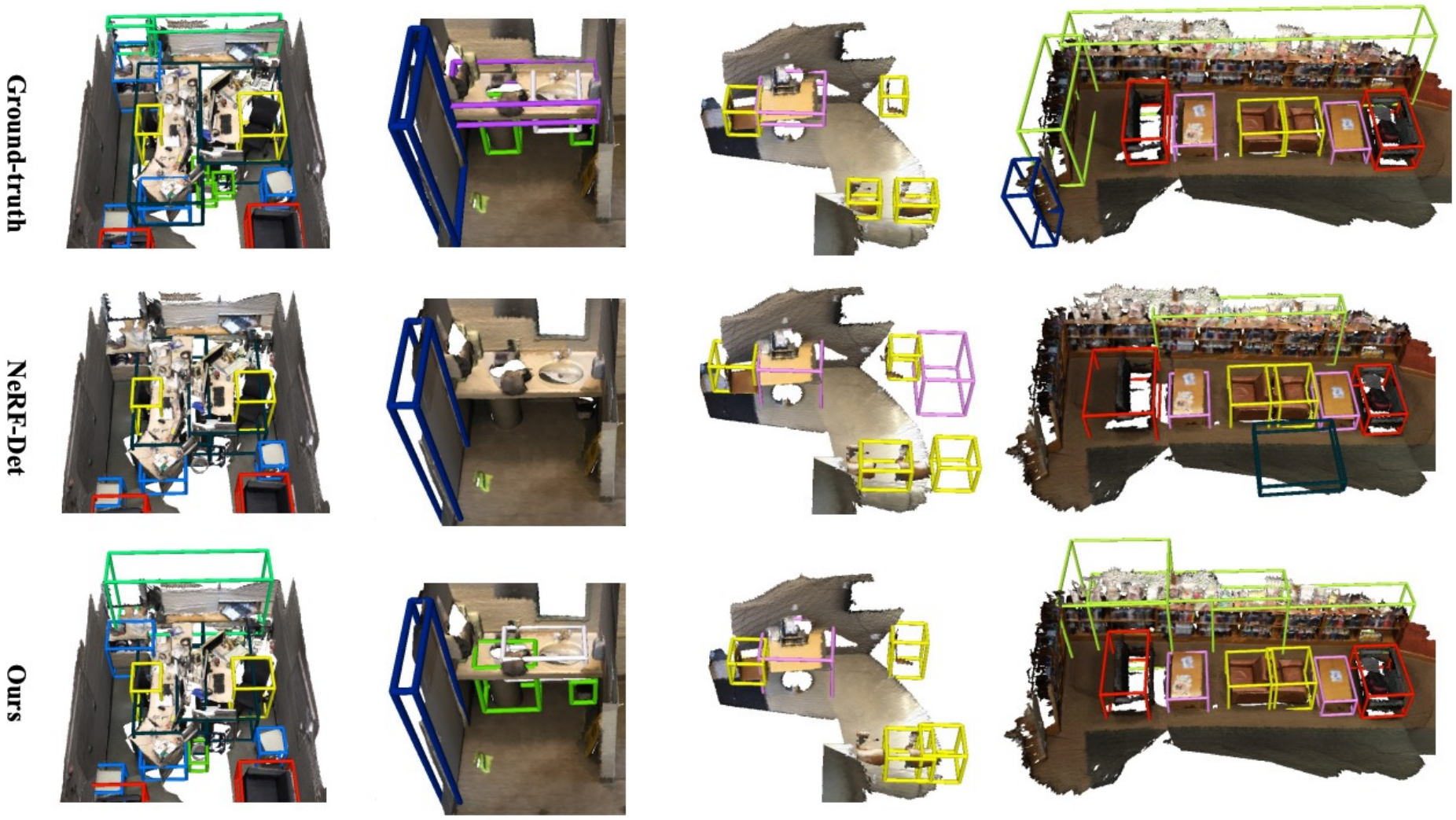
Table 1: Results on ScanNet. "GT Geo" denotes whether ground truth geometry is used as supervision during training.

Method	GT Geo	mAP@.25	mAP@.5
ImGeoNet[18]	✓	54.8	28.4
CN-RMA [16]	✓	58.6	36.8
ImVoxelNet [15]	–	46.7	23.4
NeRF-Det [21]	–	53.5	27.4
Ours	–	**56.2**	**31.3**

Table 2: Results on ARKitScenes. "GT Geo" denotes whether ground truth geometry is used as supervision during training.

Method	GT Geo	mAP@.25	mAP@.5
ImGeoNet[18]	✓	60.2	43.4
CN-RMA [16]	✓	67.6	56.5
ImVoxelNet [15]	–	27.3	4.3
NeRF-Det [21]	–	39.5	21.9
Ours	–	**42.9**	**27.0**

Qualitative Results



Thank You!