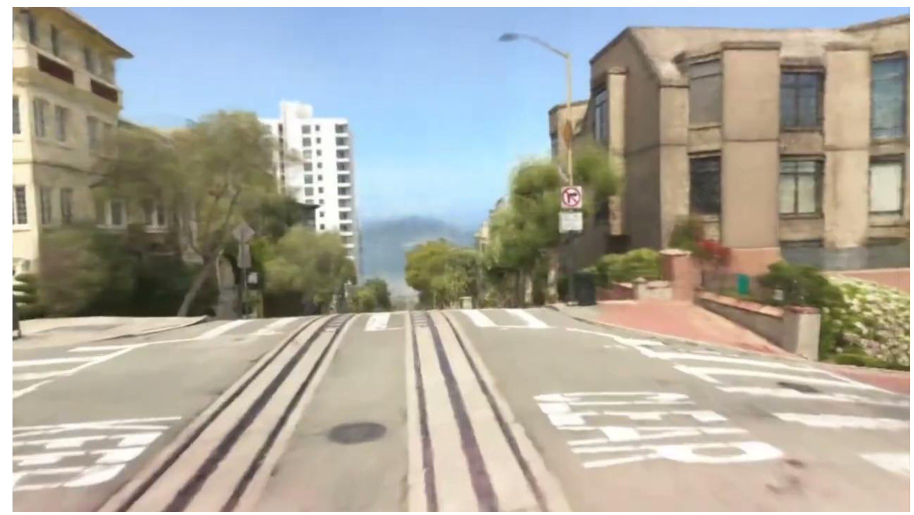
Omni-Recon: Harnessing Image-based Rendering for General-Purpose Neural Radiance Fields

ECCV 2024 Oral

Yonggan Fu, Huaizhi Qu, Zhifan Ye, Chaojian Li, Kevin Zhao,
Yingyan (Celine) Lin

Background: 3D Reconstruction



Block-NeRF, CVPR 2022

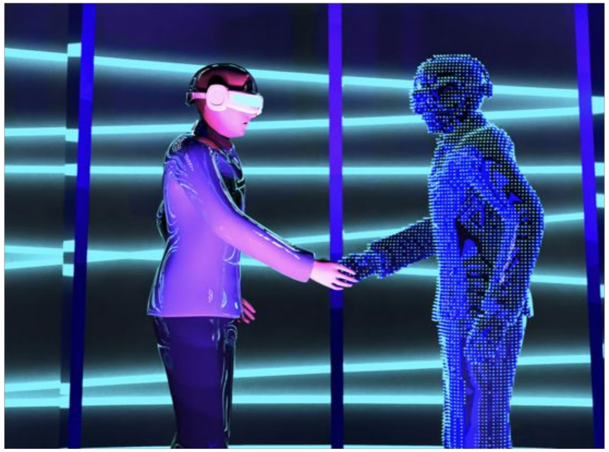
Input: Sparsely captured views

Output: Reconstructed 3D scene

A Demanding Trend: On-device 3D Recon.

Virtual Meetings

Metaverse

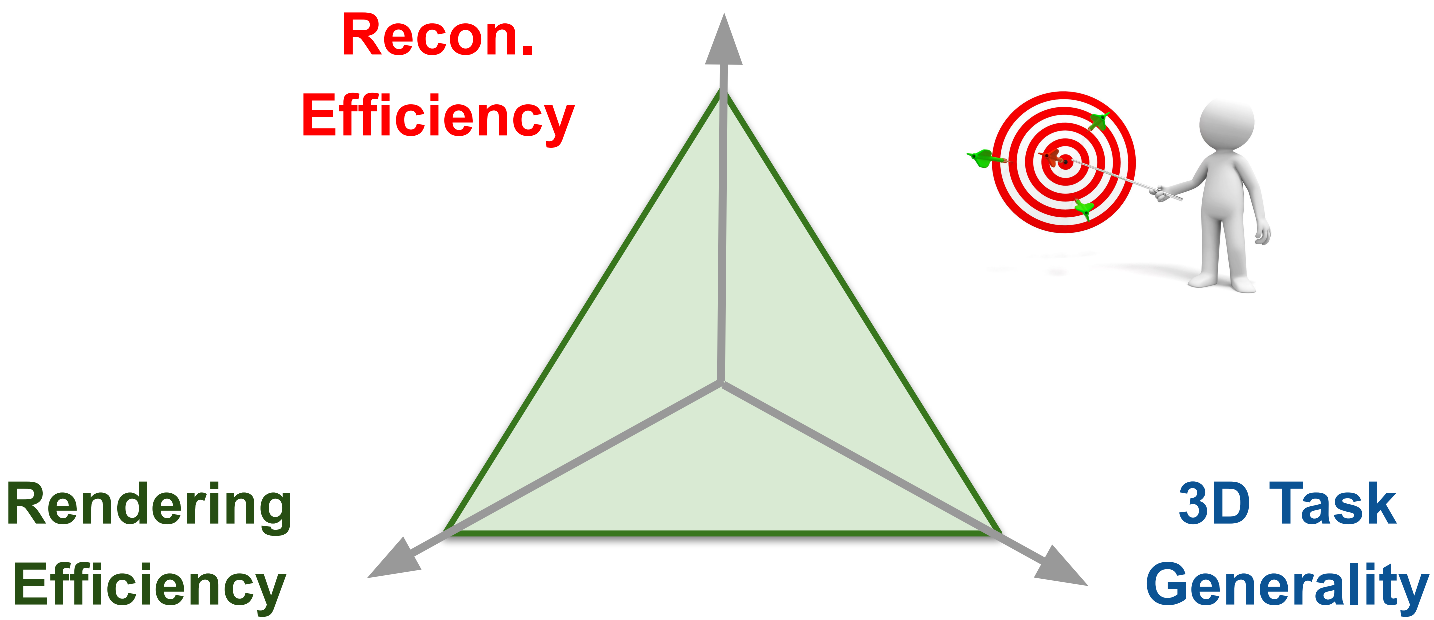


Autonomous Driving

Images from public domains

On-device 3D reconstruction: Highly desirable to enable ubiquitous 3D intelligence

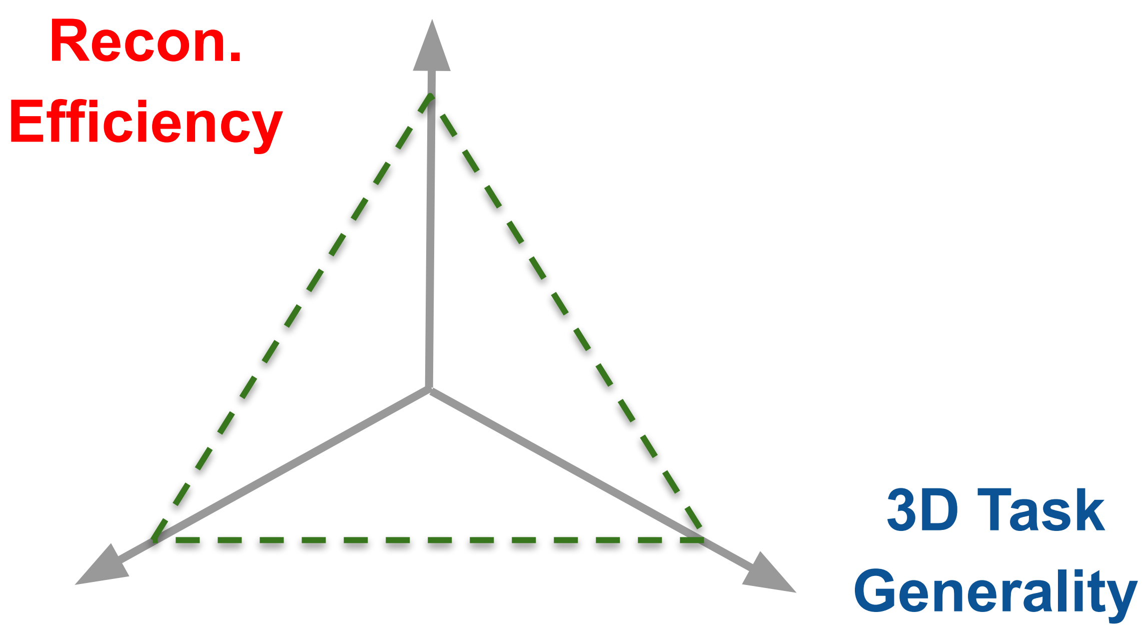
Desired Properties for Real-World 3D Applications



- **Recon. (training) efficiency:** Instantly reconstruct a new scene
- **Rendering efficiency:** Perform on-device real-time rendering
- **3D task generality:** Support general 3D understanding tasks

Limitations of Existing 3D Recon. Solutions

Existing 3D recon. solutions cannot win all the three properties simultaneously



Limitations of Existing 3D Recon. Solutions

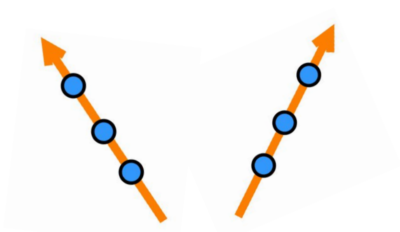
Rendering Efficiency

Recon. Efficiency

Neural Radiance Fields (NeRFs)

Require costly retraining for each new scene & task

Input



New Pos. + View dir.

<u>At Test Time</u>

Per-scene Opt. NeRFs

Output

Novel Views of the Same Scene & Task

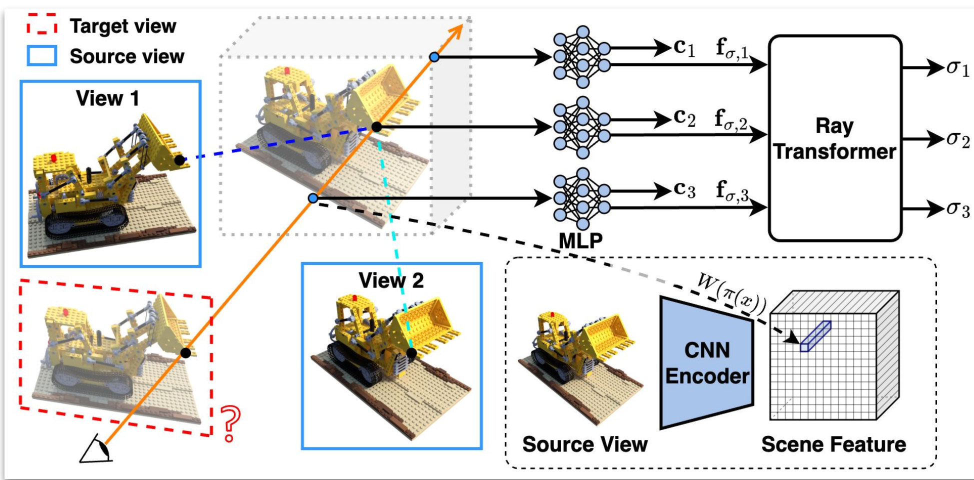
“NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”, B. Mildenhall et al., ECCV’20.

Limitations of Existing 3D Recon. Solutions

Rendering Efficiency | **Recon. Efficiency**

Generalizable NeRFs

Can only achieve < 0.25 FPS on an NVIDIA RTX 2080Ti GPU



Significant complexity of Generalizable NeRF pipelines

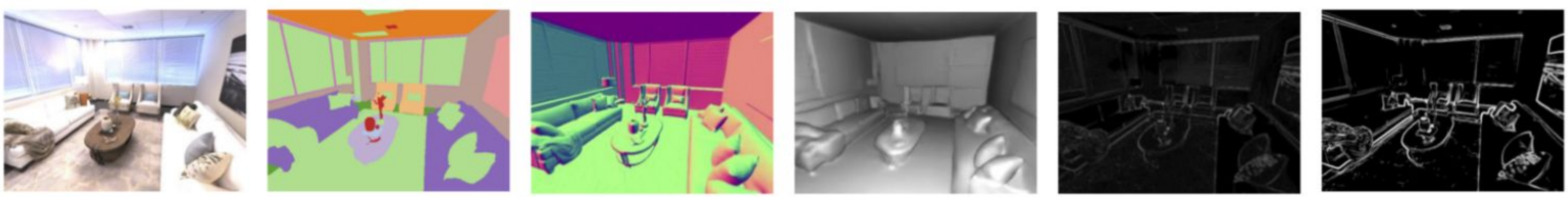
"IBRNet: Learning Multi-View Image-Based Rendering", Q. Wang et al., CVPR'21.

Limitations of Existing 3D Recon. Solutions

	Rendering Efficiency	Recon. Efficiency	3D Task Generality
NeRF-based 3D Und. Models	NO	NO	✓

Need retraining on unseen 3D understanding tasks

Target Scene



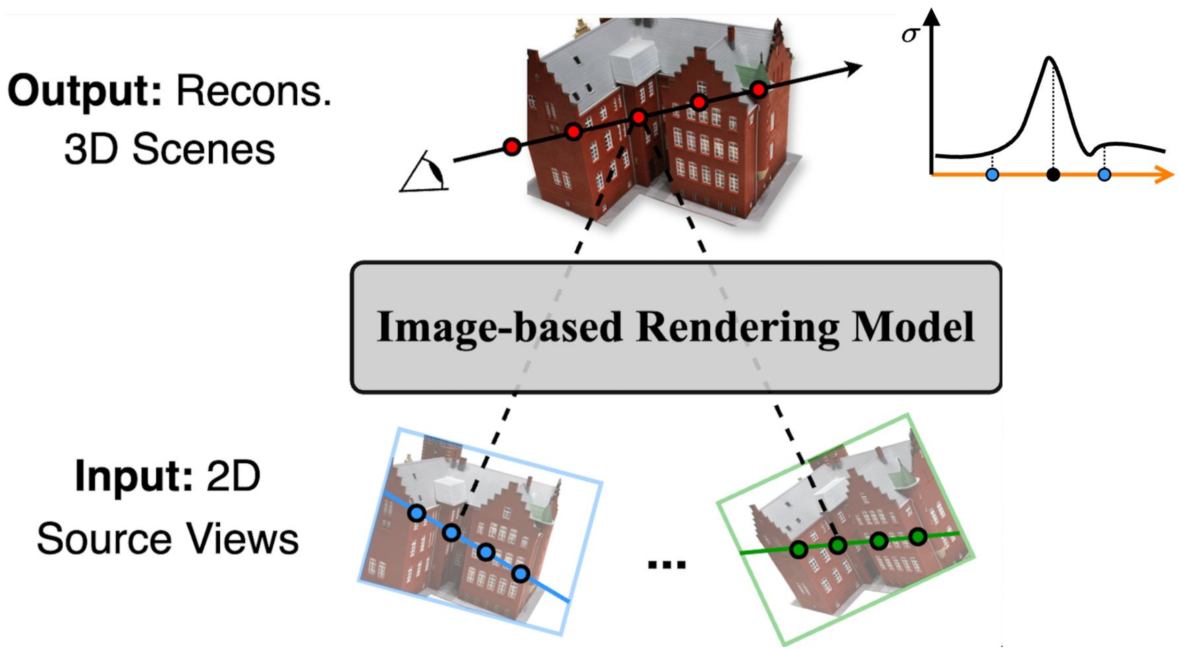
Semantic | Surface Normal | Shading | Keypoint Detection | Edge Detection

"Semantic Ray: Learning a Generalizable Semantic Field with Cross-Reprojection Attention", F. Liu et al., CVPR'23.

Our Proposed Method: Omni-Recon

- **Omni-Recon:** Harnessing image-based rendering for **ubiquitous 3D reconstruction and understanding**
  - Instant scene reconstruction
  - Rapidly enable real-time rendering
  - Zero-shot 3D scene understanding

Omni-Recon: Key Research Question



How to address the broken links in an image-based rendering pipeline?

Rendering Efficiency

Recon. Efficiency

3D Task Generality

Omni-Recon: Key Insights & Enablers

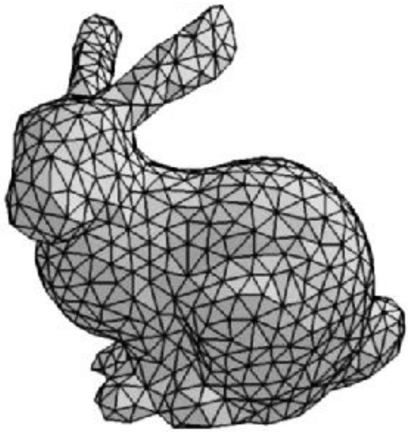
Rendering Efficiency

Recon. Efficiency

3D Task Generality

Insight 1: Pretraining in NeRF and rendering with GPU-friendly representations could win the best of both worlds

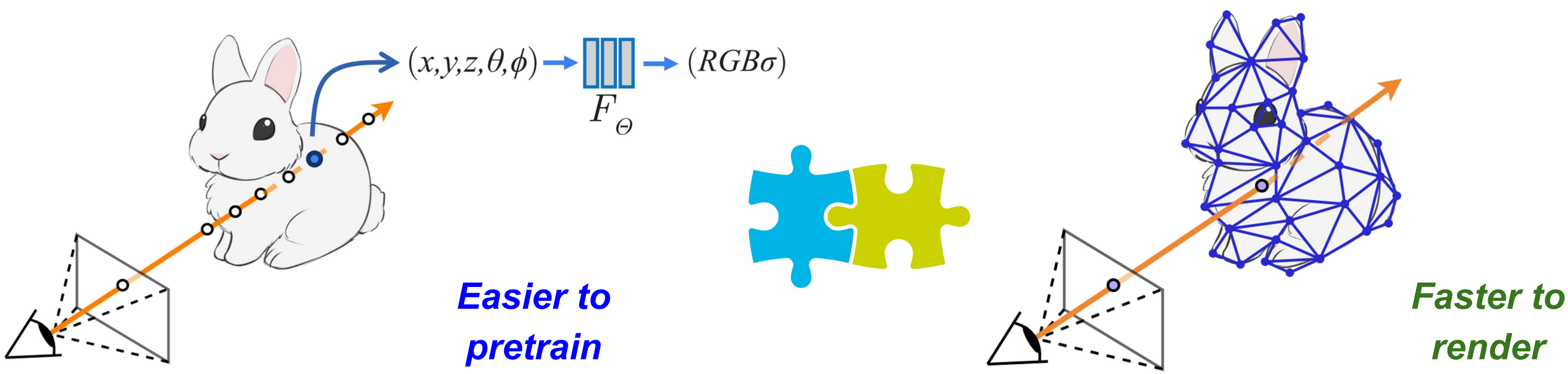
For example: Meshes are GPU-friendly due to rasterization



Omni-Recon: Key Insights & Enablers

Rendering Efficiency **Recon. Efficiency** **3D Task Generality**

Insight 1: Pretraining in NeRF and rendering with GPU-friendly representations could win the best of both worlds



NeRF: MLP inference for **>= 64 points per ray**

Mesh: MLP inference for **1 points per ray** + Well supported by **GPU rasterization**

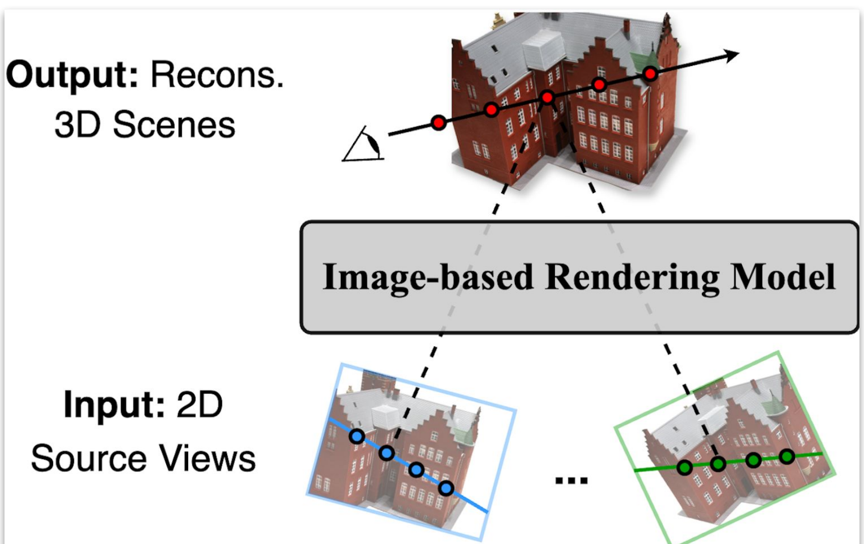
Omni-Recon: Key Insights & Enablers

Rendering Efficiency

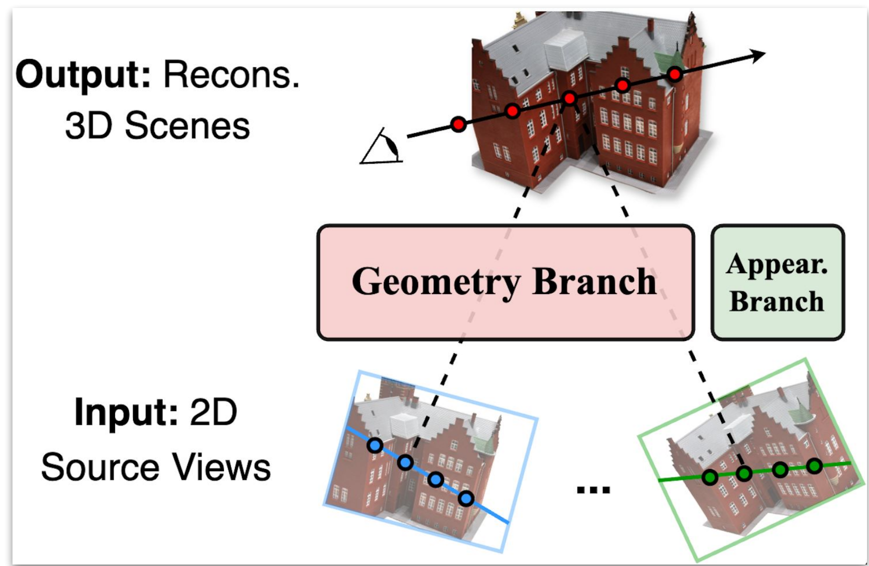
Recon. Efficiency

3D Task Generality

Enabler 1: A NeRF backbone with decoupled geo./appear. branches



Previous models

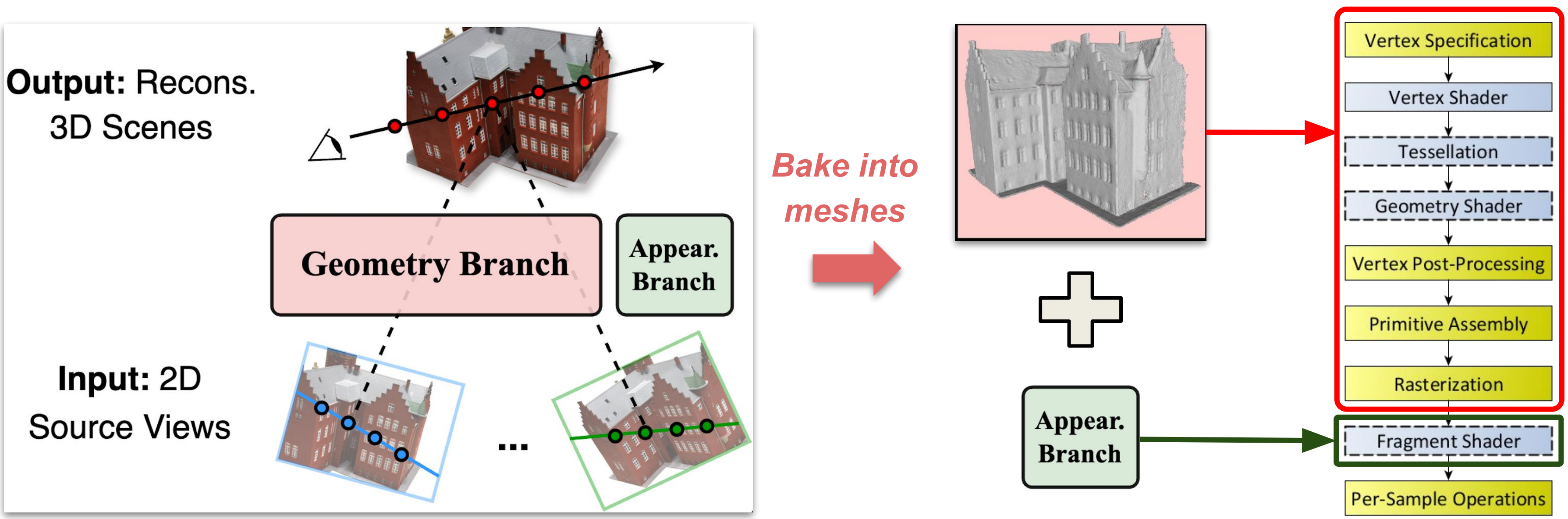


Our Omni-Recon's backbone

Omni-Recon: Key Insights & Enablers

Rendering Efficiency 🤝 **Recon. Efficiency** 🤝 3D Task Generality

Enabler 1: A NeRF backbone with decoupled geo./appear. branches

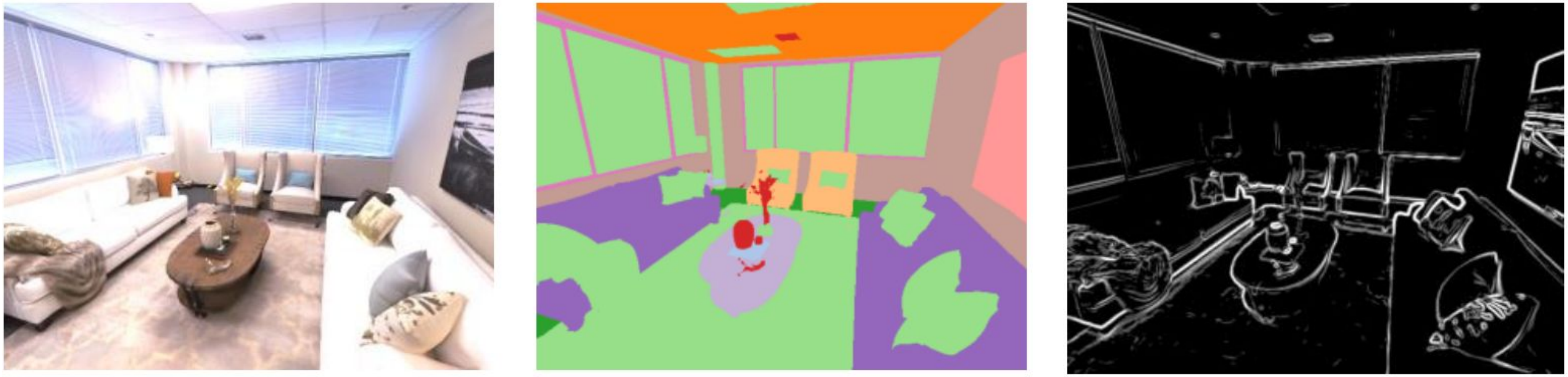


At rendering time: Well-fitted into GPU rasterization pipelines

Omni-Recon: Key Insights & Enablers

Rendering Efficiency | Recon. Efficiency | 3D Task Generality

Insight 2: Regions with similar appearance (RGB) are highly likely to have similar 3D scene properties (e.g., semantics)



Omni-Recon: Key Insights & Enablers

Rendering Efficiency

Recon. Efficiency

3D Task Generality

Enabler 2: Lift 2D task predictions to 3D in a zero-shot manner via reusing the appearance branch predictions

2D predictions using 2D Vision Models

+

Appear. Branch

Lift 2D RGB to 3D

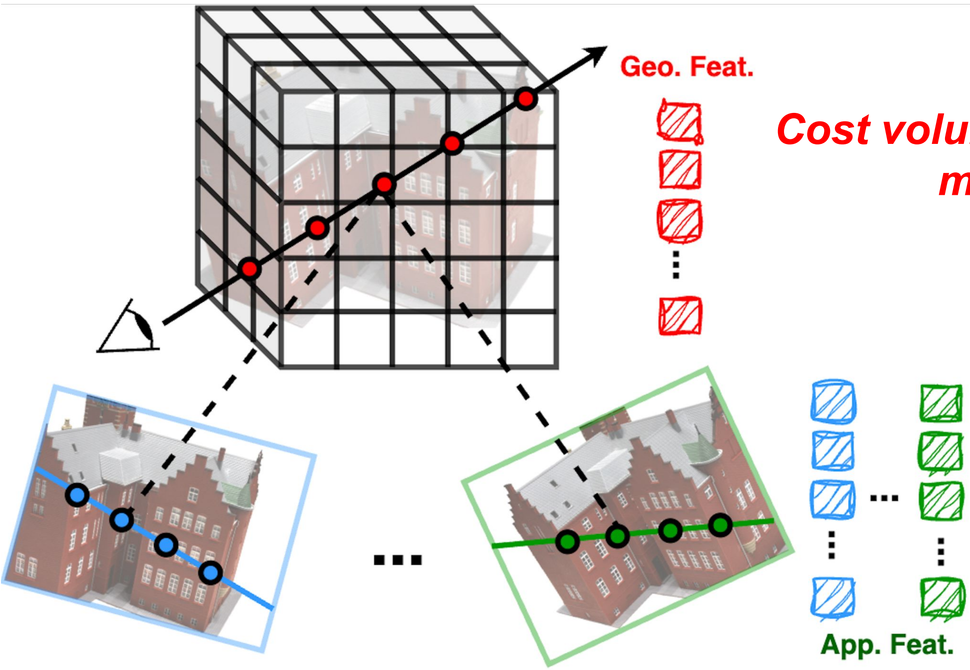
=

Lift 2D task predictions to 3D

Omni-Recon: Detailed Backbone Design

Input: Source views of a new scene

Omni-Recon: Detailed Backbone Design

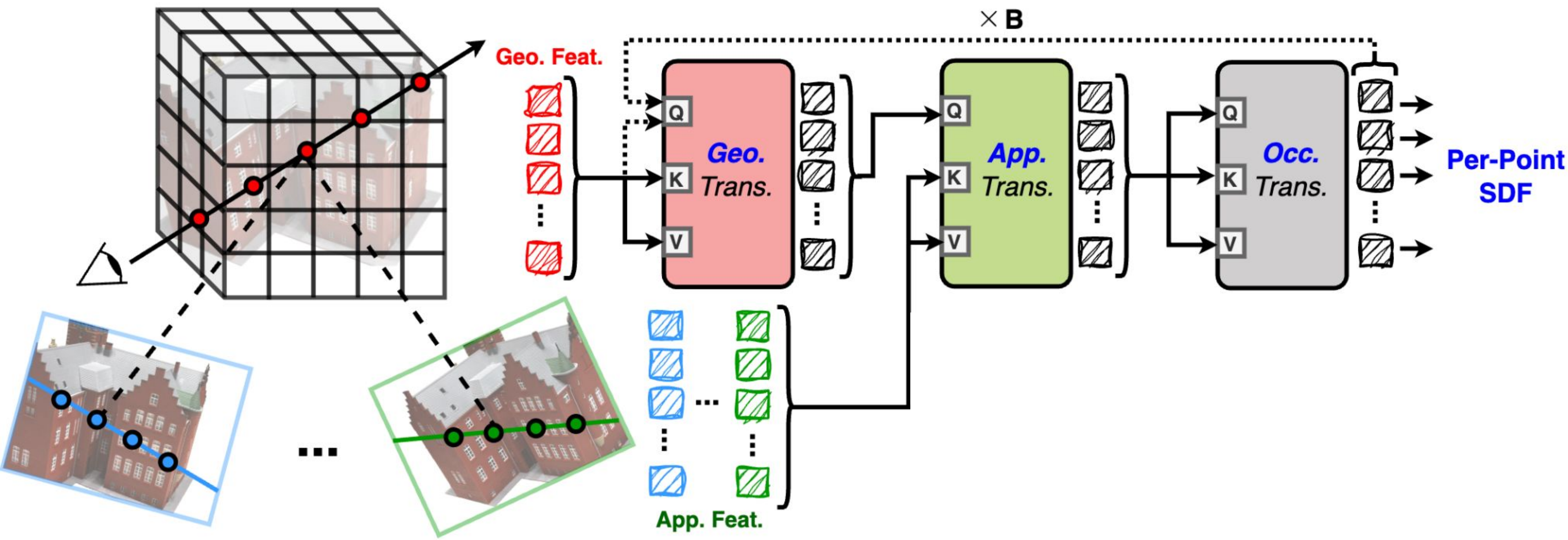


Cost volume constructed following traditional multi-view stereo methods [1]

Extract geometry and appearance features

[1] "MVSNet: Depth Inference for Unstructured Multi-view Stereo", Y. Yao et al., ECCV'18.

Omni-Recon: Detailed Backbone Design



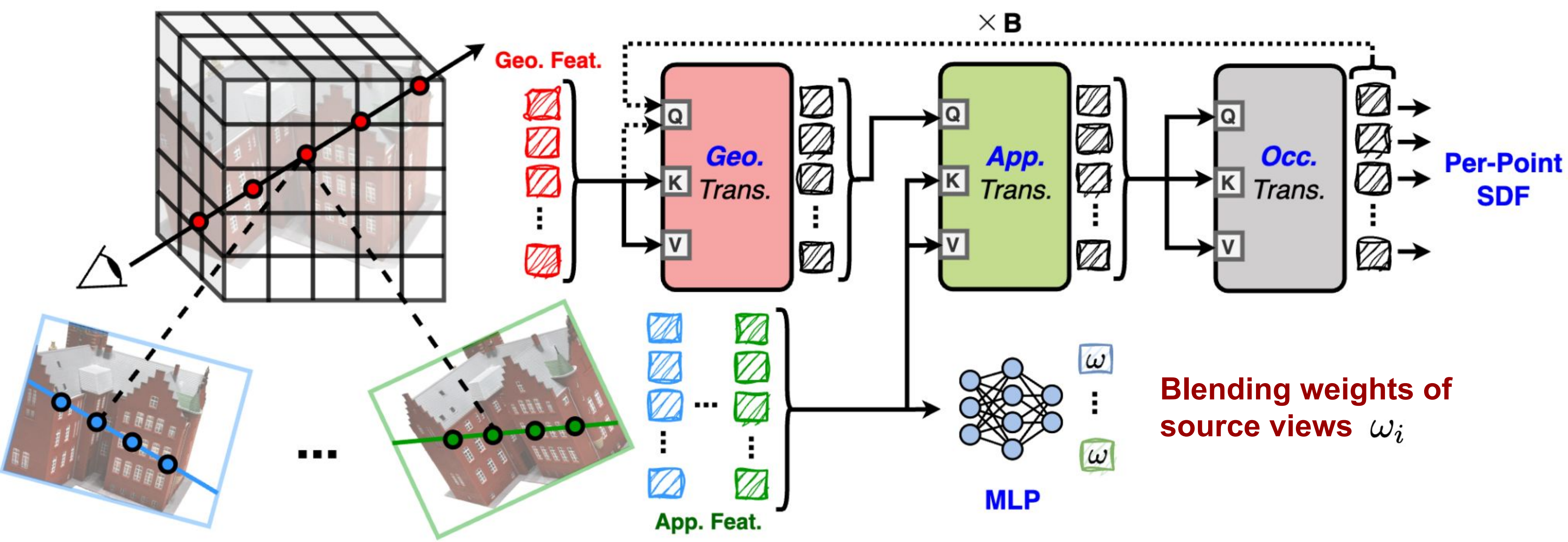
The complex geometry branch: Model the interactions with geometry and appearance features as well as the occlusion effect along the ray

$$\mathbf{M}_{sdf}^{geo}(\mathbf{x}, \{\mathbf{v}_k\}_{k=1}^{K}) = CrossAttention\left(\mathbf{q} = \mathbf{x}, \mathbf{k} = \mathbf{v} = \{\mathbf{v}_k\}_{k=1}^{K}\right)$$

$$\mathbf{M}_{sdf}^{appr}(\mathbf{x}, \{\mathbf{f}_i\}_{i=1}^{N}) = SubAttention\left(\mathbf{q} = \mathbf{x}, \mathbf{k} = \mathbf{v} = \{\mathbf{f}_i\}_{i=1}^{N}\right)$$

$$\mathbf{M}_{sdf}^{occ}(\mathbf{x}) = SelfAttention\left(\mathbf{q} = \mathbf{k} = \mathbf{v} = \mathbf{x}\right)$$

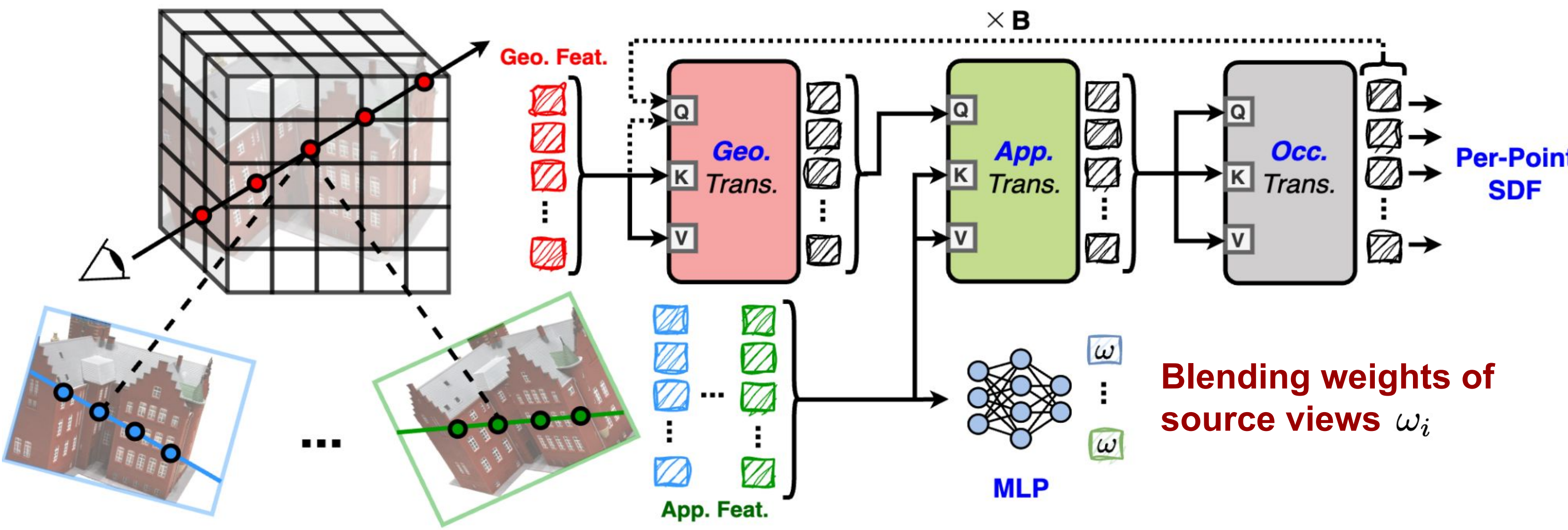
Omni-Recon: Detailed Backbone Design



The lightweight appear. branch: Model each 3D point's color by blending its 2D source view projections

$$\hat{\mathbf{c}} = \sum_{i=1}^{N} \omega_i \mathbf{c}_i$$

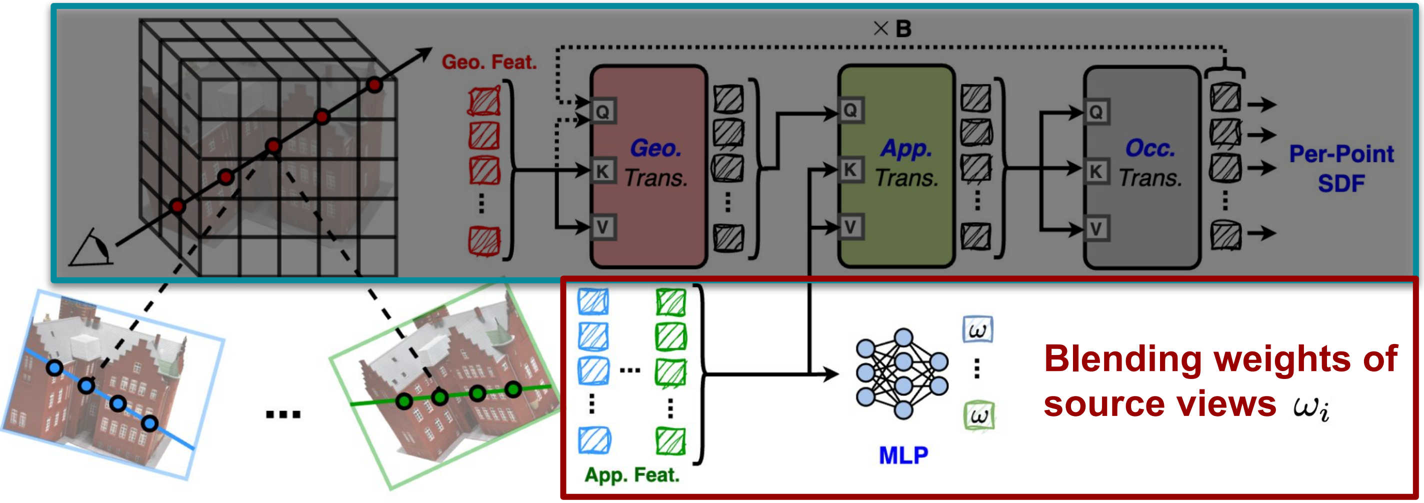
Omni-Recon: Pretraining in NeRF



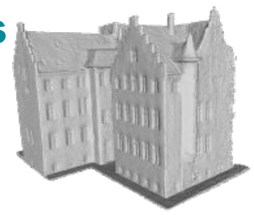
Pretrain on a set of scenes

Omni-Recon: Rendering with Mesh

Employ Marching Cube [1] for mesh extraction

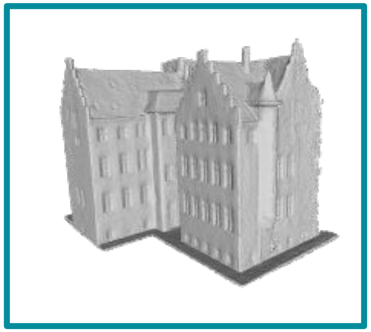


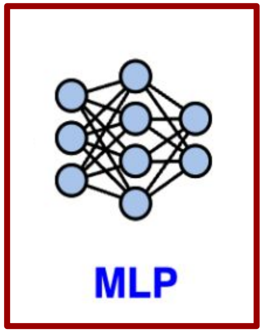
Baked into meshes



Fragment Shader

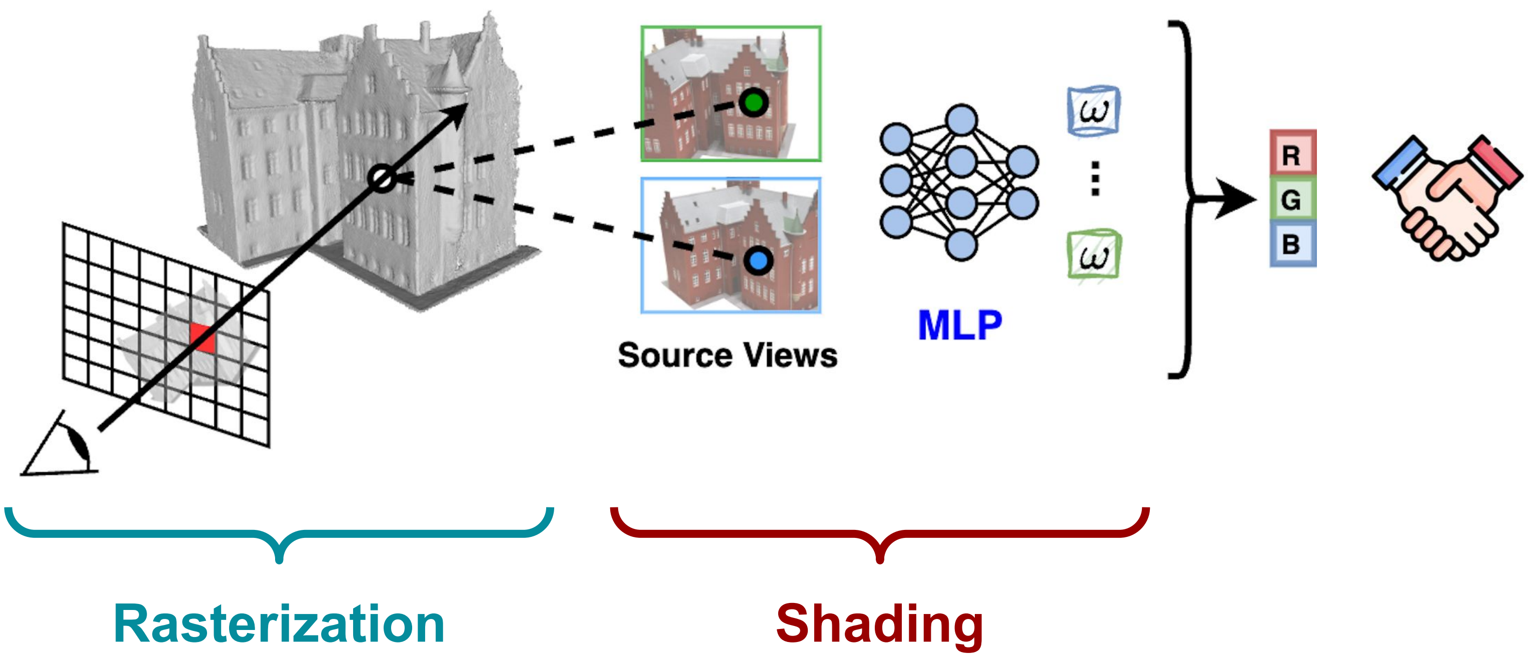
At rendering time:

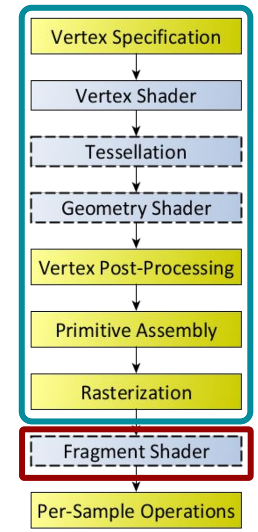




[1] "Marching cubes: A high resolution 3D surface construction algorithm", W. Lorensen et al., SIGGRAPH'87.

Omni-Recon: Rendering with Mesh



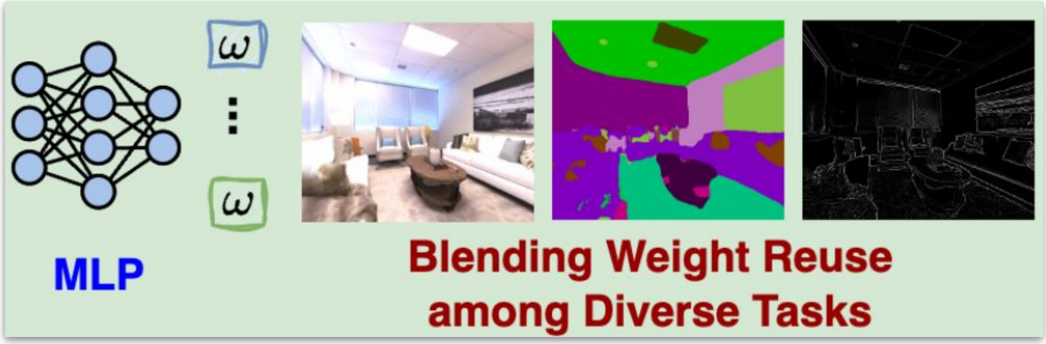


Supported by the GPU rasterization pipeline [1] for real-time rendering & rapid mesh finetuning

[1] "Modular Primitives for High-Performance Differentiable Rendering", S. Laine et al., ToG'20.

Omni-Recon: Achieve 3D Task Generality

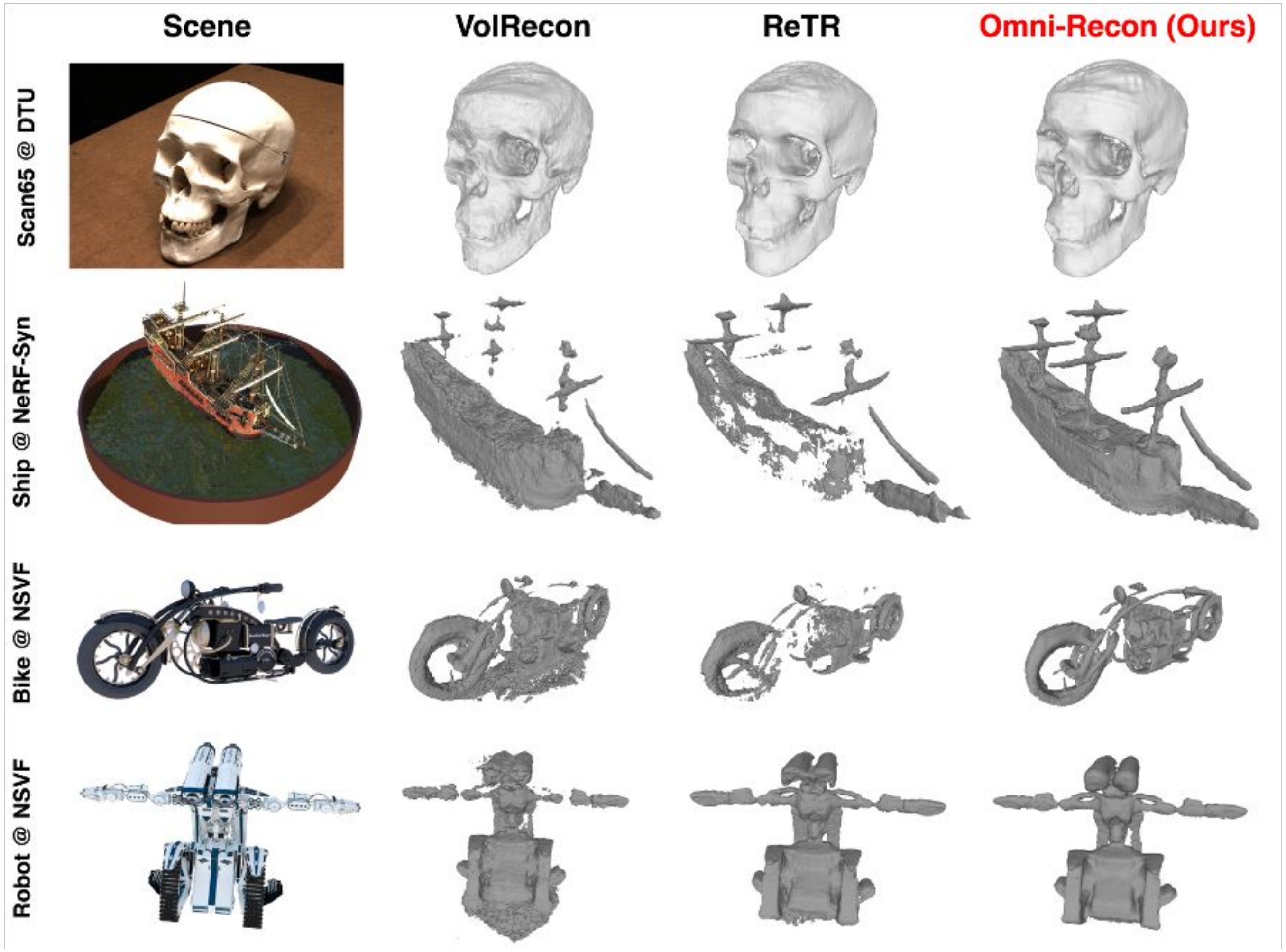
- **Zero-shot scene understanding:** Predict-then-Blend
  - Predict 2D properties of each source view
  - Lift to 3D via reusing the blending weight of RGB



3D scene understanding $\hat{\mathbf{p}} = \sum_{i=1}^{N} \underbrace{\omega_i}_{\text{Reuse from RGB}} \underbrace{\mathbf{p}_i}_{\text{Predict in 2D}}$

Omni-Recon: Experimental Results

- **Omni-Recon:** SOTA generalizable 3D surface extraction accuracy



Mesh reconstruction from 3 views of a new scene

Omni-Recon: Experimental Results

- **Omni-Recon:** SOTA generalizable 3D surface extraction accuracy

Method	Mean	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122
COLMAP [57]	1.52	**0.90**	2.89	1.63	1.08	2.18	1.94	1.61	<u>1.30</u>	2.34	1.28	1.10	1.42	0.76	1.17	1.14
MVSNet [83]	1.22	1.05	2.52	1.71	1.04	1.45	**1.52**	<u>0.88</u>	**1.29**	1.38	1.05	**0.91**	**0.66**	0.61	1.08	1.16
IDR [86]	3.39	4.01	6.40	3.52	1.91	3.96	2.36	4.85	1.62	6.37	5.97	1.23	4.73	0.91	1.72	1.26
VolSDF [84]	3.41	4.03	4.21	6.12	0.91	8.24	1.73	2.74	1.82	5.14	3.09	2.08	4.81	0.60	3.51	2.18
UNISURF [48]	4.39	5.08	7.18	3.96	5.30	4.61	2.24	3.94	3.14	5.63	3.40	5.09	6.38	2.98	4.05	2.81
NeuS [76]	4.00	4.57	4.49	3.97	4.32	4.63	1.95	4.68	3.83	4.15	2.50	1.52	6.47	1.26	5.57	6.11
PixelNeRF [88]	6.18	5.13	8.07	5.85	4.40	7.11	4.64	5.68	6.76	9.05	6.11	3.95	5.92	6.26	6.89	6.93
IBRNet [77]	2.32	2.29	3.70	2.66	1.83	3.02	2.83	1.77	2.28	2.73	1.96	1.87	2.13	1.58	2.05	2.09
MVSNeRF [10]	2.09	1.96	3.27	2.54	1.93	2.57	2.71	1.82	1.72	2.29	1.75	1.72	1.47	1.29	2.09	2.26
SparseNeuS [40]	1.96	2.17	3.29	2.74	1.67	2.69	2.42	1.58	1.86	1.94	1.35	1.50	1.45	0.98	1.86	1.87
VolRecon [55]	1.38	1.20	2.59	1.56	1.08	1.43	1.92	1.11	1.48	1.42	1.05	1.19	1.38	0.74	1.23	1.27
ReTR [34]	<u>1.17</u>	1.05	<u>2.31</u>	**1.44**	<u>0.98</u>	<u>1.18</u>	**1.52**	<u>0.88</u>	1.35	<u>1.30</u>	<u>0.87</u>	1.07	<u>0.77</u>	<u>0.59</u>	**1.05**	<u>1.12</u>
Omni-Recon (Ours)	**1.13**	<u>0.91</u>	**2.13**	<u>1.52</u>	**0.93**	**1.09**	<u>1.70</u>	**0.84**	**1.29**	**1.20**	**0.83**	<u>1.04</u>	0.81	**0.55**	**1.05**	**1.05**

Setting: Mesh reconstruction from 3 views of a new scene from DTU

Metric: Chamfer Distance (↓)

Omni-Recon: Experimental Results

- **Omni-Recon with mesh baking & finetuning**
  - Enable **real-time rendering (2458 × faster)**
  - Surpass generalizable recon. baselines with **a 10s finetuning**
  - **A +3.43 PSNR improvement** after 5min finetuning

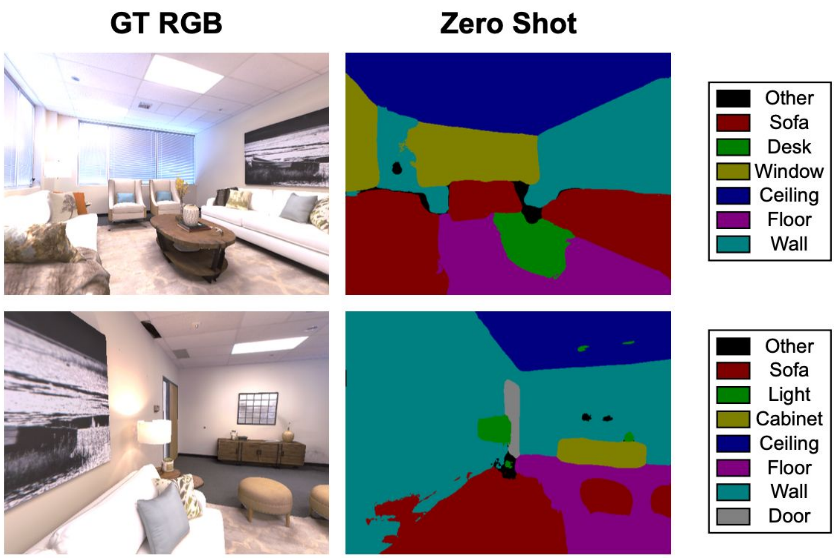
Method	FPS	Mean	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122
VolRecon [55]	0.029	24.58	22.33	20.59	21.53	23.72	24.2	23.65	24.47	22.77	23.54	22.62	26.89	27.44	25.76	30.14	29.19
ReTR [34]	0.024	25.59	24.32	21.84	23.4	24.56	26.31	24.5	24.63	24.3	24.58	23.85	27.84	27.97	26.76	30.03	28.96
Ours w/o ft.		22.96	20.12	19.71	22.27	22.78	24.55	21.77	21.51	26.72	22.33	24.49	22.52	22.93	24.68	23.84	24.12
Ours (ft. 10s)		25.68	21.42	21.68	24.06	24.12	28.19	24.10	23.95	31.65	24.41	28.15	25.63	25.85	26.15	26.82	26.89
Ours (ft. 20s)		27.21	22.63	22.92	25.12	25.42	30.03	25.95	26.16	33.19	26.22	30.36	27.44	27.04	26.93	29.15	29.65
Ours (ft. 30s)	71.3 (40.82)	27.78	23.2	23.26	25.54	25.7	30.59	26.83	26.96	33.66	26.47	30.73	28.14	27.70	27.1	30.17	30.65
Ours (ft. 1min)		28.34	24.69	24.11	25.76	26.05	30.93	27.66	27.49	33.68	27.07	30.97	28.51	28.54	27.35	31.17	31.54
Ours (ft. 3min)		28.95	25.2	24.32	**25.94**	26.16	32.15	28.99	**27.88**	34.94	**27.35**	31.62	28.93	28.97	27.56	31.70	32.49
Ours (ft. 5min)		**29.02**	**25.34**	**24.36**	25.63	**26.21**	**32.16**	**29.33**	27.81	**34.94**	27.32	**31.74**	**29.04**	**29.05**	**27.69**	**31.74**	**32.89**

Rendering PSNR (↑) on test scenes @ DTU

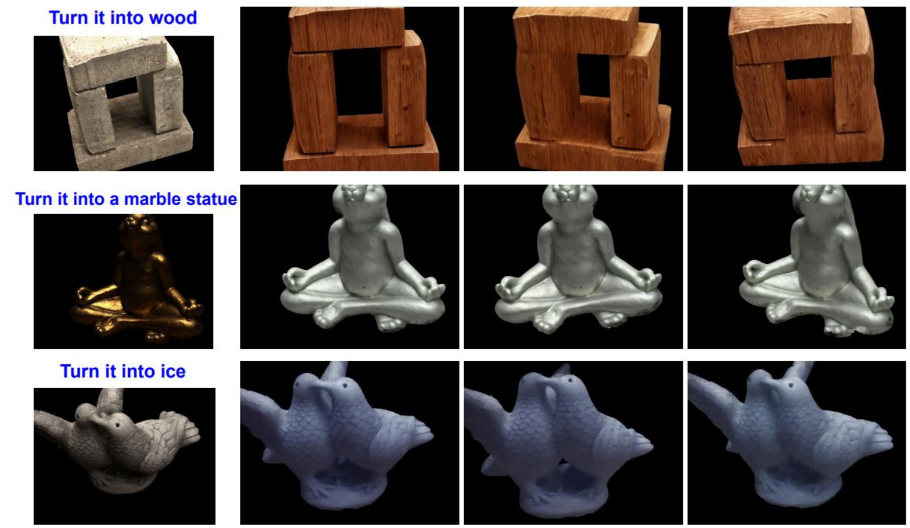
FPS measured on an NVIDIA RTX 2080Ti GPU

Omni-Recon: Experimental Results

- **Omni-Recon:** Support diverse 3D understanding & editing tasks leveraging our rendering pipeline



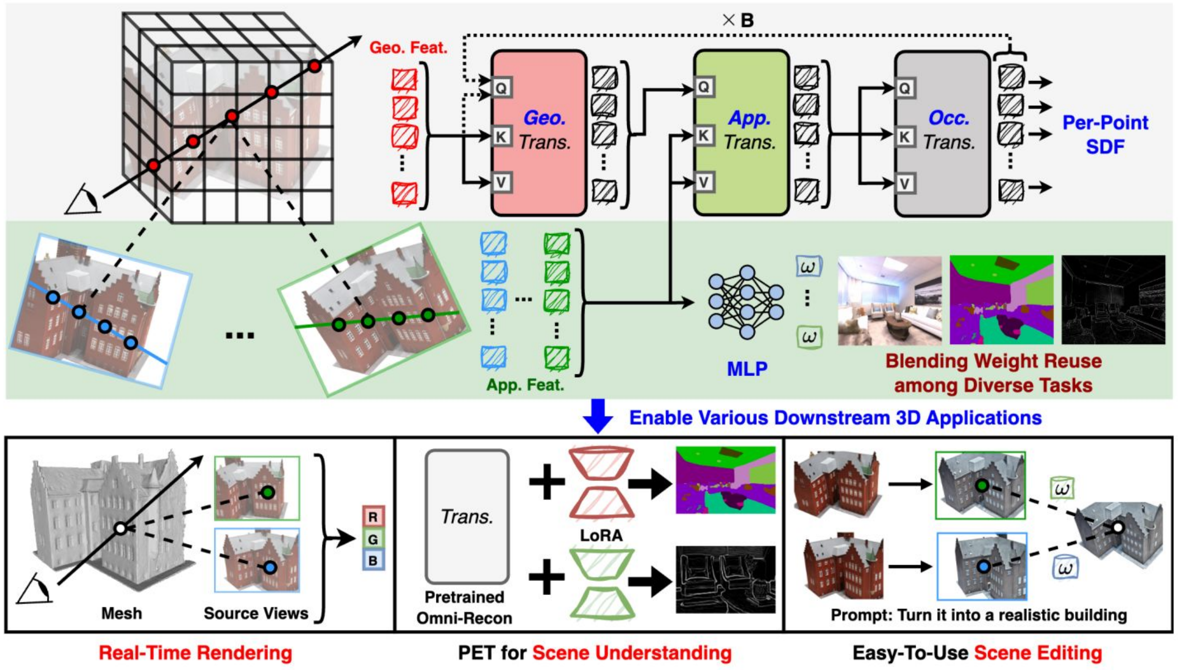
Language-driven open-set semantic segmentation



3D scene editing

Omni-Recon: Key Takeaways

- **Pretraining in image-based NeRF** and **rendering with mesh** could win both recons. and rendering efficiency
- The correlation between appear. and scene properties makes the **zero-shot 2D-to-3D task lifting** feasible



Omni-Recon: Harnessing Image-based Rendering for General-Purpose Neural Radiance Fields

ECCV 2024 Oral

Yonggan Fu, Huaizhi Qu, Zhifan Ye, Chaojian Li, Kevin Zhao,

Yingyan (Celine) Lin