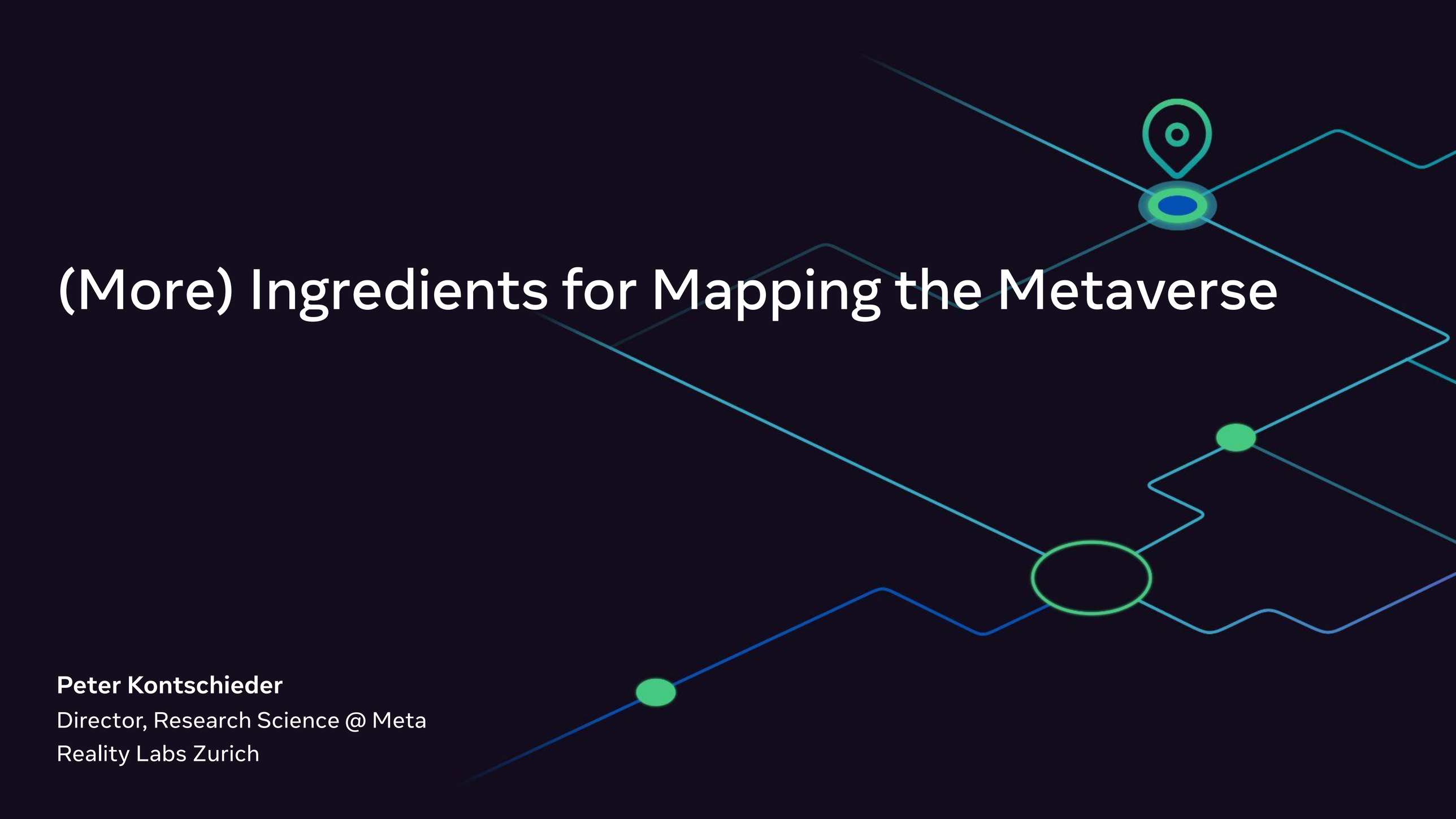


(More) Ingredients for Mapping the Metaverse

Peter Kontschieder
Director, Research Science @ Meta
Reality Labs Zurich

OBJECTIVE

Develop the next generation of CV/ML algorithms for building high-fidelity and holistic 3D semantic scenes from images for the metaverse.


Jonathon Luiten


Katja Schwarz


Duncan Zauss


Norman Mueller


Andrea Simonelli

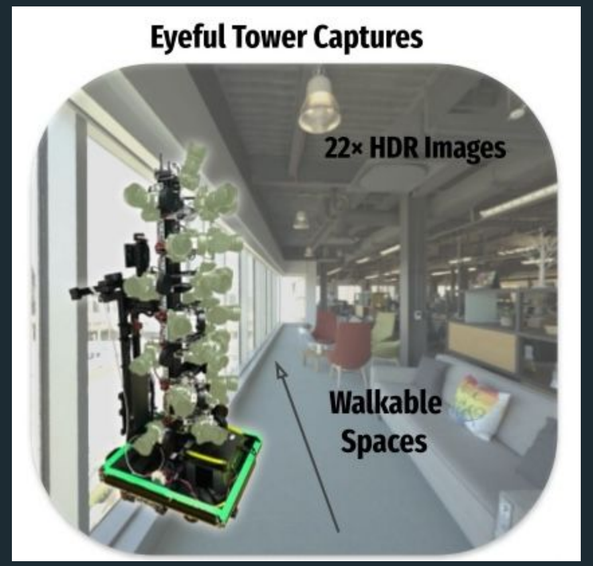

Arno Knapitsch


Samuel Rota Bulò


Lorenzo Porzi


Corinne Stucker


Peter Kontschieder



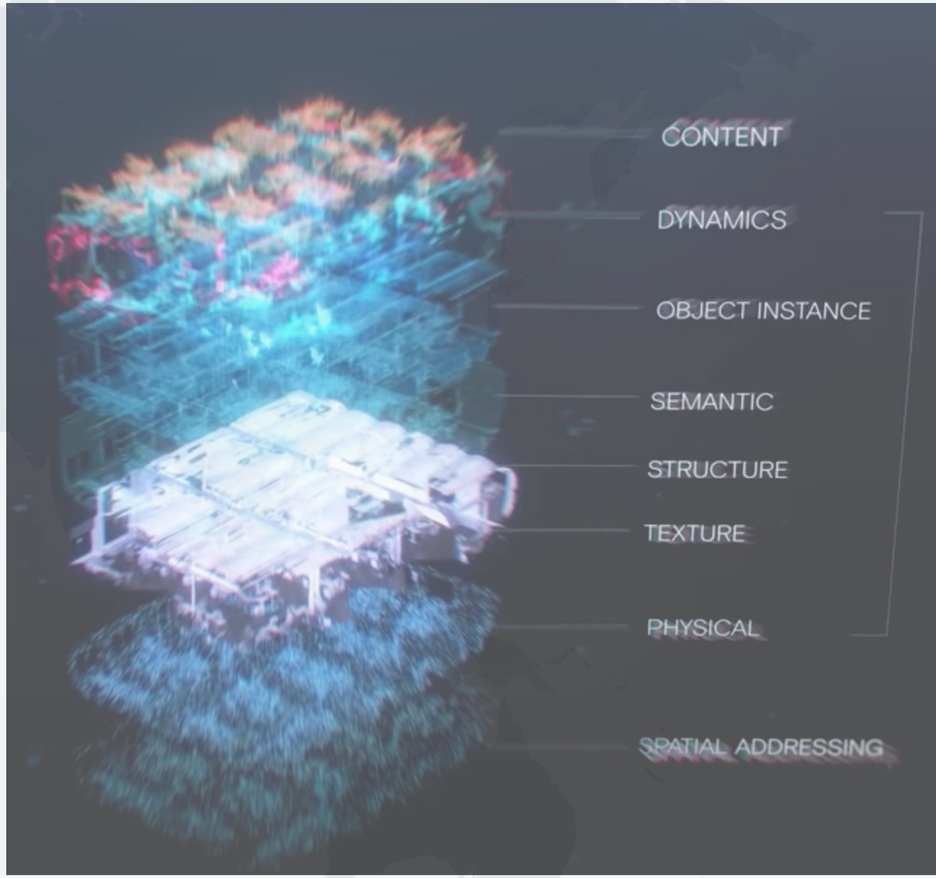
Xu et al. The VR-NeRF Eyeful Tower Dataset. ACM SIGGRAPH Asia 2023.

Mapillary. The Swiss National Museum. 2024

Data
Indoor &
outdoor

Algorithms
Performance,
robustness,
editability, …

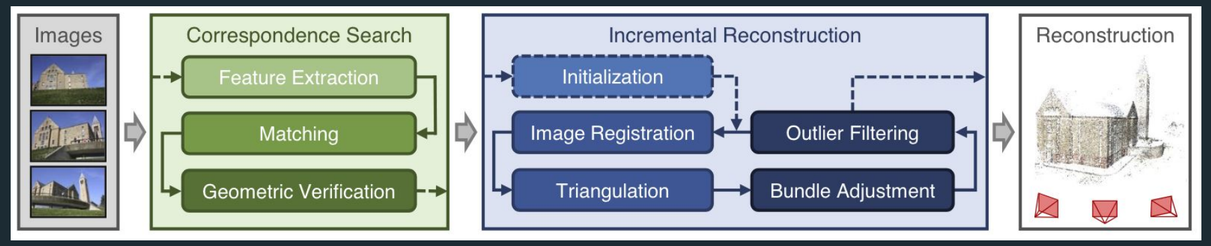
Representation
Efficiency,
photorealism,
dynamics

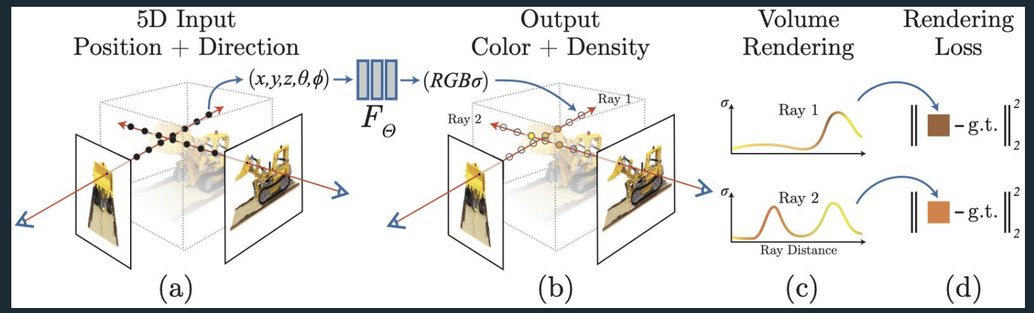


Before we start...

some prerequisites

Structure-from-Motion and Neural Radiance Fields





Schönberger et al. Pixelwise View Selection for Unstructured Multi-View Stereo. ECCV'16.

Mildenhall et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV'20.

Object Recognition for Scene Understanding

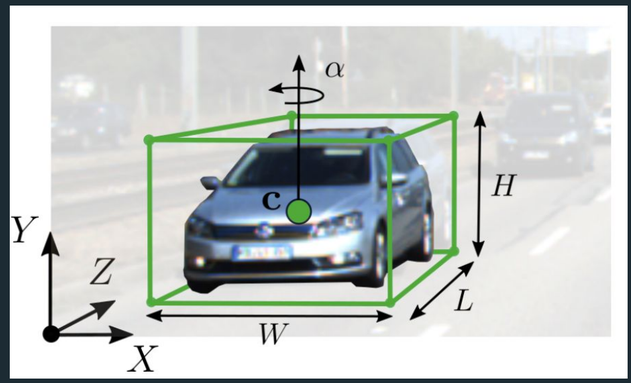
Panoptic Segmentation

Segment an image into *things* and *stuff*

- **Things**: Countable objects
- **Stuff**: Amorphous areas (road, sky)

Porzi et al. Improving Panoptic Segmentation at All Scales. CVPR'21.

Monocular 3D Object Detection

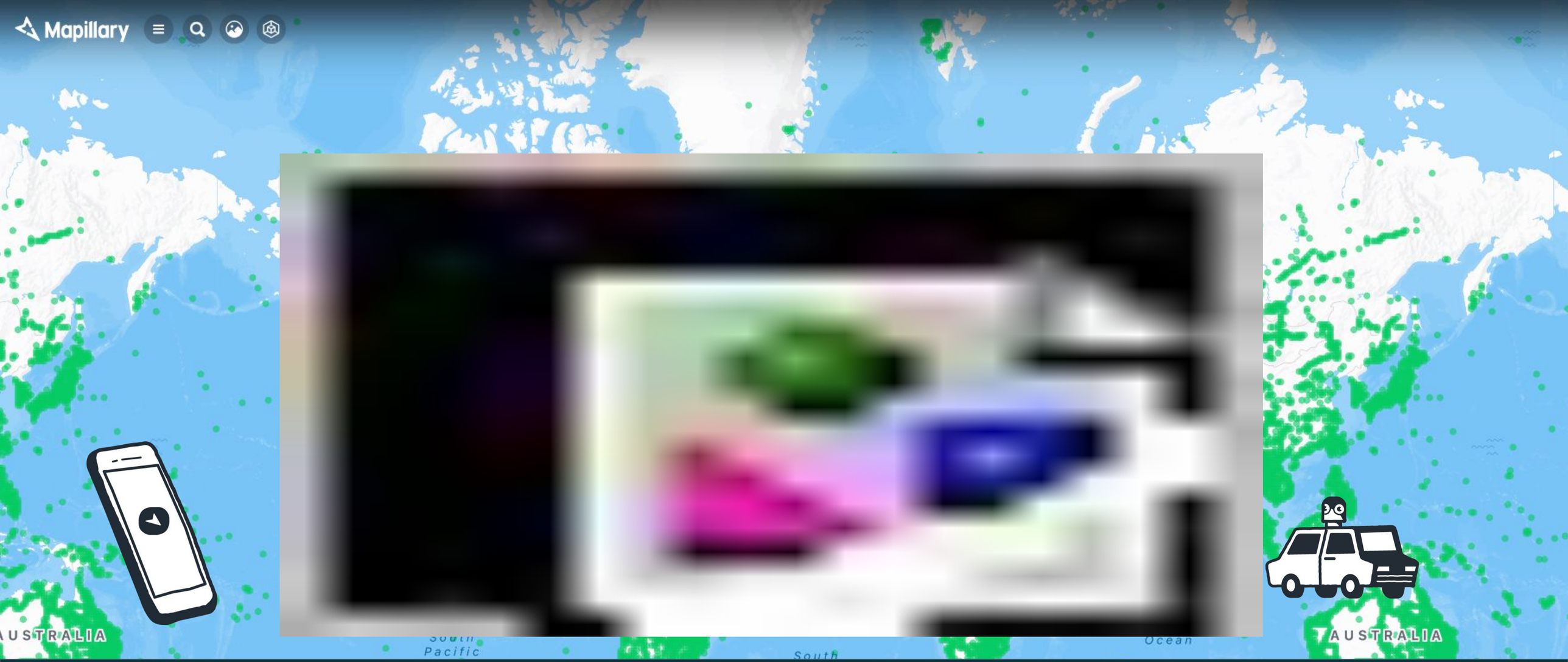


Input: Single RGB image & camera parameters

Output: 3D bounding boxes in camera coords

Simonelli et al. Are we Missing Confidence in Pseudo-LiDAR Methods for Monocular 3D Object Detection?. ICCV'21.

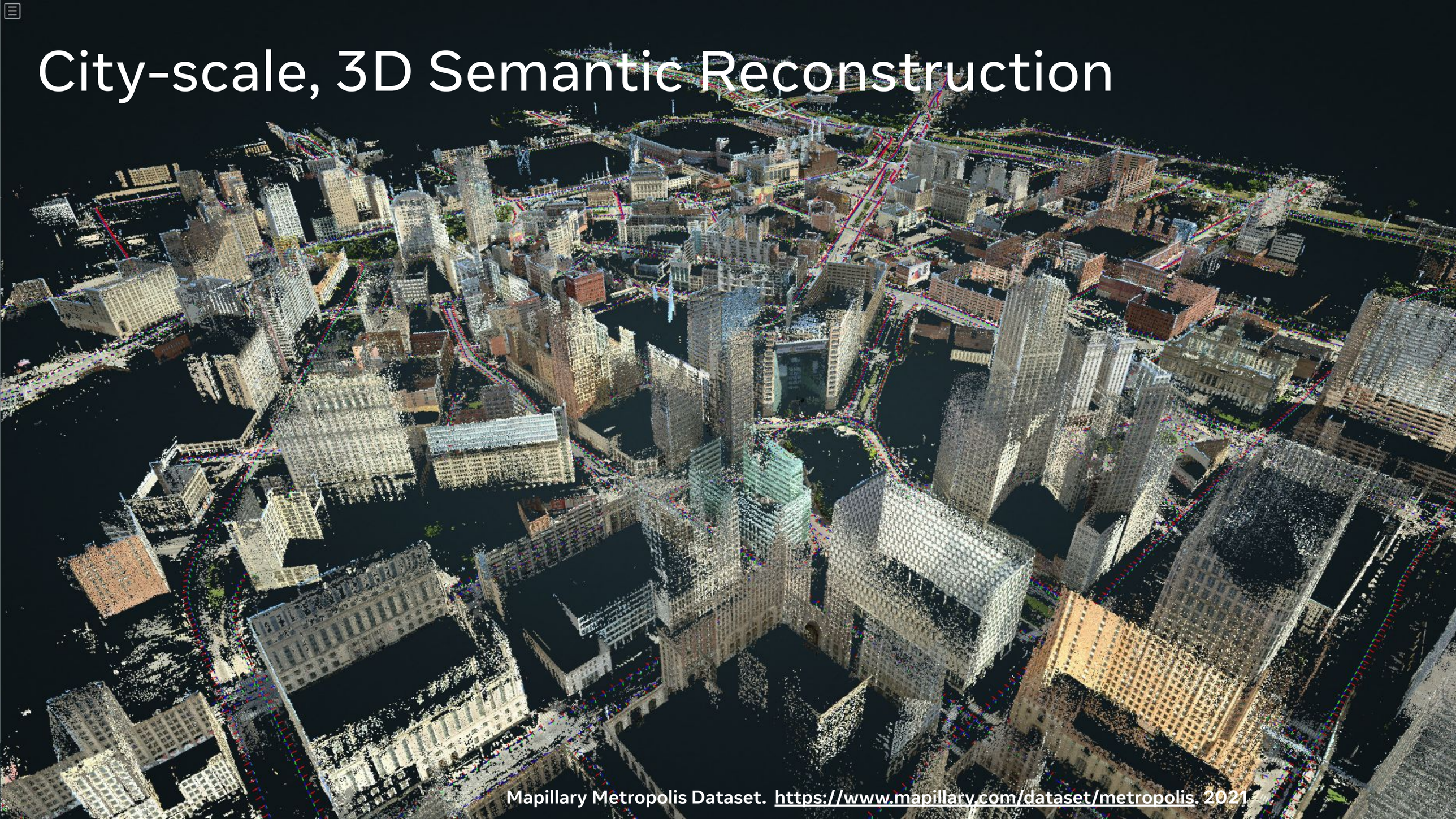
Challenges we’ve taken on!



Mapillary

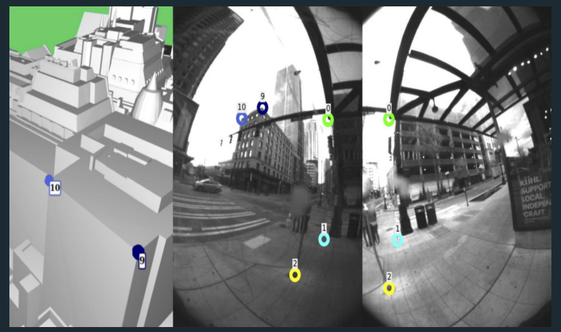
Building a digital twin of the world

City-scale, 3D Semantic Reconstruction



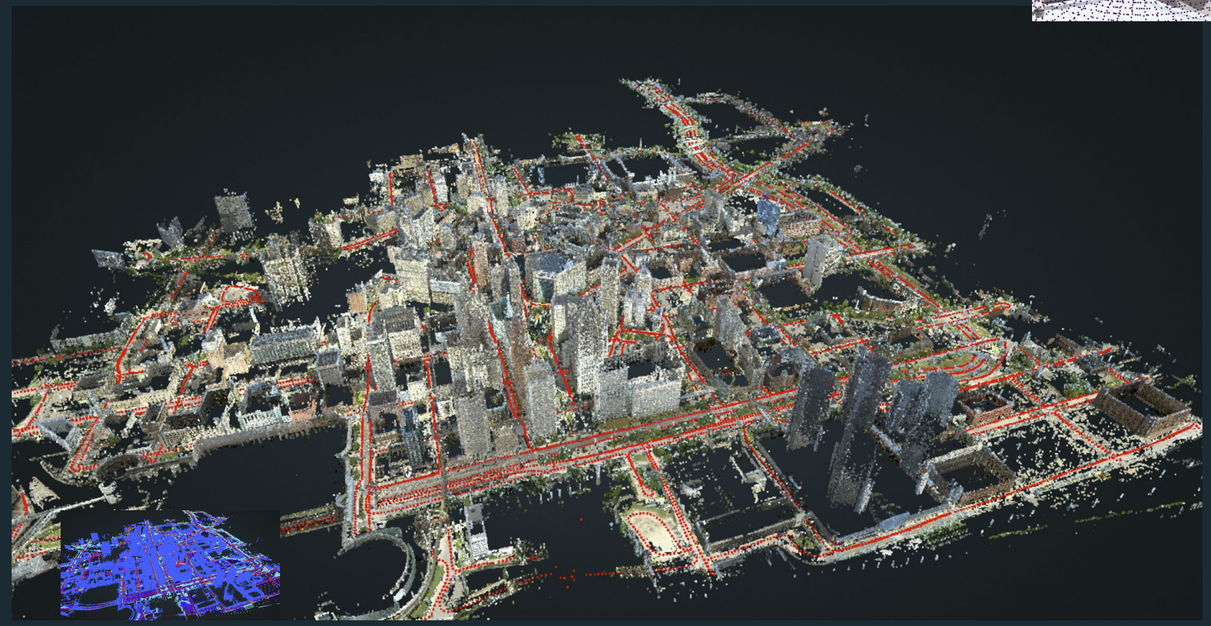
Mapillary Metropolis Dataset. https://www.mapillary.com/dataset/metropolis. 2021

Registered CAD models

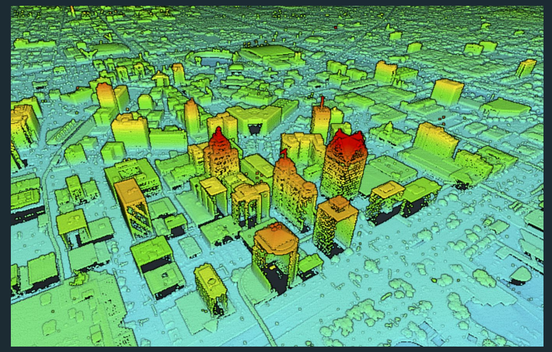


Human-annotated 2D, 3D, tracking, etc.

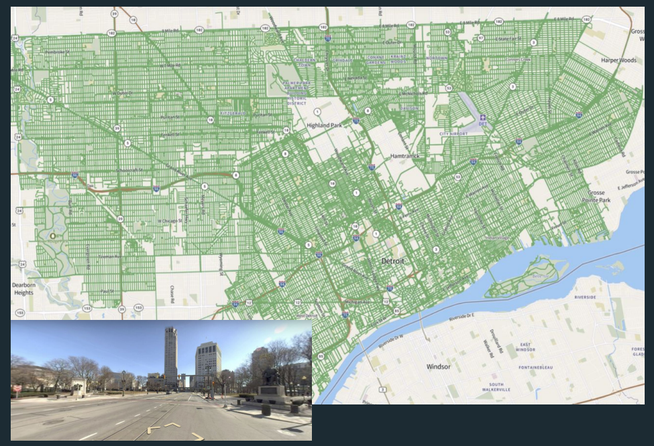
Registered point cloud data

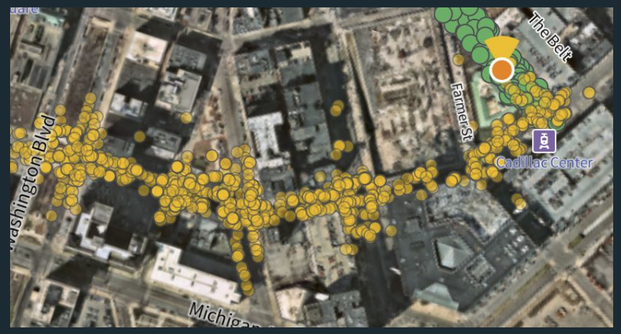


Aerial LiDAR+DEM



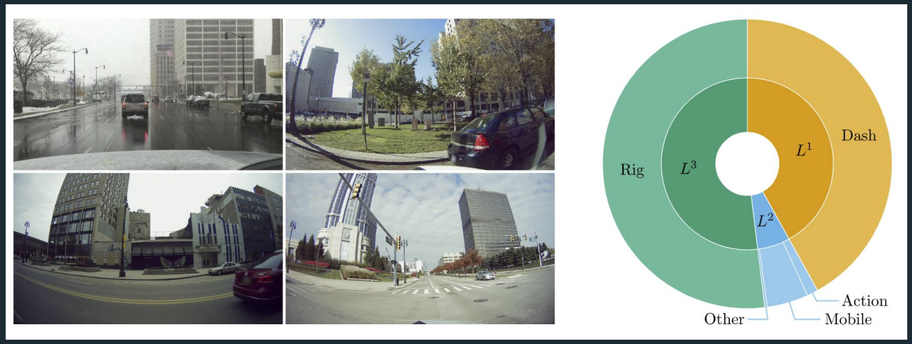
2.5M High-res 360° images >25MP





Aerial Imagery

Precisely measured ground control pts



2.3M consumer-grade & low-cost images

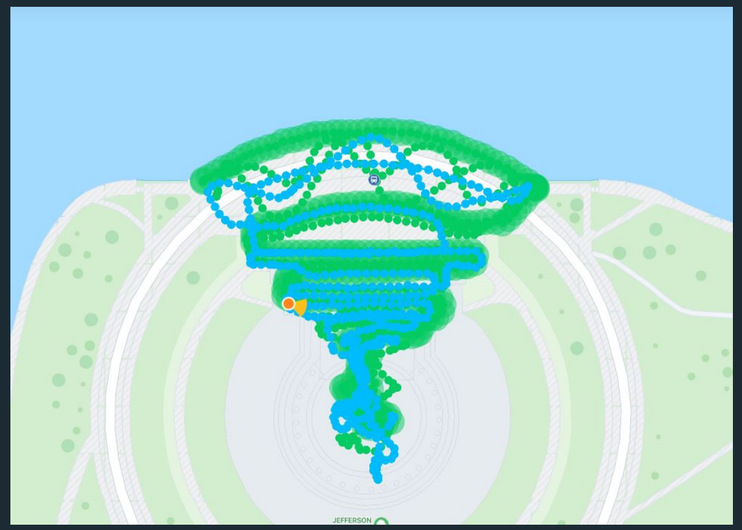
NeRFs for Mapillary - Challenges Everywhere?

Neural Radiance Fields (NeRFs)

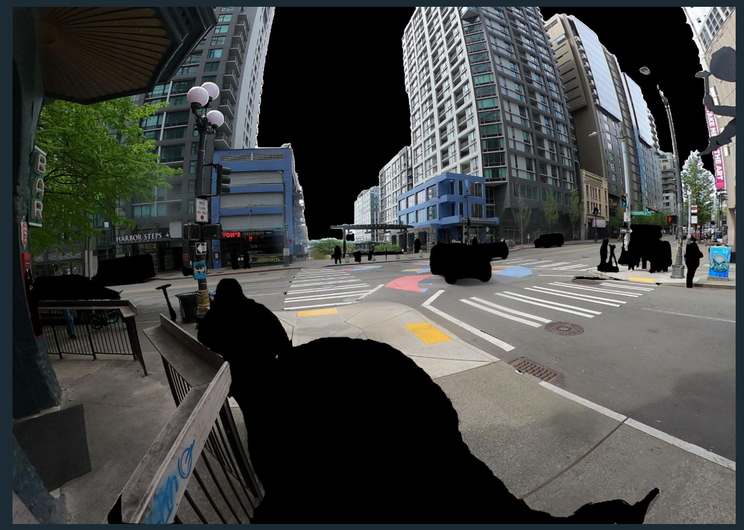
Encode scenes as volumes of density & color in a Neural Net

+ Photo-realistic quality for novel view synthesis
+ Able to model view-dependent effects like reflections
– Difficulties with dynamics (pedestrians, trees, clouds, …)
– Very dependent on high capture coverage from all angles
– Extremely picky about correct camera poses
– Challenging wrt real-time rendering

NeRFs for Mapillary - Taking Pragmatic Choices

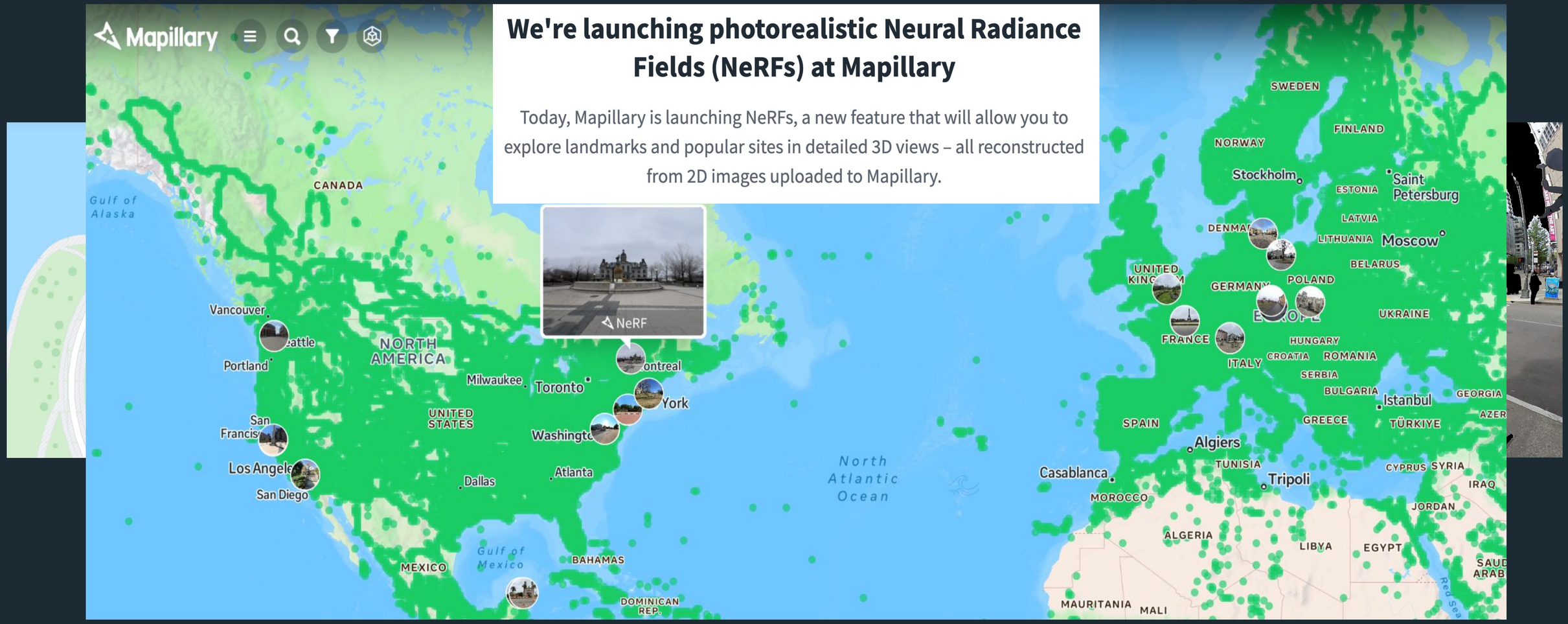


Find supportive data



Leverage semantics

NeRFs for Mapillary - Launch it!

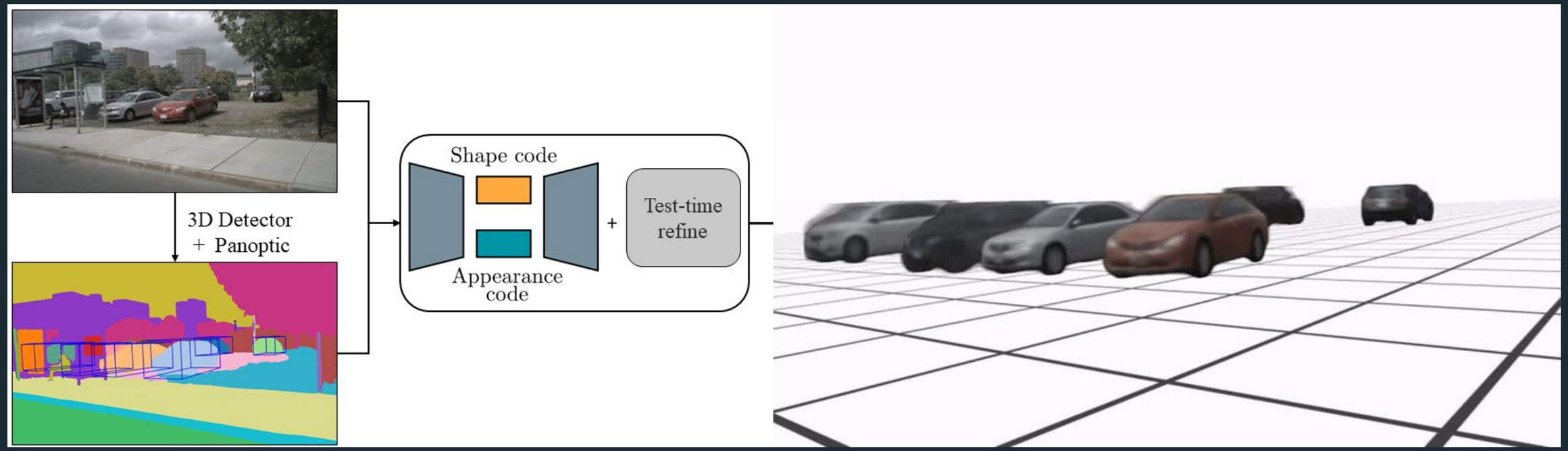







3D Objects from Single View Observations

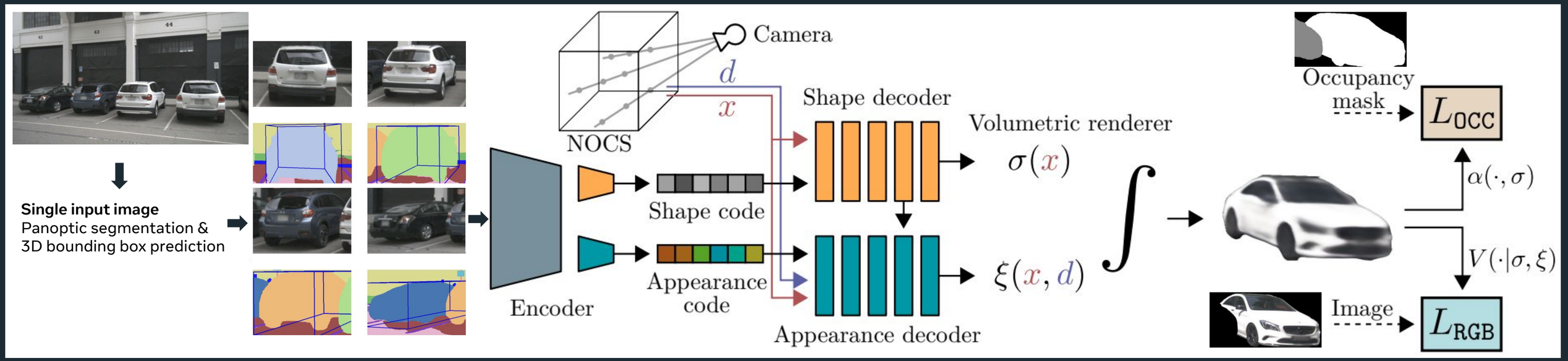
Single-shot novel view synthesis from street-level images



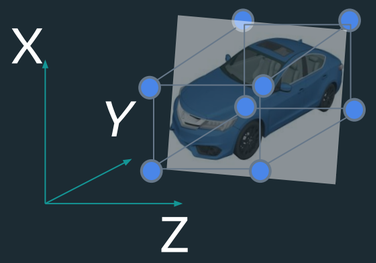
Müller et al. AutoRF: Learning 3D Object Radiance Fields from Single View Observations. CVPR'22

3D Objects from Single View Observations

AutoRF: Basic training pipeline

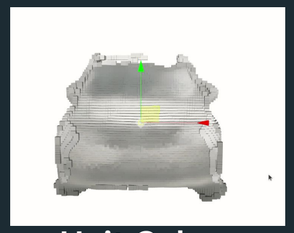


Camera Space

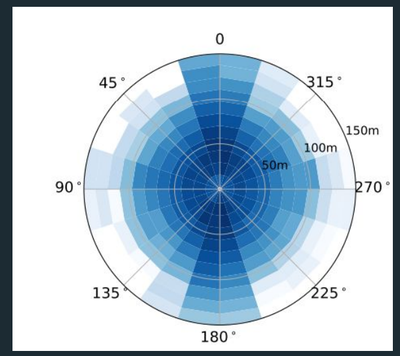


Transformation
f (3D Box)

Normalised Space



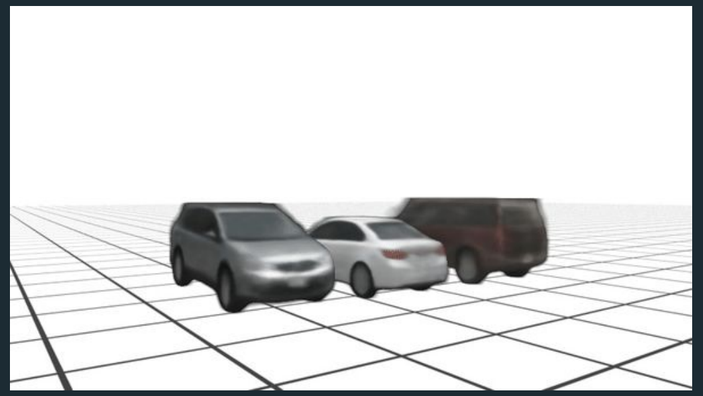
Unit Cube
$[-0.5, +0.5]^3$

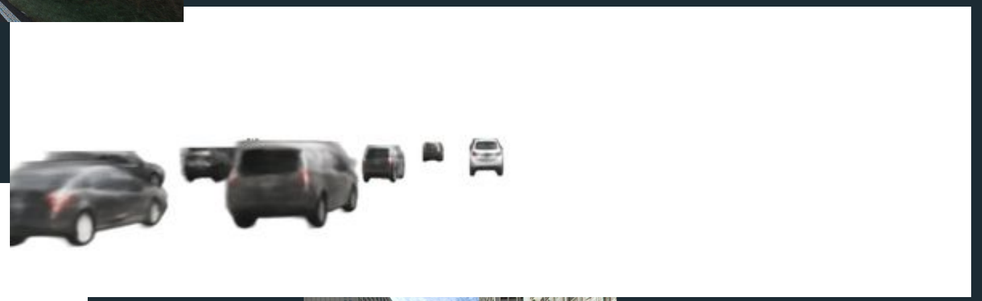


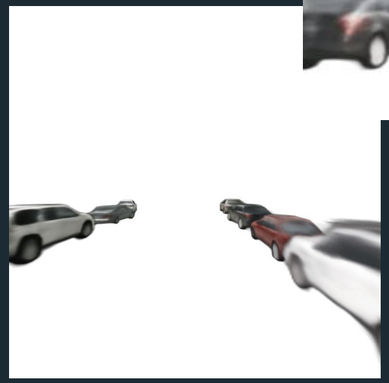
Viewpoint distance and orientation coverage for nuScenes cars

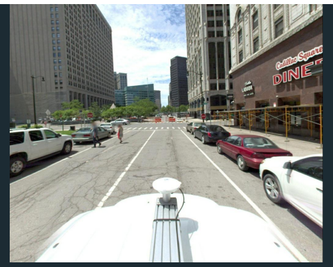
Qualitative Results: Novel-view synthesis, Editing

Unseen images from nuScenes, Kitti, Metropolis datasets









Jointly modeling static and dynamic scenes?

In-the-wild captures like street scene data typically contains fast object motion and limited viewpoint coverage.

Many previous focus on reconstructions of static scene elements only, obtaining incomplete reconstruction results

Tancik et al. Block-NeRF: Scalable Large Scene Neural View Synthesis. CVPR’22.

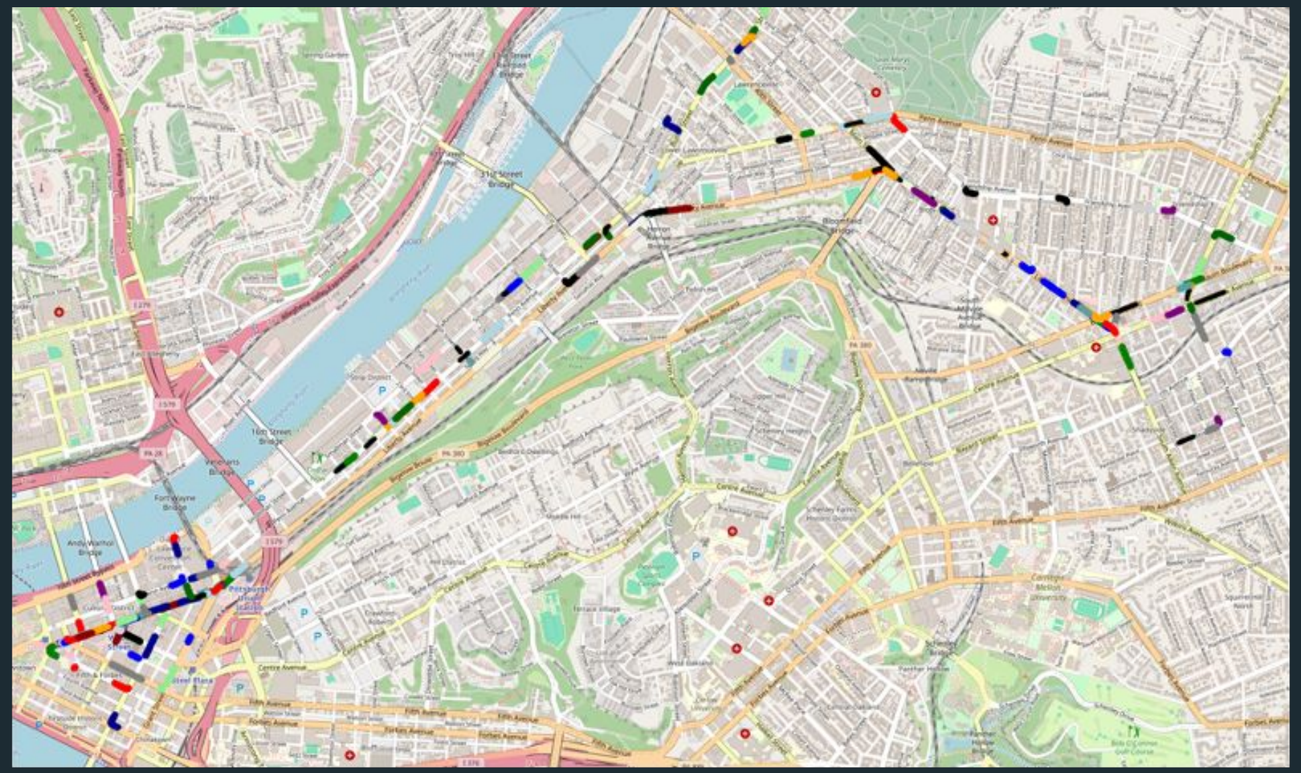
Benchmark Setup

Based on Argoverse 2 Vehicle Fleet Dataset

- Different weather, season, time of day
- Synchronized sensors:
  - 7 global shutter cameras
  - LiDAR
  - GPS

→ 2 geographic regions

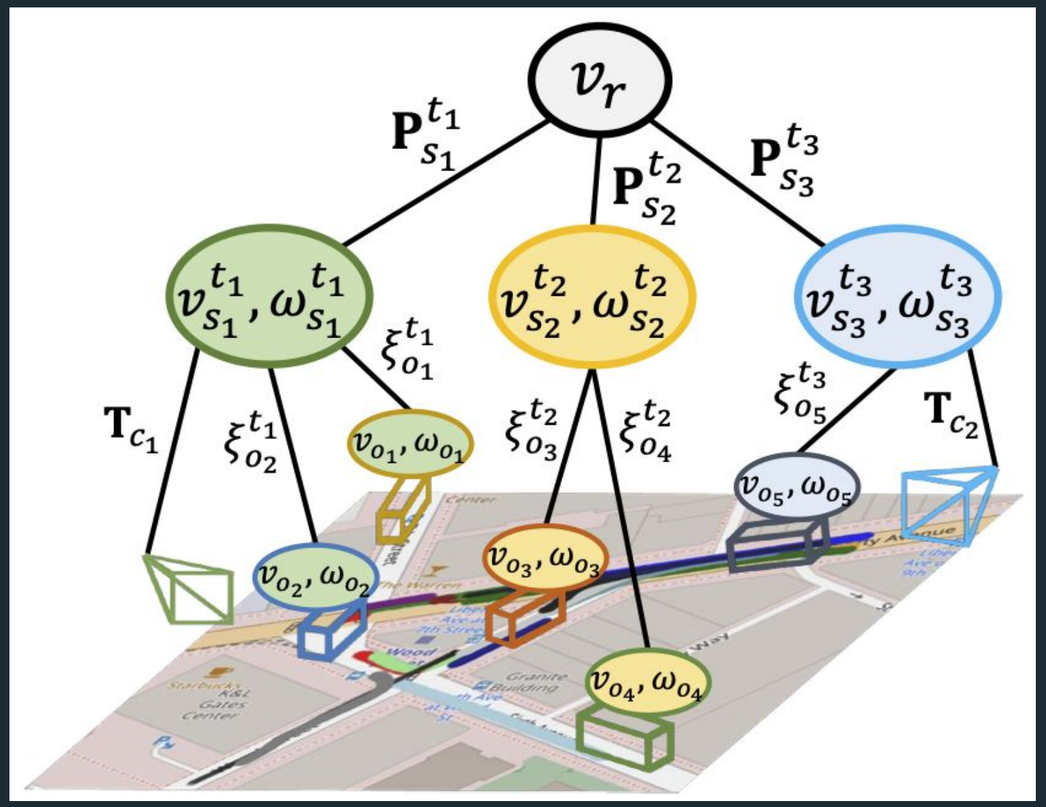
37 vehicle captures



Wilson et al.Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv, 2023.

Multi-Level Neural Scene Graphs for Dynamic Urban Environments

We represent sequences captured from moving vehicles in a shared geographic area with a multi-level scene graph



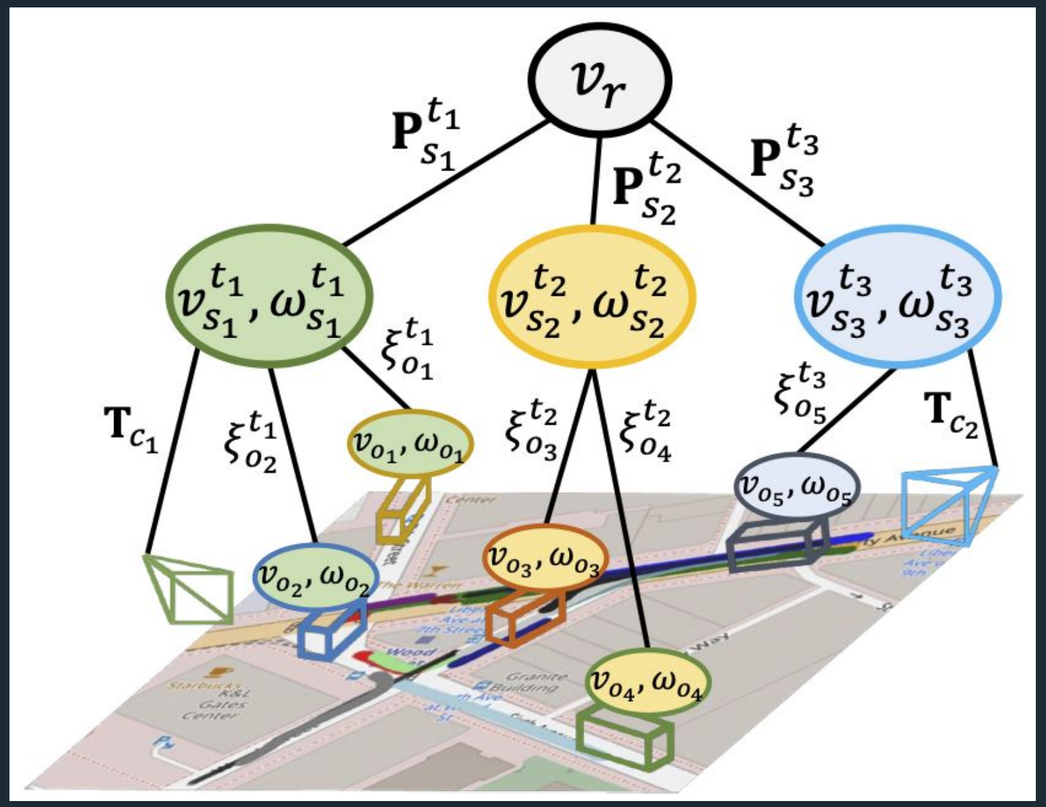
- For each sequence $s \in S$
  - Timesteps $t \in T_s$
  - Ego-vehicle poses $\mathbf{P}_s^t$
  - Cameras C_s calibrated w.r.t. ego-vehicle via $\mathbf{T}_c$
- For each object $o \in O_s$
  - Poses w.r.t. ego-vehicle ξ_o^t
  - Object dimensions s_o

Goal: Estimate radiance field $f_\theta(\mathbf{x}, \mathbf{d}, t, s) = (\sigma(\mathbf{x}, t, s), \mathbf{c}(\mathbf{x}, \mathbf{d}, t, s))$

Fischer et al. Multi-Level Neural Scene Graphs for Dynamic Urban Environments. CVPR 2024

Multi-Level Neural Scene Graphs for Dynamic Urban Environments

We represent sequences captured from moving vehicles in a shared geographic area with a multi-level scene graph



We create a scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Nodes $\mathcal{V}$
  - Sequence nodes v_s^t
    - Latent code ω_s^t
  - Object nodes v_o
    - Latent code ω_o
  - Camera nodes v_c
- Edges $\mathcal{E}$
  - Rigid transformations $e_{v_s^t \rightarrow v_r} = \mathbf{P}_s^t, \ldots$

- Given $\mathcal{G}$, we model f_θ with
  - Static radiance field: $\phi(\mathbf{x}, \mathbf{d}, \omega_s^t)$
  - Dynamic radiance field: $\psi(\mathbf{x}, \mathbf{d}, \omega_s^t, \omega_o)$

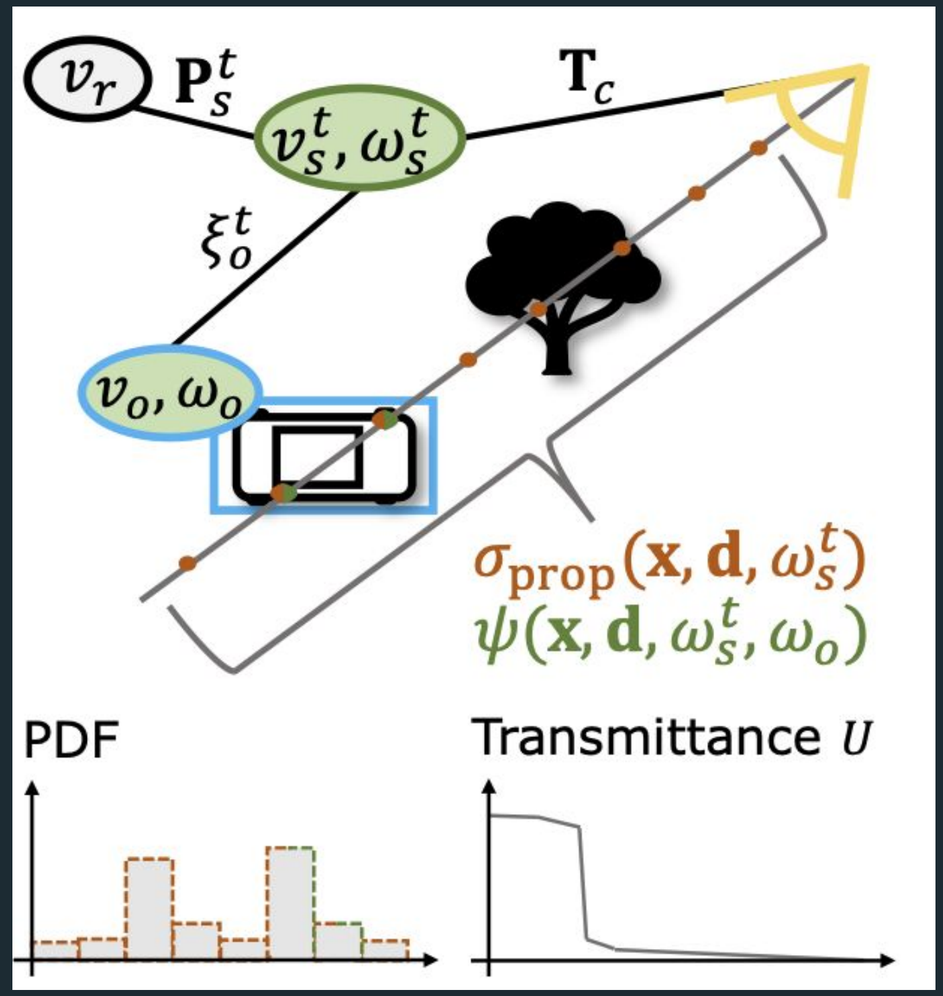
Fischer et al. Multi-Level Neural Scene Graphs for Dynamic Urban Environments. CVPR 2024

Efficient Scene Graph Rendering

- Efficient CUDA implementation for ray-box intersection [u_in, u_out]

- Composite proposal sampling
  - Efficient proposal network σ_{prop}
  - Composite density: σ_{prop} + more expensive radiance field ψ

- 2x proposal sampling before rendering



Fischer et al. Multi-Level Neural Scene Graphs for Dynamic Urban Environments. CVPR 2024

Experimental Results

Rendering novel trajectories - including a slow-motion part

SUDS - Scalable Urban Dynamic Scenes

Ours

Turki et al. SUDS - Scalable Urban Dynamic Scenes. CVPR 2023
Fischer et al. Multi-Level Neural Scene Graphs for Dynamic Urban Environments. CVPR 2024

Experimental Results (cont’d)

Mix + match for changing scene appearance and configurations

Fischer et al. Multi-Level Neural Scene Graphs for Dynamic Urban Environments. CVPR 2024

Experimental Results (cont'd)

Method	Residential			Downtown			Mean			Train time (h)
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
Nerfacto + Emb.	19.83	0.637	0.562	18.05	0.655	0.625	18.94	0.646	0.594	8.0
Nerfacto + Emb. + Time	20.05	0.641	0.562	18.66	0.656	0.603	19.36	0.654	0.583	13.2
SUDS [61]	21.76	0.659	0.556	19.91	0.665	0.645	20.84	0.662	0.601	54.8
Ours	**22.29**	**0.678**	**0.523**	**20.01**	**0.681**	**0.586**	**21.15**	**0.680**	**0.555**	17.2

Argoverse 2 results (top). Kitti and Virtual Kitti dataset results (bottom).

Method	KITTI [75%]			KITTI [50%]			KITTI [25%]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [32]	18.56	0.557	0.554	19.12	0.587	0.497	18.61	0.570	0.510
NeRF + Time	21.01	0.612	0.492	21.34	0.635	0.448	19.55	0.586	0.505
NSG [36]	21.53	0.673	0.254	21.26	0.659	0.266	20.00	0.632	0.281
Nerfacto + Emb.	22.75	0.801	0.156	22.38	0.793	0.160	21.24	0.758	0.178
Nerfacto + Emb. + Time	23.19	0.804	0.155	23.18	0.803	0.155	21.98	0.777	0.172
SUDS [61]	22.77	0.797	0.171	23.12	0.821	0.135	20.76	0.747	0.198
Ours	**28.38**	**0.907**	**0.052**	**27.51**	**0.898**	**0.055**	**26.51**	**0.887**	**0.060**

Method	VKITTI2 [75%]			VKITTI2 [50%]			VKITTI2 [25%]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [32]	18.67	0.548	0.634	18.58	0.544	0.635	18.17	0.537	0.644
NeRF + Time	19.03	0.574	0.587	18.90	0.565	0.610	18.04	0.545	0.626
NSG [36]	23.41	0.689	0.317	23.23	0.679	0.325	21.29	0.666	0.317
Nerfacto + Emb.	22.15	0.847	0.145	21.88	0.843	0.148	21.28	0.827	0.155
Nerfacto + Emb. + Time	22.11	0.849	0.144	21.78	0.844	0.147	21.00	0.825	0.157
SUDS [61]	23.87	0.846	0.150	23.78	0.851	0.142	22.18	0.829	0.160
Ours	**29.73**	**0.912**	**0.065**	**29.19**	**0.906**	**0.066**	**28.29**	**0.901**	**0.067**

Fischer et al. Multi-Level Neural Scene Graphs for Dynamic Urban Environments. CVPR 2024

Sneak Preview: Dynamic 3D Gaussian Fields

Single dynamic scene representation from heterogeneous input sequences for NVS at interactive speed.

Fischer et al. Dynamic 3D Gaussian Fields. arXiv. 2024

Panoptic Lifting

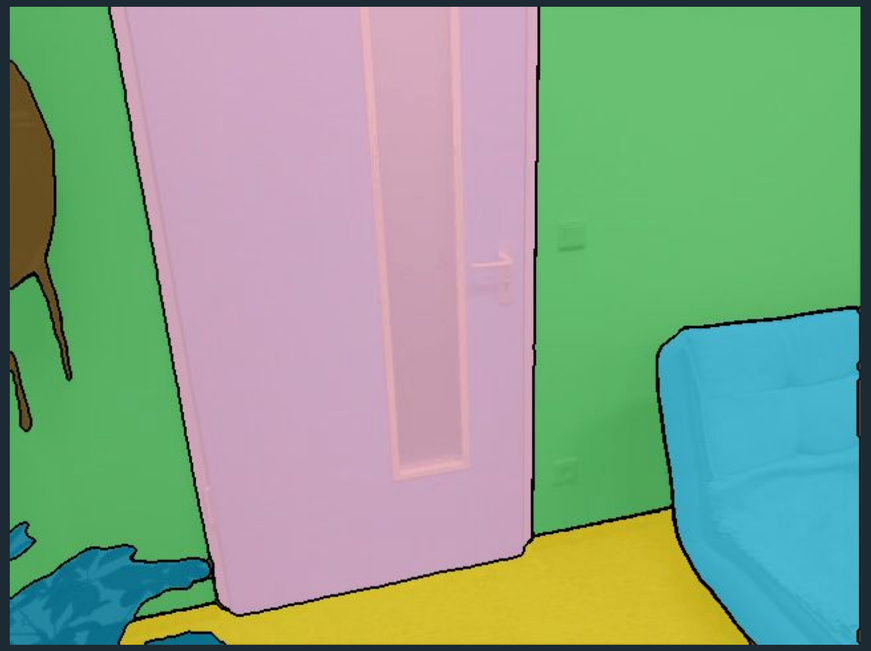
Reconstructing panoptic 3D volumetric representations from posed RGB images

Siddiqui et al. Panoptic Lifting for 3D Scene Understanding with Neural Fields. CVPR'23

Panoptic Lifting

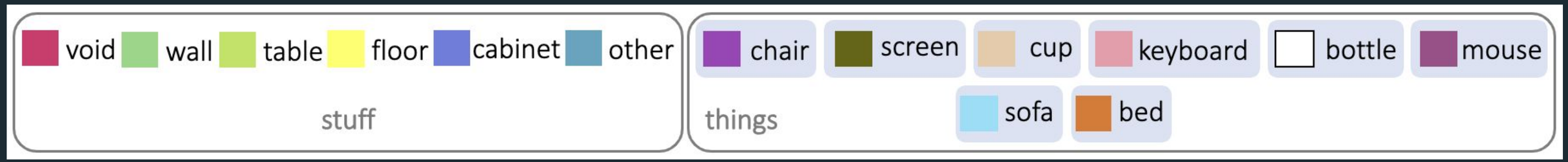
Leveraging 2D panoptic segmentation requires handling of inconsistencies

RGB input



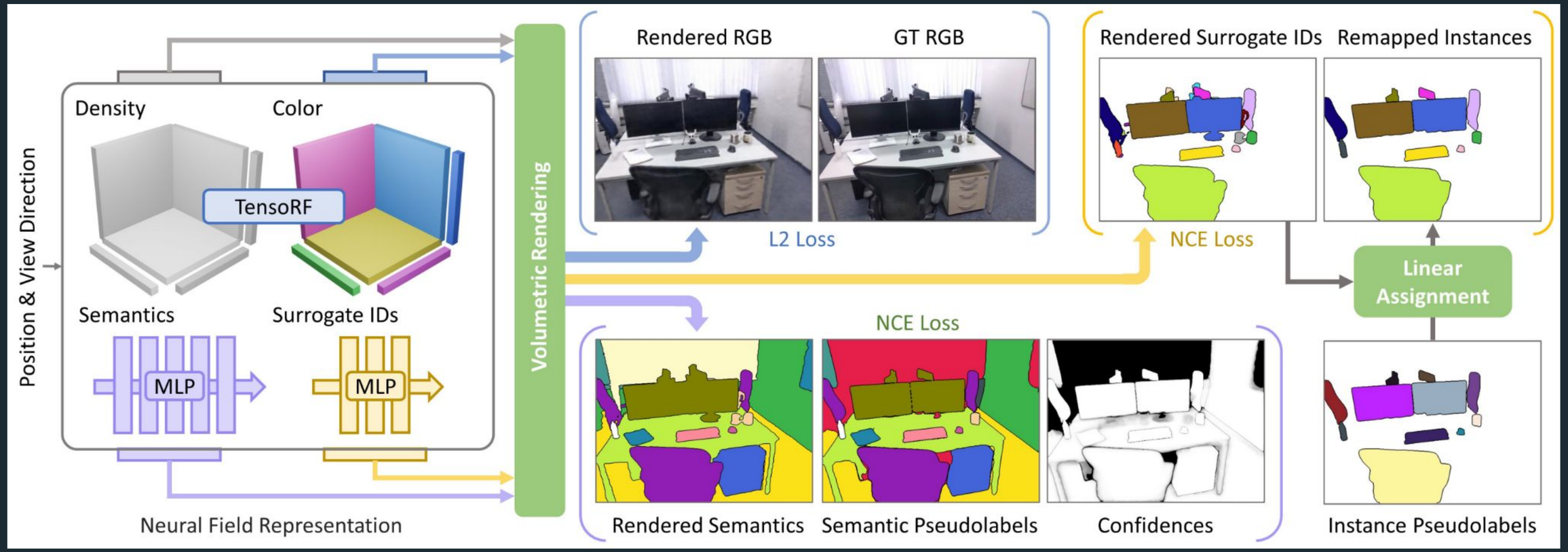
2D semantic segmentation

2D instance segmentation



Panoptic Lifting

Basic training pipeline



Siddiqui et al. Panoptic Lifting for 3D Scene Understanding with Neural Fields. CVPR'23

Panoptic Lifting

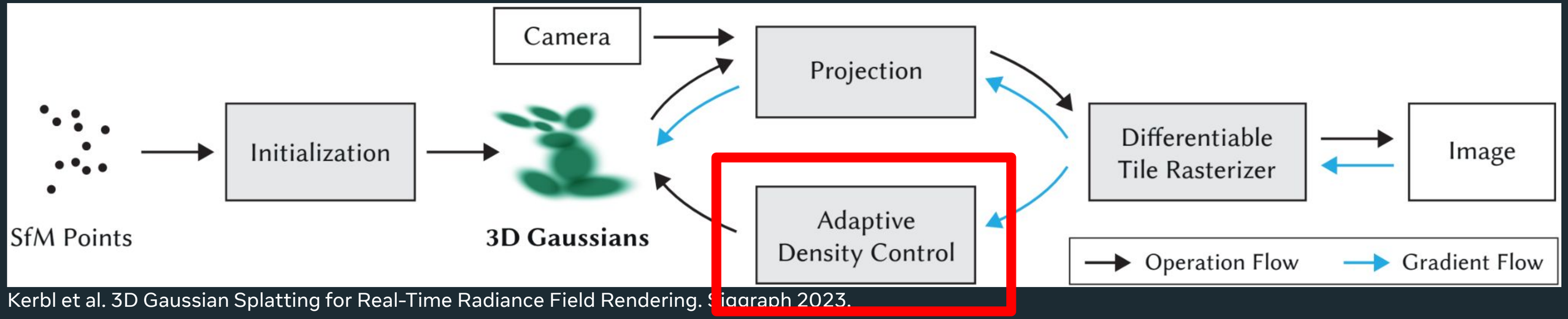
Further refinement leveraging GroundingDino + Segment Anything

3D Gaussian Splatting

Gaussian primitive. A Gaussian primitive $\gamma_k := (\boldsymbol{\mu}_k, \Sigma_k, \alpha_k, \boldsymbol{f}_k)$ geometrically resembles a 3D Gaussian kernel

$$\mathcal{G}_k(\boldsymbol{x}) := \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)$$

centered in $\boldsymbol{\mu}_k \in \mathbb{R}^3$ and having Σ_k as its 3×3 covariance matrix. Each primitive additionally entails an opacity factor $\alpha_k \in [0,1]$ and a feature vector $\boldsymbol{f}_k \in \mathbb{R}^d$ (*e.g.* RGB color or spherical harmonics coefficients).



Kerbl et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. Siggraph 2023.

3D Gaussian Splatting: Adaptive Density Control

ADC handles densification and pruning in 3DGS:

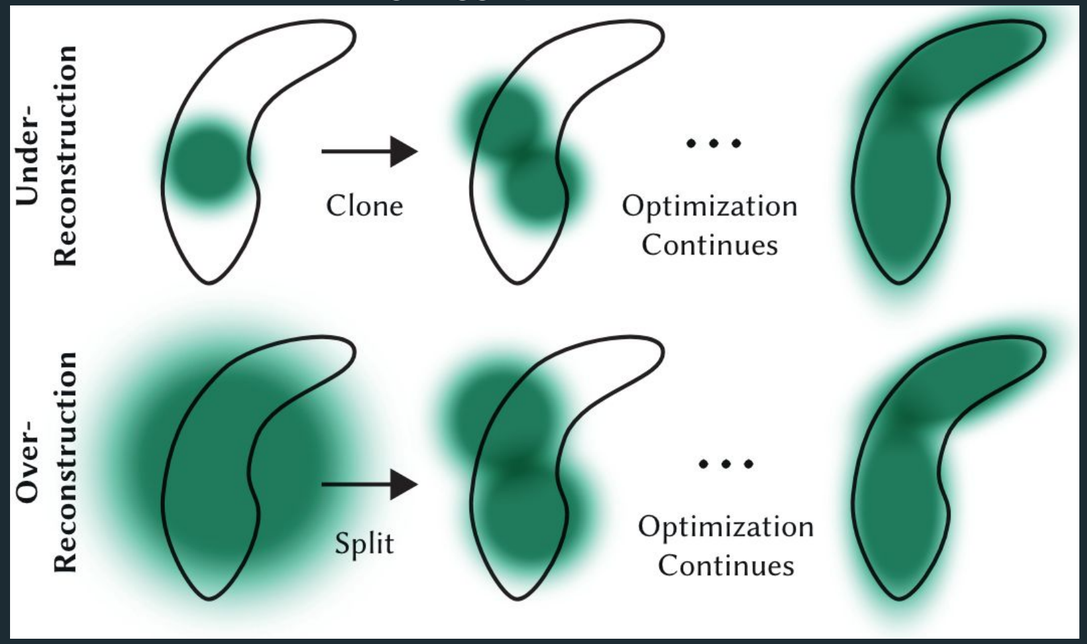
Growing depends on accumulated positional gradients of existing primitives and is conditioned on the size of the Gaussians (ie., cloning small ones, splitting large ones)

Pruning is activated when opacity drops below a threshold

Issues:

- It's non-intuitive to estimate a threshold for a gradient magnitude-based quantity
- Fails in areas where few large Gaussians model high-frequency patterns
- Lacks control for max. number of Gaussians

Kerbl et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. Siggraph 2023.



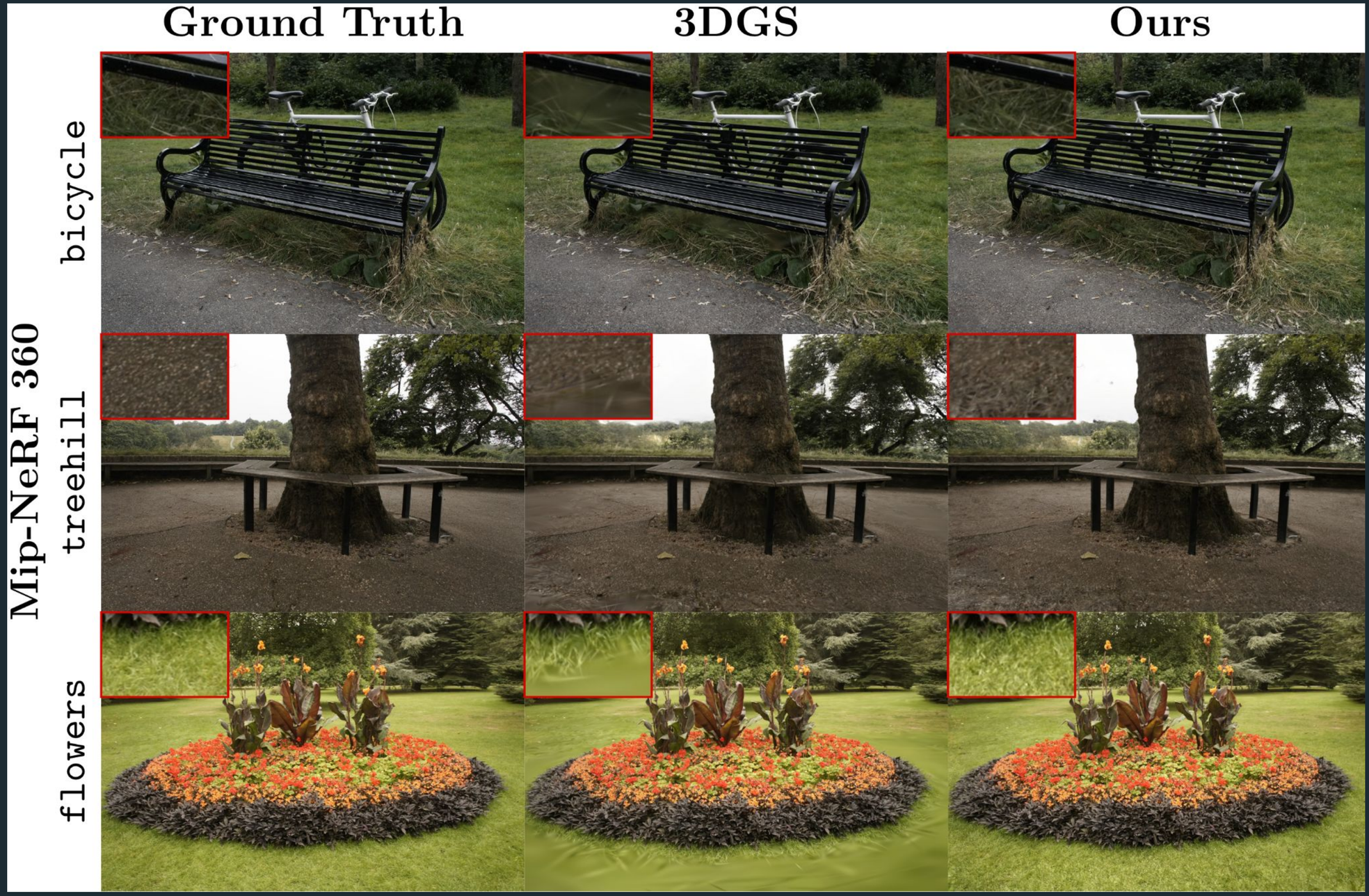
Rota Bulò et al. Revising Densification in Gaussian Splatting. arXiv 2024

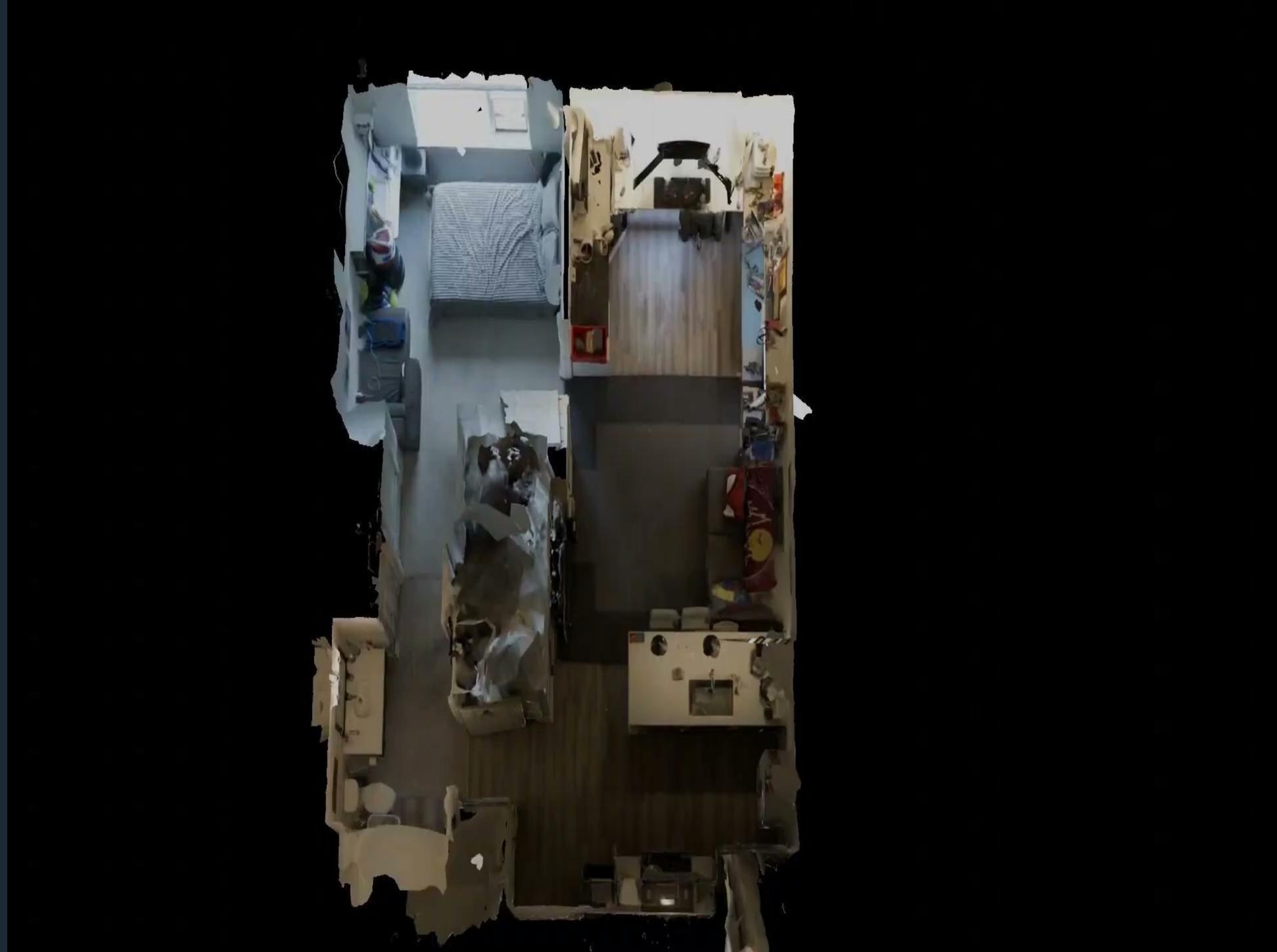
3D Gaussian Splatting: Error-based densification

We propose an auxiliary, per-pixel error function (Structural similarity based)

Turning per-pixel errors into per-Gaussian primitive errors. This is done by re-distributing per-pixel errors proportionally to each Gaussian wrt contributions to the rendered pixels.

Rota Bulò et al. Revising Densification in Gaussian Splatting. arXiv 2024


Ground Truth
3DGS
Ours
Mip-NeRF 360
bicycle
treehill
flowers



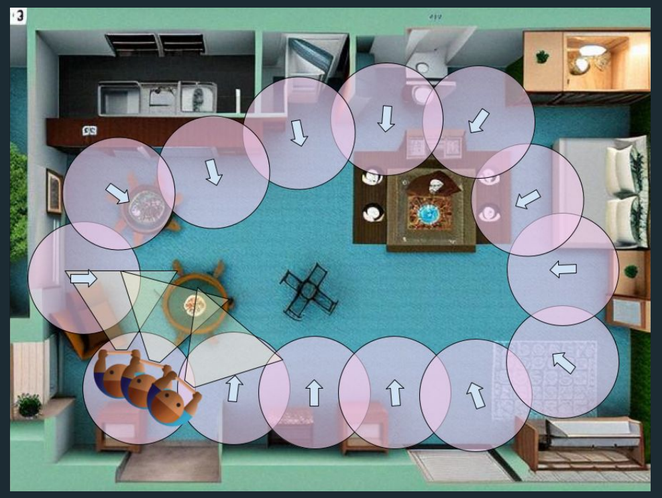
Xu et al. The VR-NeRF Eyeful Tower Dataset. ACM SIGGRAPH Asia 2023.

Data Capture vs. Content Generation

Let’s assume the representation problem is solved… What’s next?

Only Captured Data
(Reconstruction)

Only Generated Data
(Synthesis)



Very strong dependencies on input data quality and coverage!

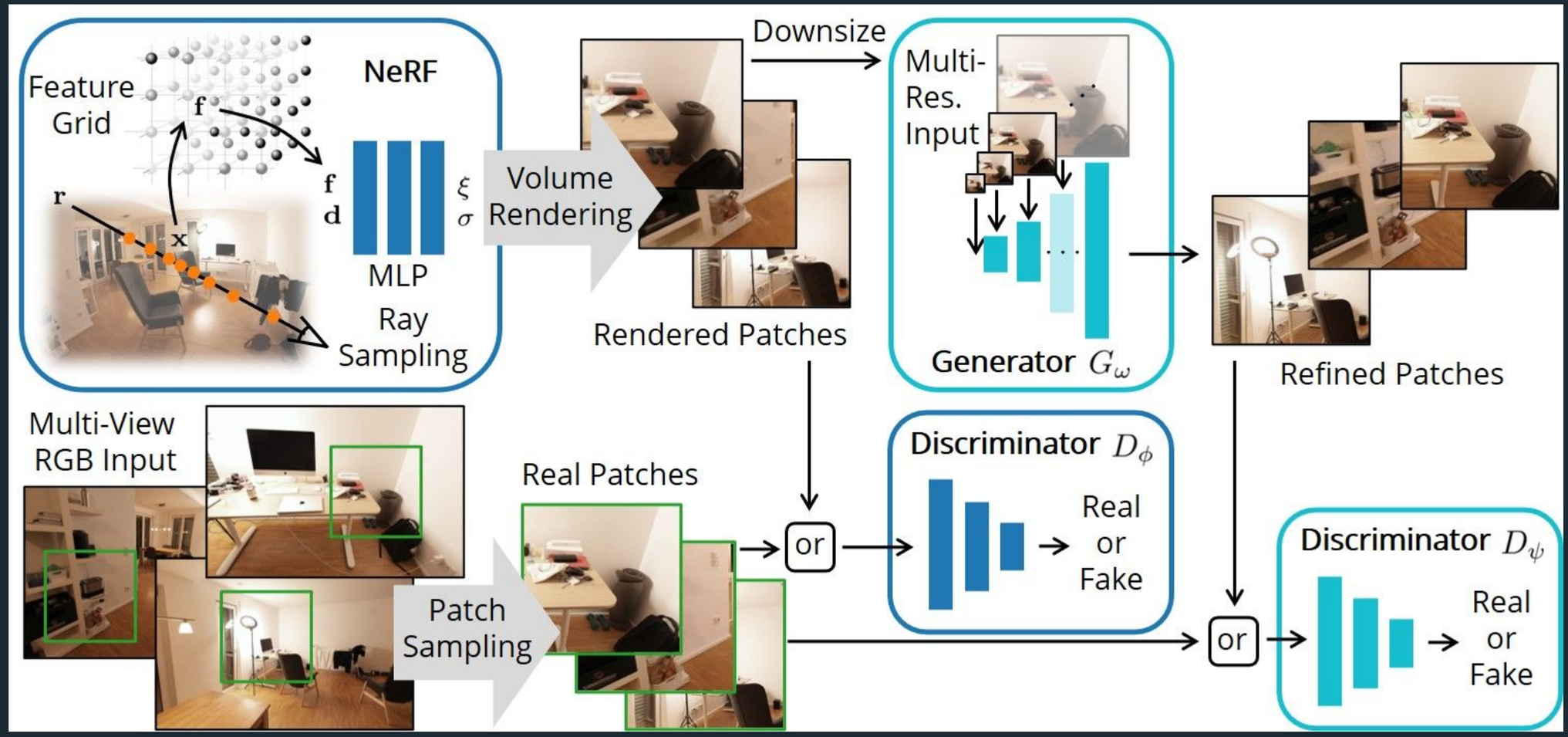
600+ images required for a great result (per scene)

GANeRF: Leveraging Discriminators to Optimize NeRFs

With imperfect input data, NeRF reconstructions will introduce artefacts like floaters. GANeRF introduces an adversarial formulation where gradient flow encourages 3D-consistency and additional constraints enabling more realistic novel view synthesis

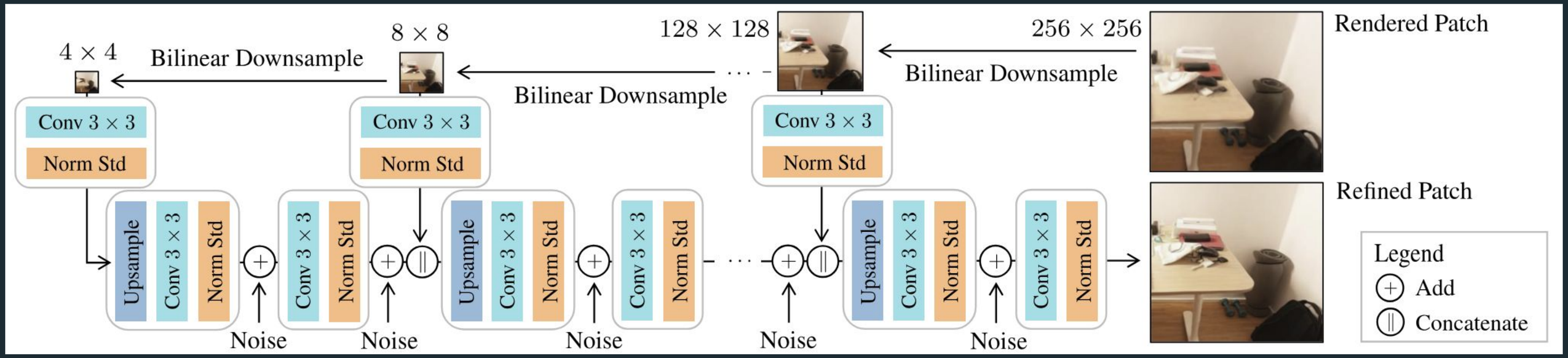
Rössle et al. GANeRF: Leveraging Discriminators to Optimize Neural Radiance Fields. Siggraph Asia, 2023

GANeRF: Architecture

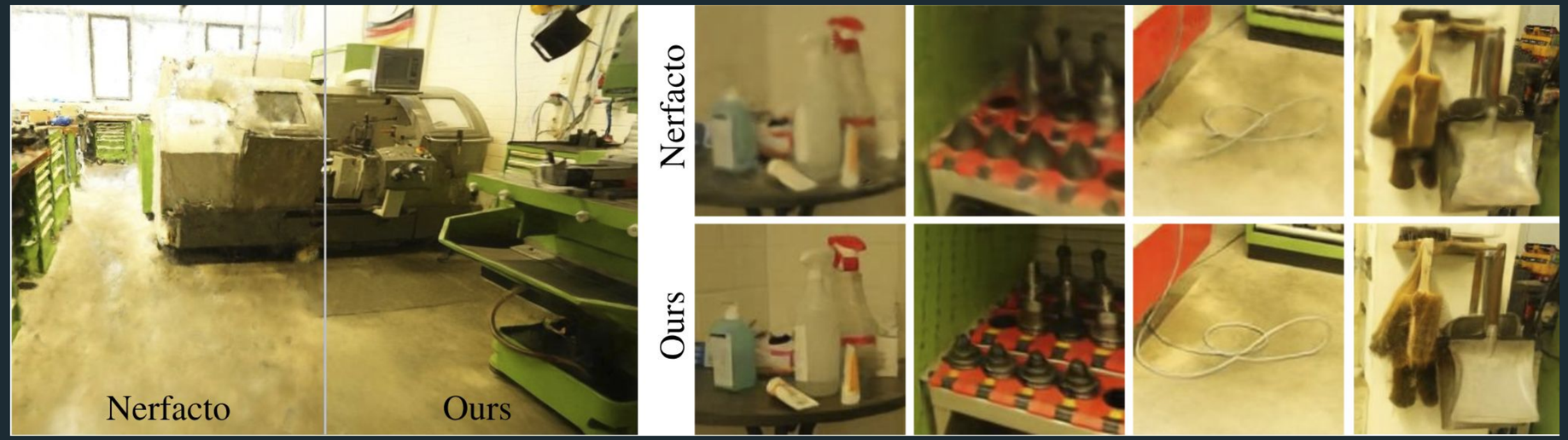


Rössle et al. GANeRF: Leveraging Discriminators to Optimize Neural Radiance Fields. ACM SIGGRAPH Asia 2023

GANeRF: Additional 2D Refinement



Full model qualitative comparisons vs. NeRFacto baseline



GANeRF: Experimental Evaluations

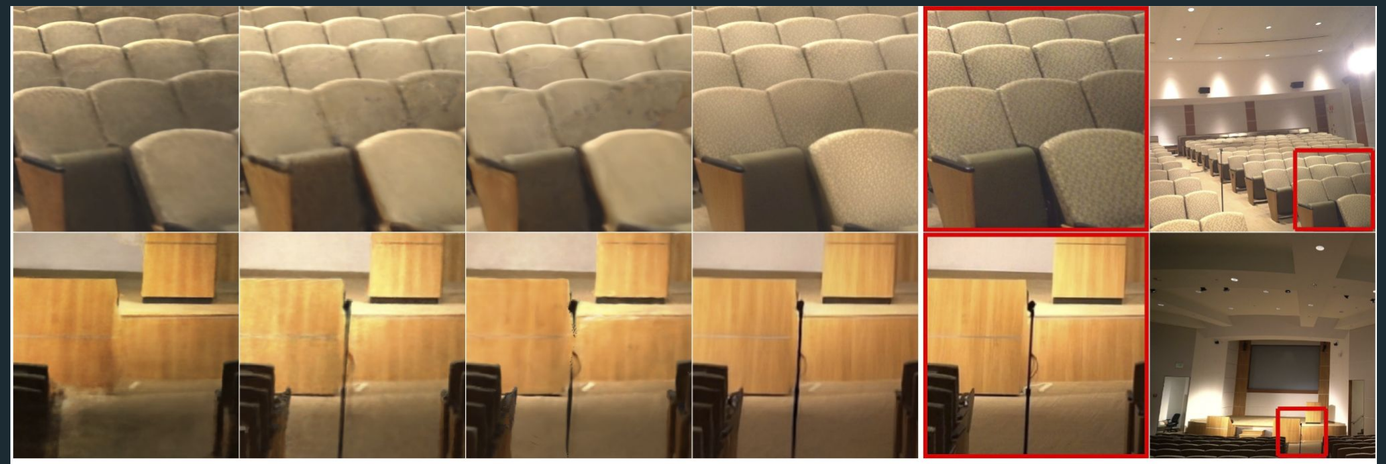
Scannet++ (5 scenes)

Method	PSNR↑	SSIM↑	LPIPS↓	KID ↓
Mip-NeRF 360 [Barron et al. 2022]	24.9	0.862	0.225	0.0241
Instant NGP [Müller et al. 2022]	25.3	0.844	0.269	0.0511
4K-NeRF [Wang et al. 2022]	22.7	0.807	0.254	0.0350
Nerfacto [Tancik et al. 2023]	25.6	0.848	0.245	0.0398
Nerfacto + extra capacity	25.9	0.854	0.228	0.0314
Nerfacto + pix2pix [Isola et al. 2017]	24.9	0.848	0.193	0.0162
Nerfacto + ControlNet [Zhang and Agrawala 2023]	23.1	0.827	0.174	**0.0097**
Ours w/o discriminator	25.8	0.857	0.177	0.0143
Ours w/o generator	25.9	0.860	0.198	0.0169
Ours	**26.1**	**0.864**	**0.161**	0.0113

Tanks and Temples Dataset

Method	PSNR↑	SSIM↑	LPIPS↓	KID ↓
Mip-NeRF 360 [Barron et al. 2022]	18.5	0.709	0.327	0.0277
Instant NGP [Müller et al. 2022]	19.3	0.700	0.369	0.0466
4K-NeRF [Wang et al. 2022]	19.4	0.656	0.356	0.0353
Nerfacto [Tancik et al. 2023]	19.5	0.716	0.329	0.0432
Nerfacto + extra capacity	19.6	0.733	0.291	0.0314
Nerfacto + pix2pix [Isola et al. 2017]	20.6	0.739	0.242	0.0115
Nerfacto + ControlNet [Zhang and Agrawala 2023]	19.6	0.706	0.213	0.0085
Ours w/o discriminator	20.6	0.745	0.192	0.0102
Ours w/o generator	19.9	0.739	0.251	0.0130
Ours	**20.9**	**0.776**	**0.169**	**0.0065**

Tanks and Temples: Qualitative Results



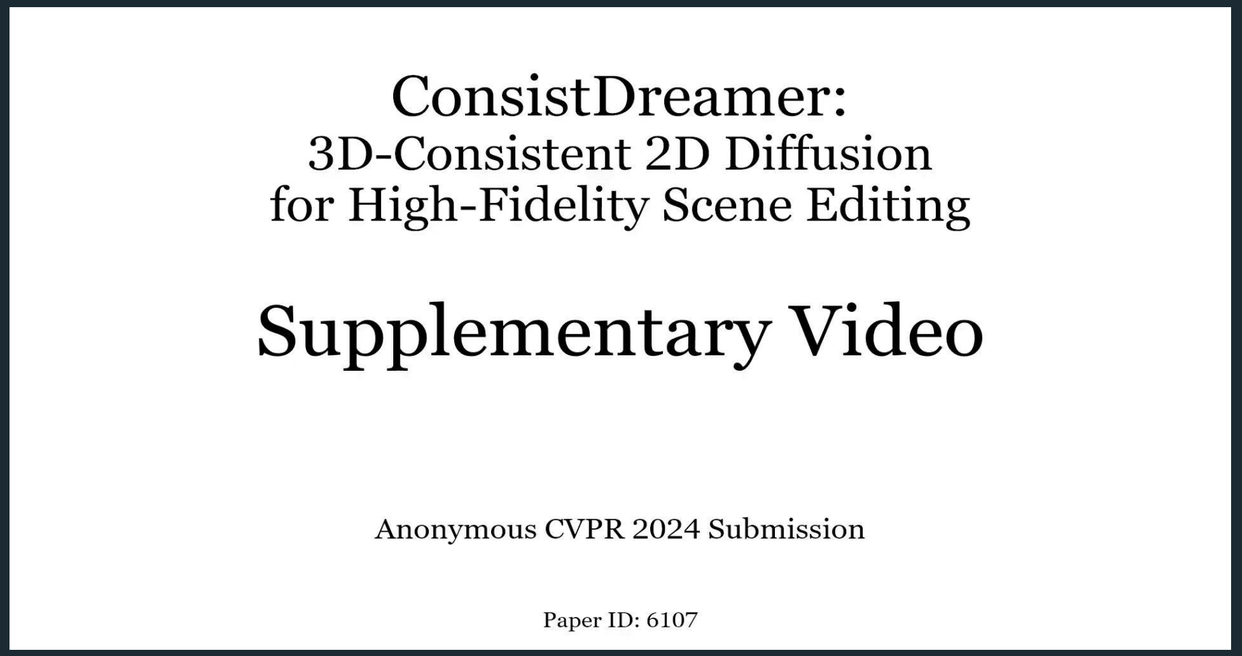
Mip-NeRF 360 | Instant NGP | Nerfacto | Ours | Ground truth

Scannet++ (reduced #images)

Method	# images	PSNR↑	SSIM↑	LPIPS↓	KID ↓
Nerfacto [Tancik et al. 2023]	800	24.2	0.844	0.247	0.0661
Ours		**24.7**	**0.870**	**0.169**	**0.0157**
Nerfacto [Tancik et al. 2023]	400	24.2	0.843	0.252	0.0766
Ours		**24.3**	**0.864**	**0.169**	**0.0176**
Nerfacto [Tancik et al. 2023]	200	23.5	0.825	0.274	0.0867
Ours		**23.9**	**0.862**	**0.182**	**0.0206**
Nerfacto [Tancik et al. 2023]	100	22.2	0.805	0.307	0.1069
Ours		**22.5**	**0.842**	**0.216**	**0.0289**
Nerfacto [Tancik et al. 2023]	50	20.2	0.771	0.347	0.1564
Ours		**20.5**	**0.788**	**0.282**	**0.0858**
Nerfacto [Tancik et al. 2023]	25	15.5	0.686	0.524	0.2432
Ours		**16.7**	**0.727**	**0.408**	**0.1738**

More GenAI gems at CVPR 2024!

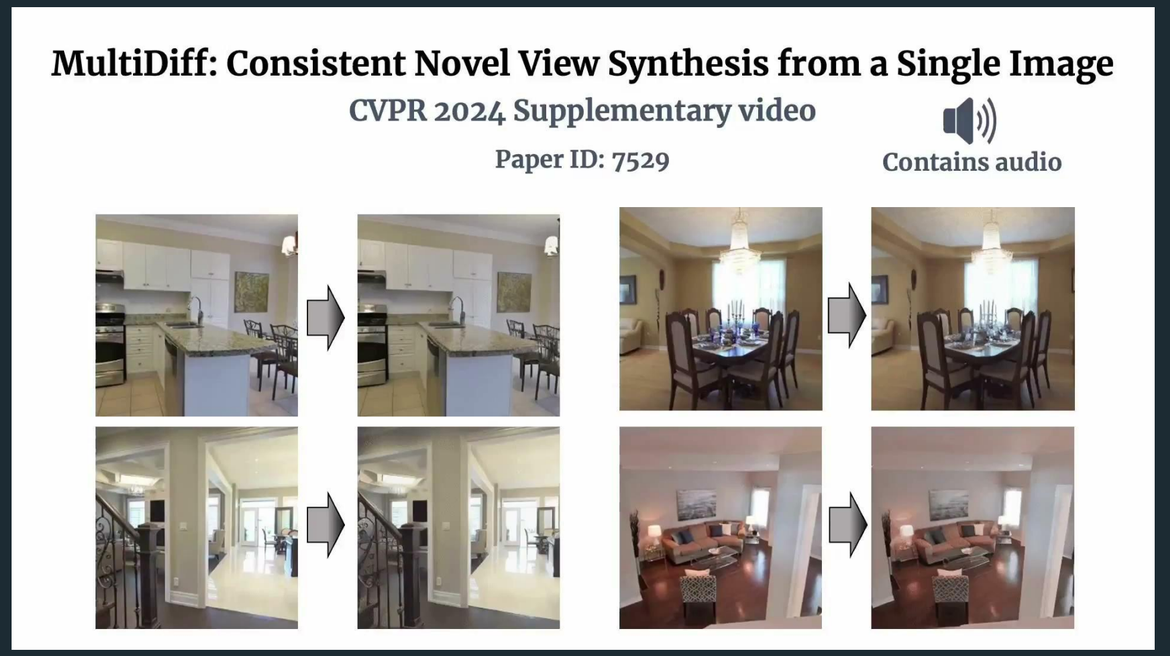
ConsistDreamer: Poster Session 5 & Exhibit Hall



High-fidelity stylization and editing from 3D-consistent 2D Diffusion models

Chen et al. ConsistDreamer: 3D-Consistent 2D Diffusion for High-Fidelity Scene Editing. CVPR 2024

MultiDiff: Poster Session 3 & Exhibit Hall



Single-image 3D synthesis!

Müller et al. MultiDiff: Consistent Novel View Synthesis from a Single Image. CVPR 2024

Summary

- We need to be able to generate AR/VR mapping content for the metaverse from consumer devices at scale
- ML-based, semantic, and photorealistic 3D representations are awesome - can we make them even more robust and efficient?
- Generative models are playing crucial roles for scene completion/stylization/synthesis

My team is hiring for multiple roles in Zurich!
Contact me to learn more!


Meta