EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision

Jiawei Yang[*,¶], Boris Ivanovic[¶], Or Litany[†,¶], Xinshuo Weng[¶], Seung Wook Kim[¶], Boyi Li[¶], Tong Che[¶], Danfei Xu[$,¶], Sanja Filder[§,¶], Marco Pavone[‡,¶], Yue Wang[*,¶]

[*]University of Southern California, [†]Technion, [$]Georgia Institute of Technology, [§]University of Toronto,

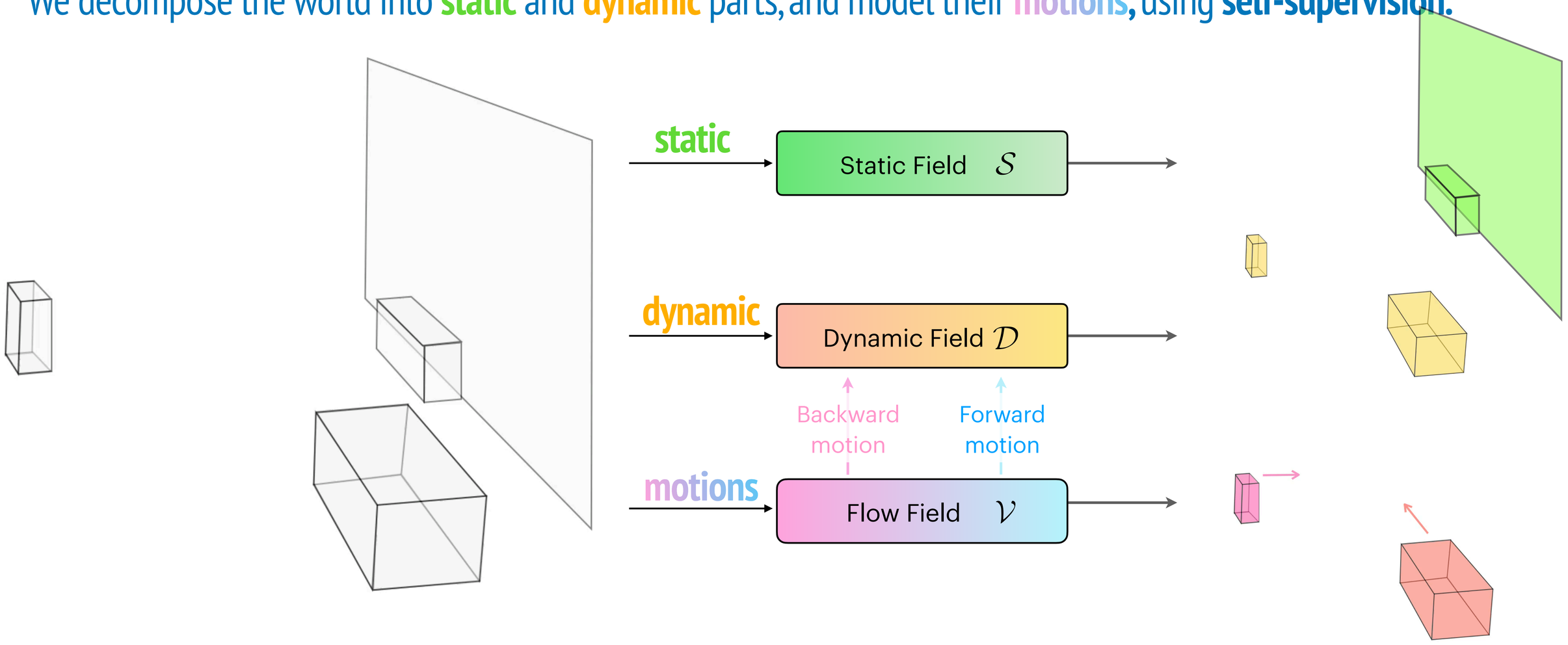
[‡]Stanford University, [¶]Nvidia Research

EmerNeRF

- Scalable: Everything is self-supervised. No human annotation is needed.
- Realistic: State-of-the-art photorealistic reconstruction. Perfect for simulations.
- All-in-one-pipeline:
  - Differentiating static/dynamic objects
  - Estimating 3D scene flows
  - Auto-labeling semantic occupancies

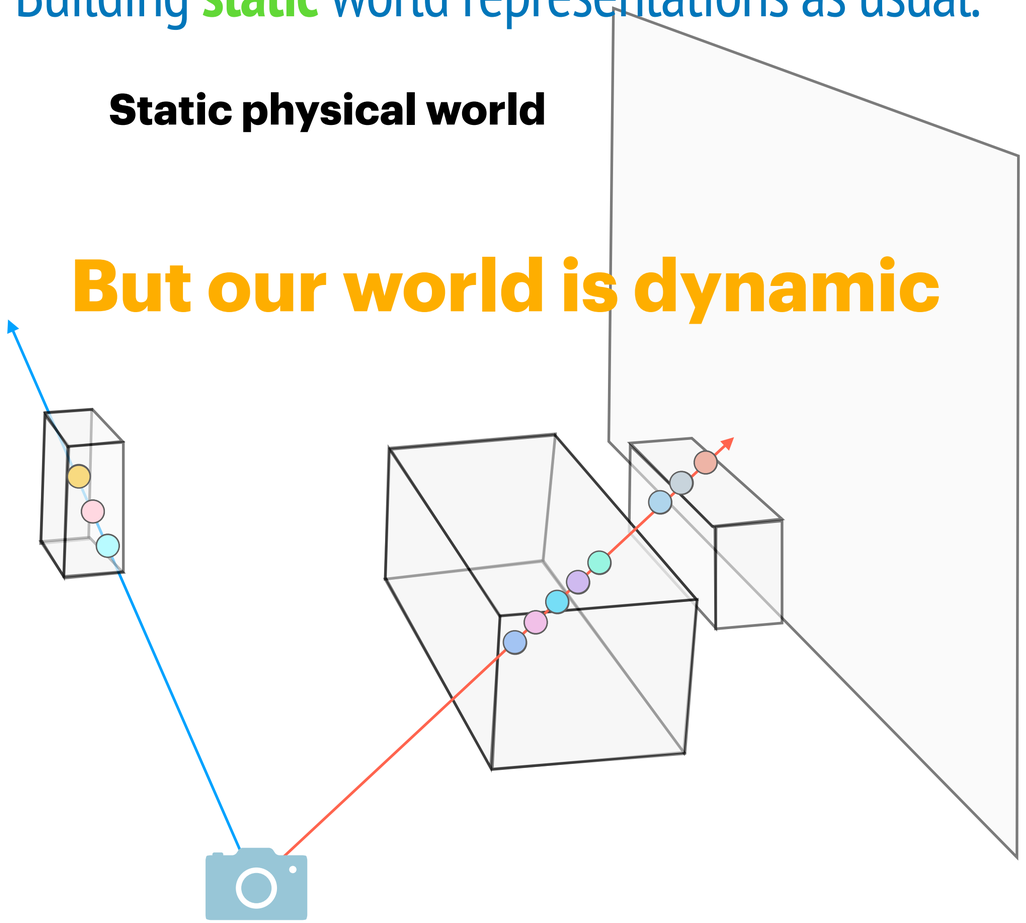
EmerNeRF Overview

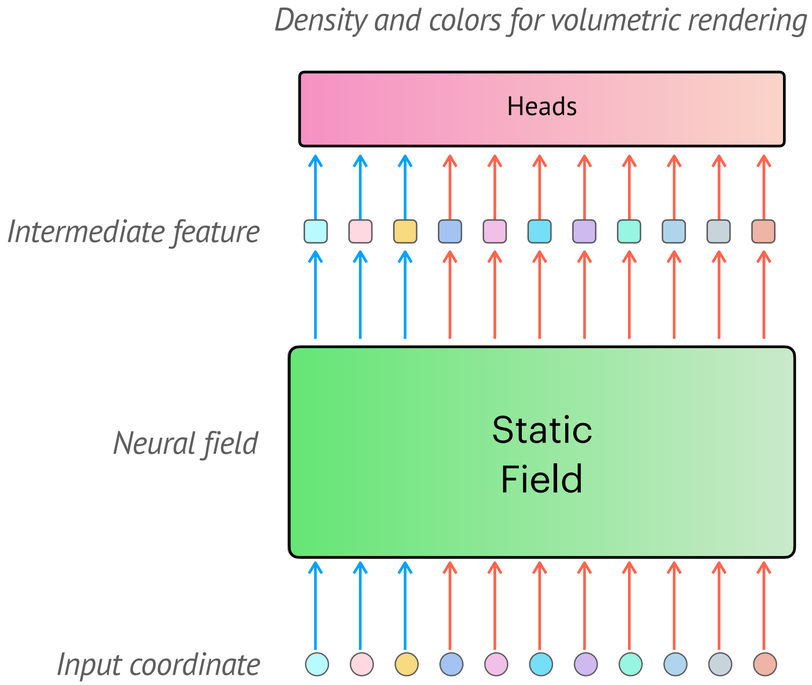
We decompose the world into **static** and **dynamic** parts, and model their **motions**, using **self-supervision.**



EmerNeRF Overview

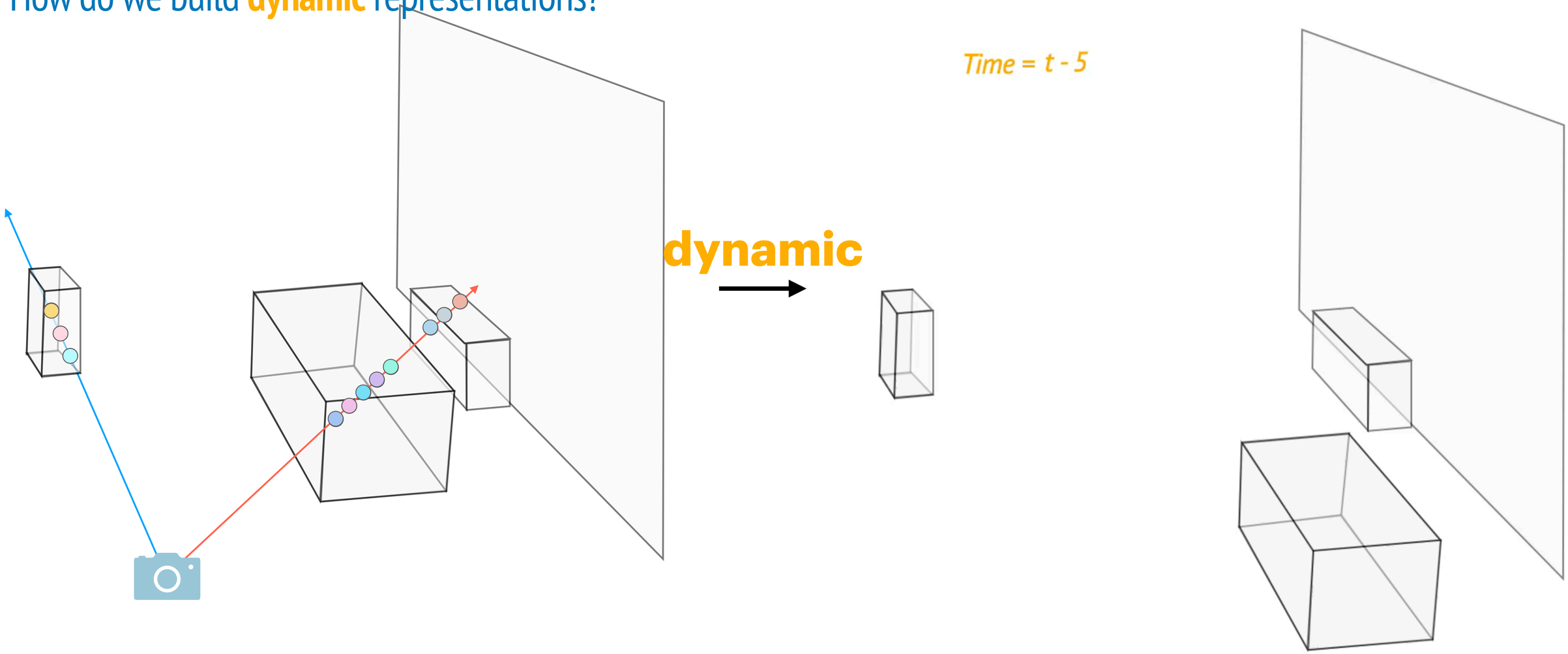
Building **static** world representations as usual.





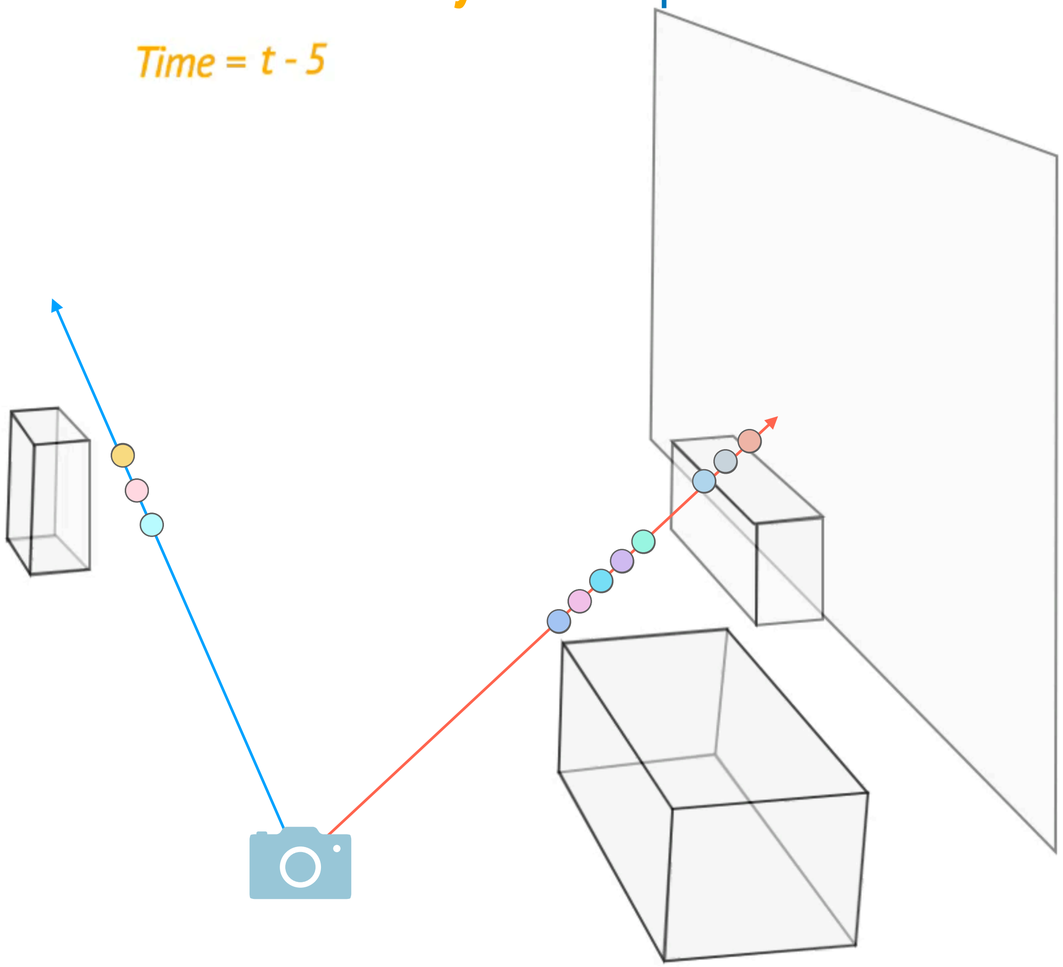
EmerNeRF Overview

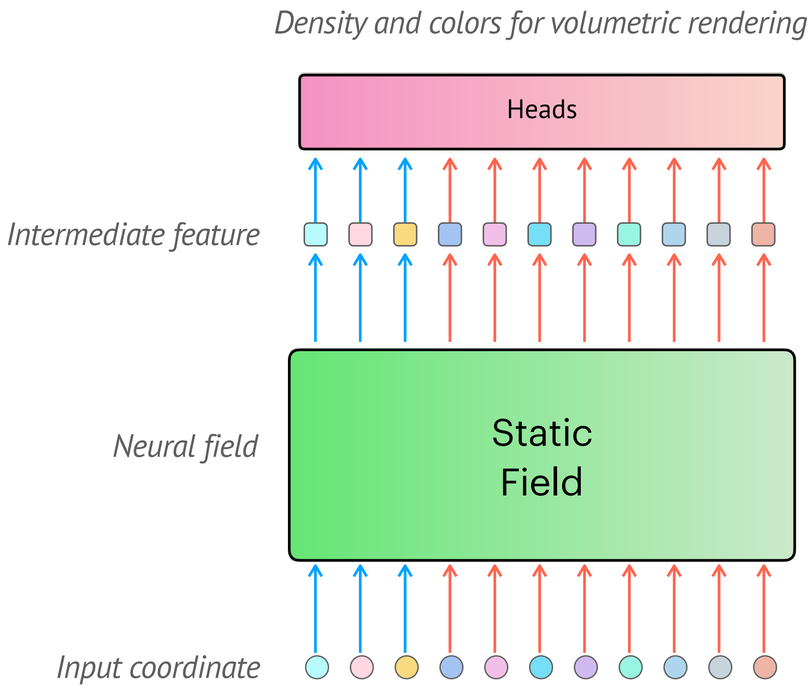
How do we build **dynamic** representations?



EmerNeRF Overview

How do we build **dynamic** representations?



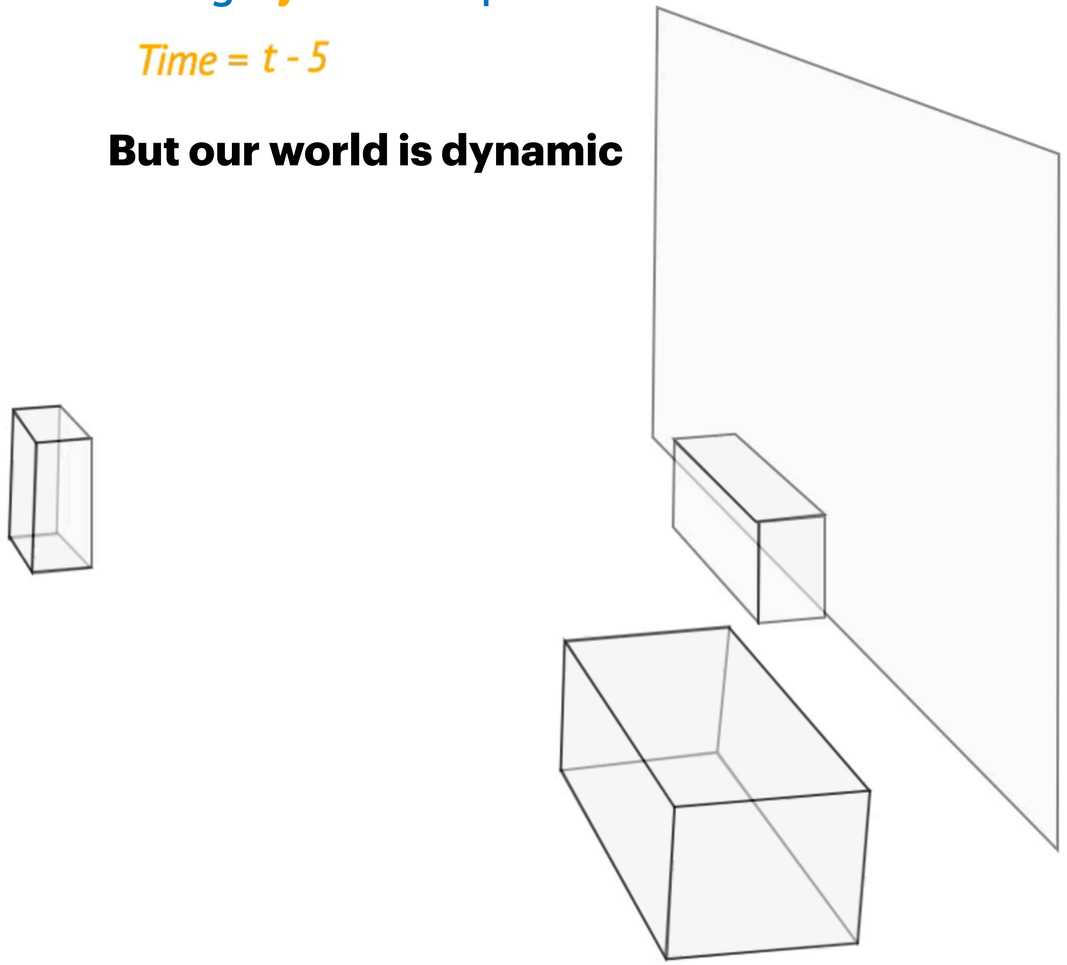


EmerNeRF Overview

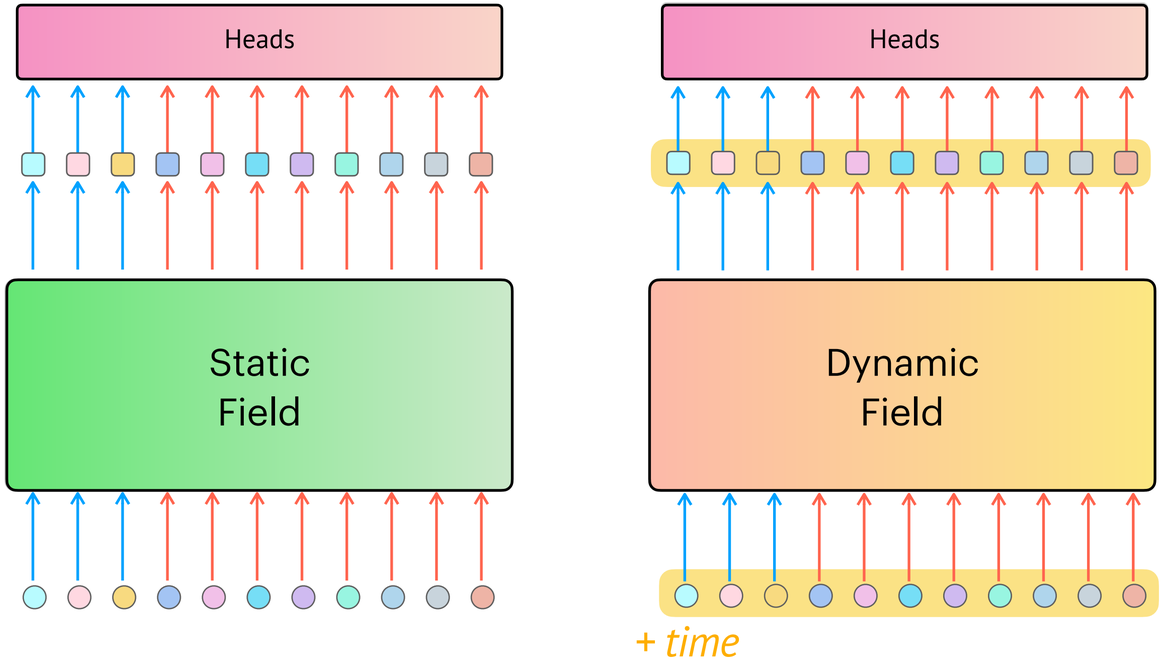
Building dynamic representations

Time = t - 5

But our world is dynamic



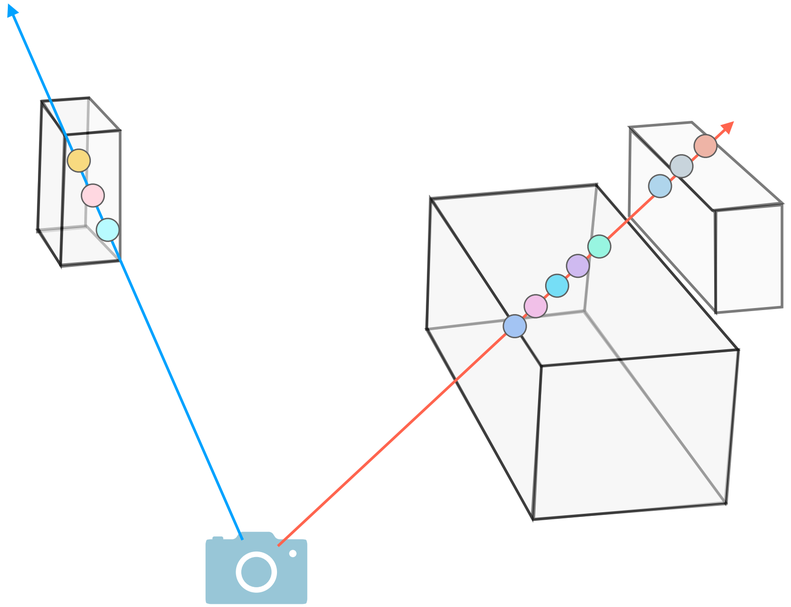
*Building **Spatial-temporal** scene representations*

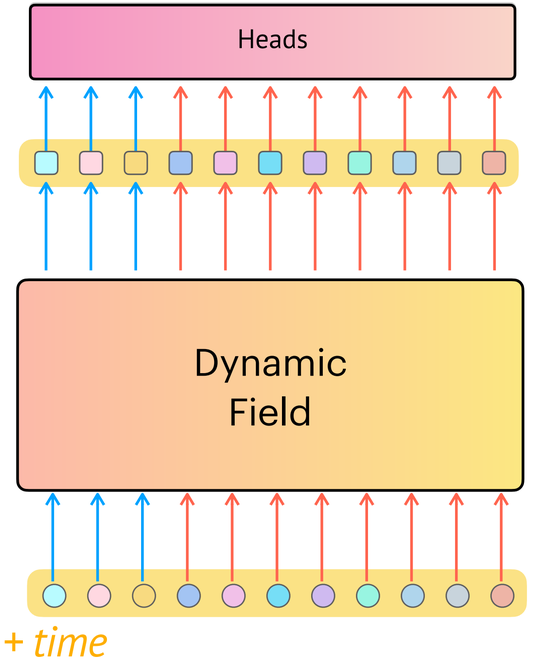


Add time-dimension to the dynamic field

EmerNeRF Overview

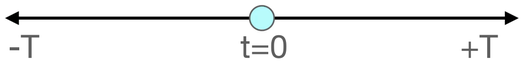
Building motion for temporal correspondence



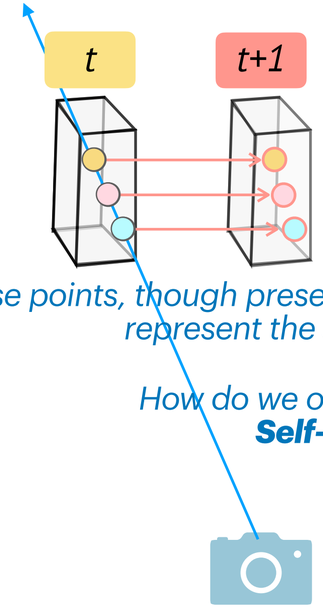


EmerNeRF Overview

Building motion for temporal correspondence

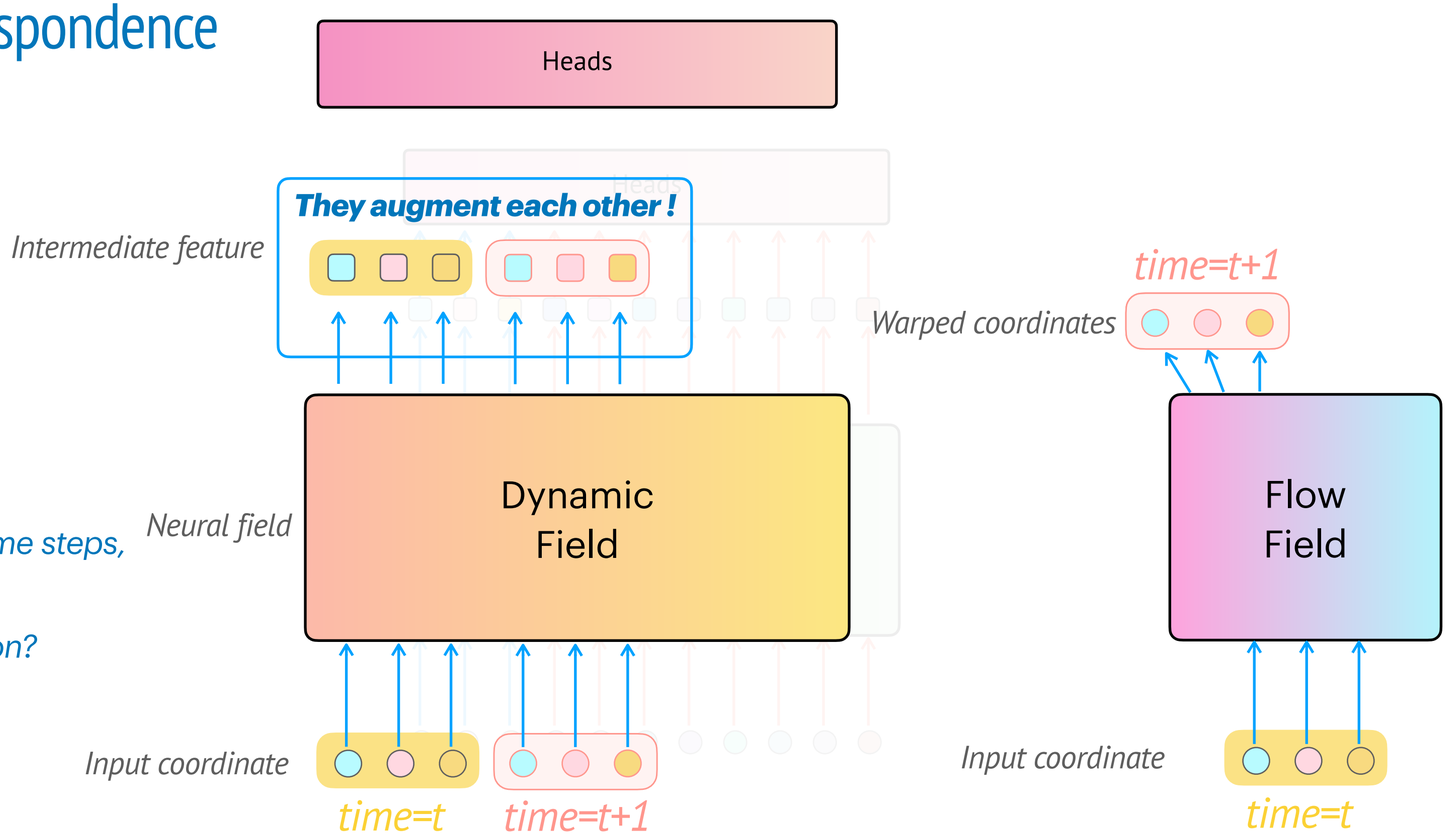


Temporal Equivariance



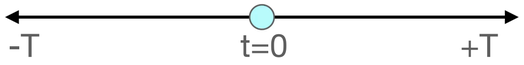
These points, though presented at different time steps, represent the same particles

How do we obtain transformation?
Self-supervision!

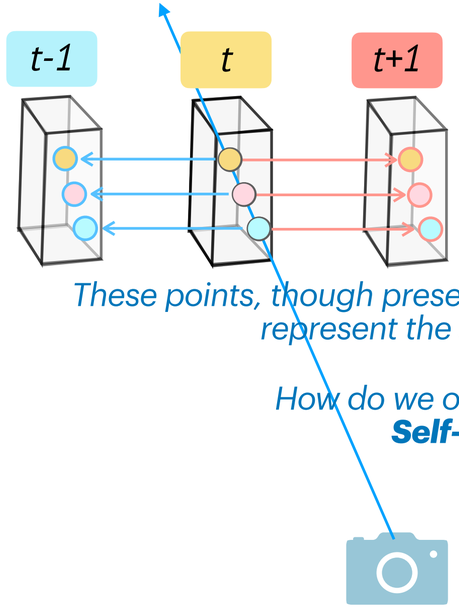


EmerNeRF Overview

Building motion for temporal correspondence

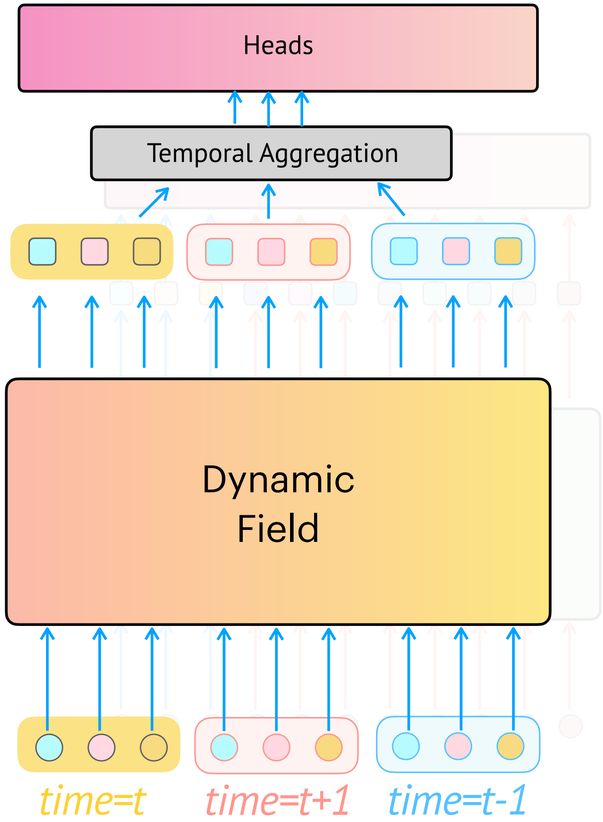


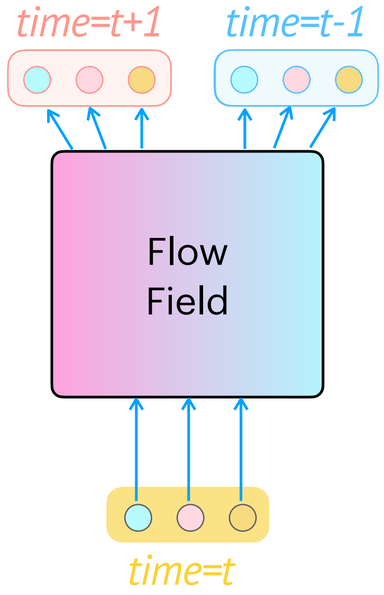
Temporal Equivariance



These points, though presented at different time steps, represent the same particles

How do we obtain transformation?
Self-supervision!





EmerNeRF Capabilities

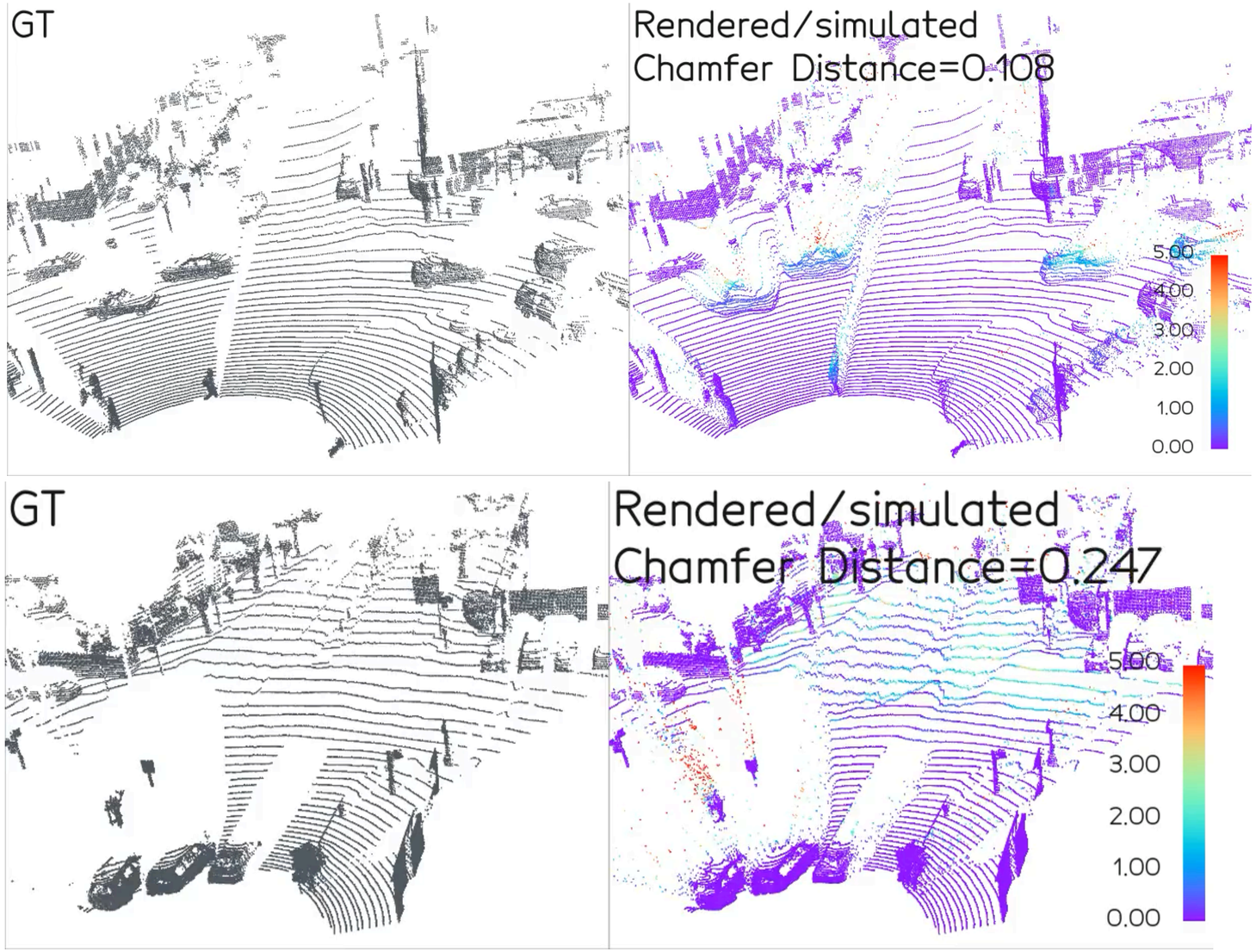
- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction

Original Camera Log

Rendered Camera Log

EmerNeRF Capabilities

- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction



Why neural field scene representation?

EmerNeRF Capabilities

- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction

Original Camera Log

Static RGB

Dynamic RGB

EmerNeRF Capabilities

- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction

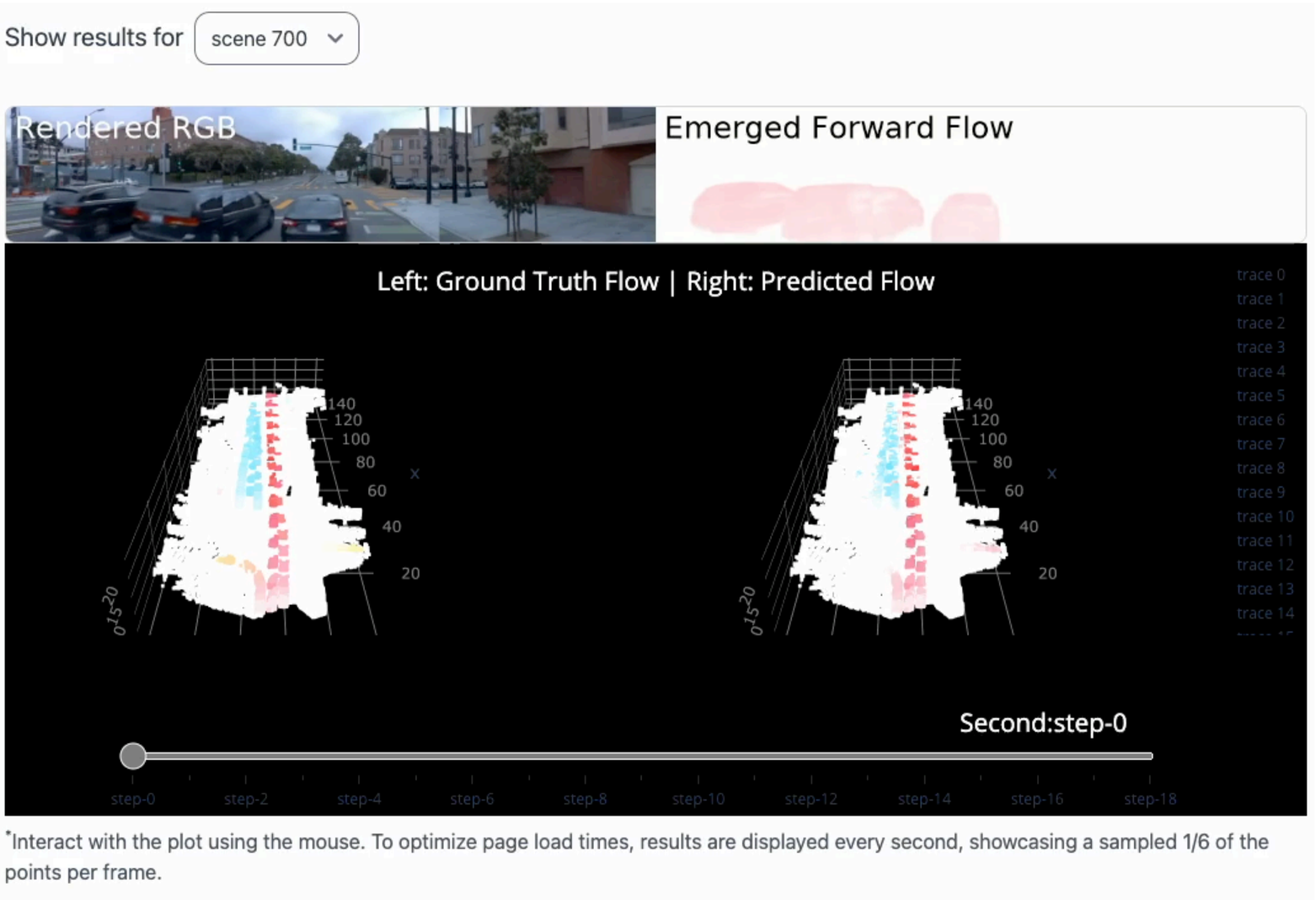
Original Camera Log

Scene Flow estimation.

Why neural field scene representation?

EmerNeRF Capabilities

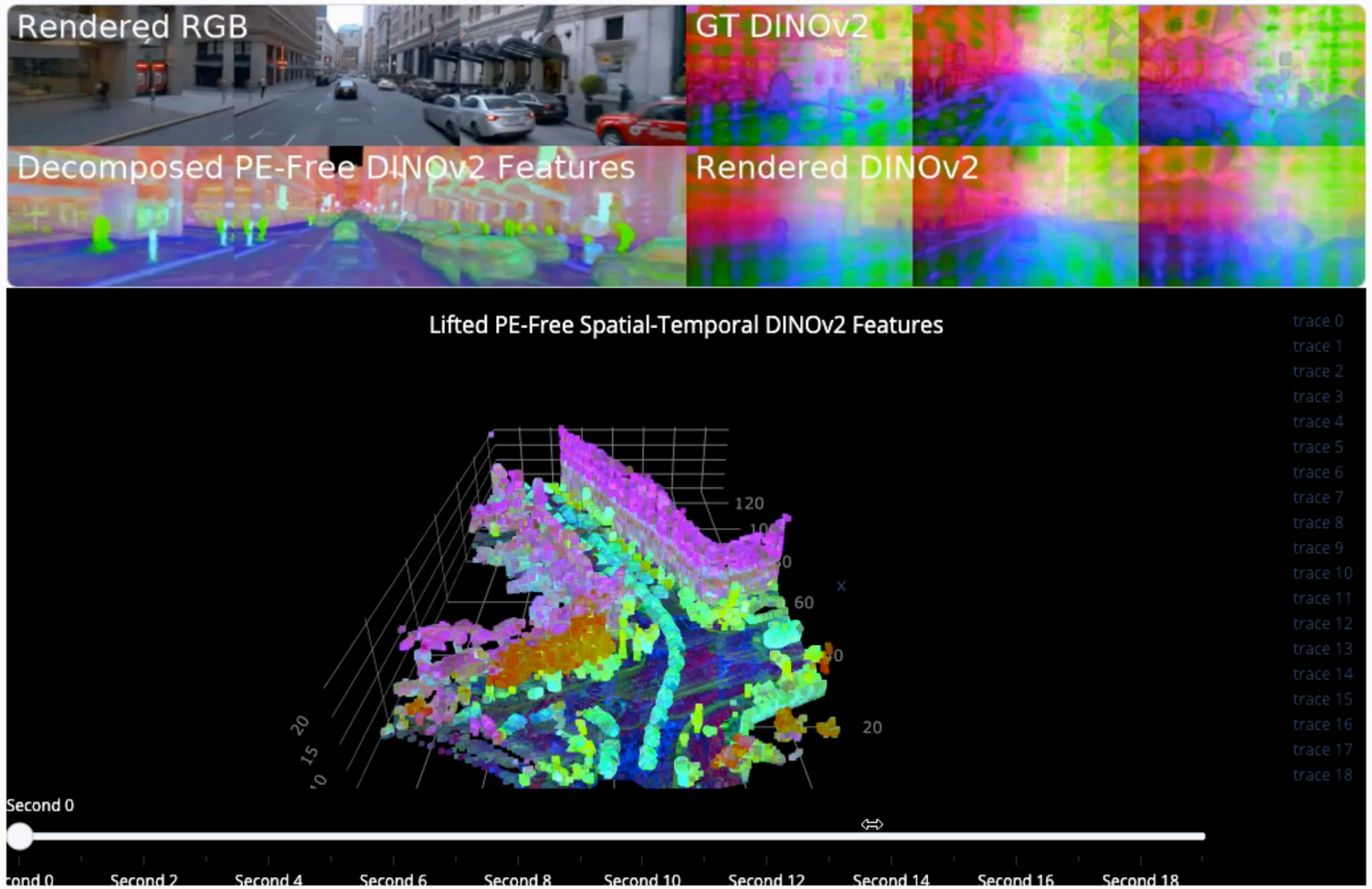
- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction



*Interact with the plot using the mouse. To optimize page load times, results are displayed every second, showcasing a sampled 1/6 of the points per frame.

EmerNeRF Capabilities

- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction

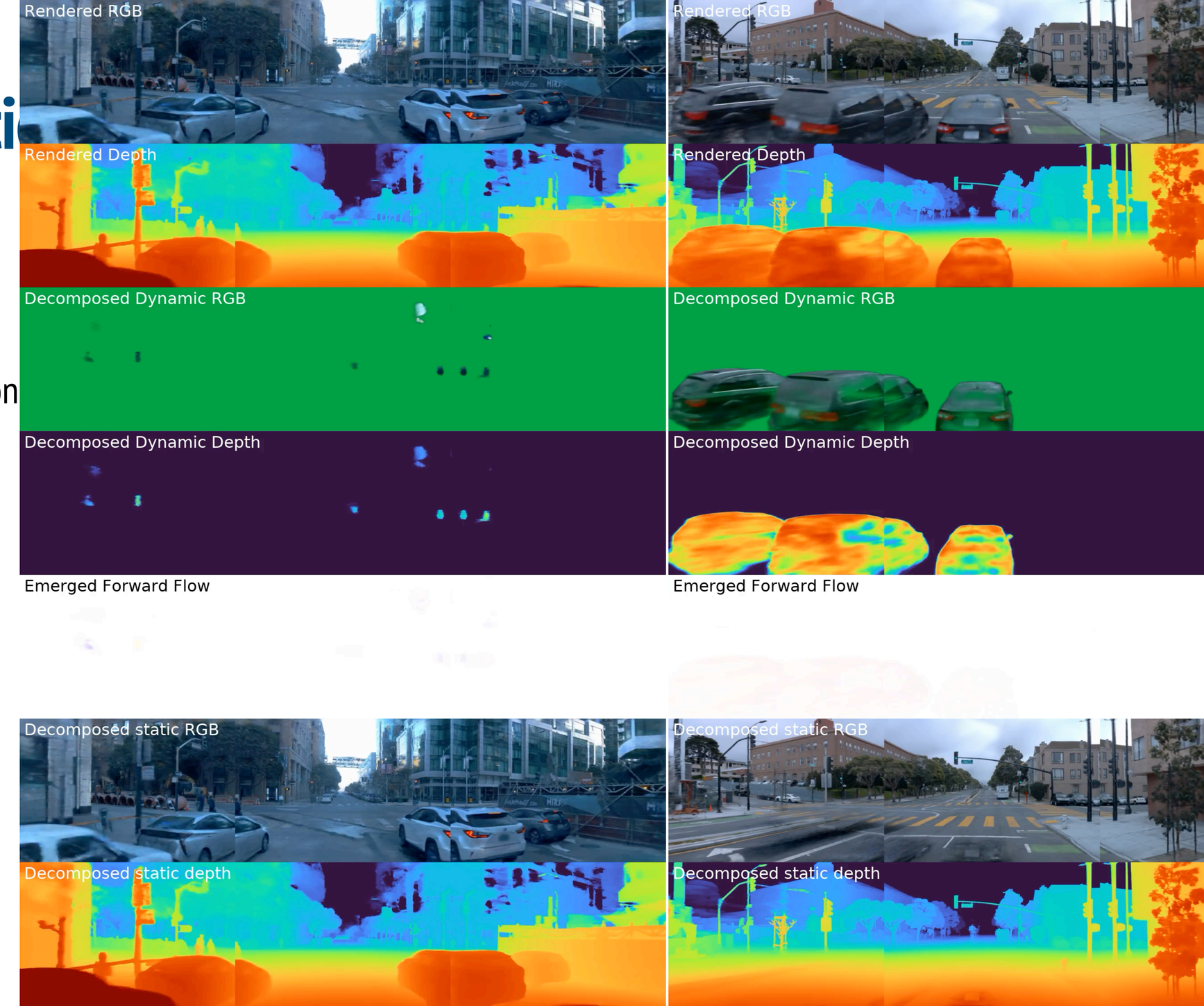


*Interact with the plot using the mouse. To optimize page load times, results are displayed every second. Note: 2D and 3D features are visualized distinctly and may have different color representations. Voxel size is 0.15m.

Why neural field scene representation?

EmerNeRF Capabiliti

- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction



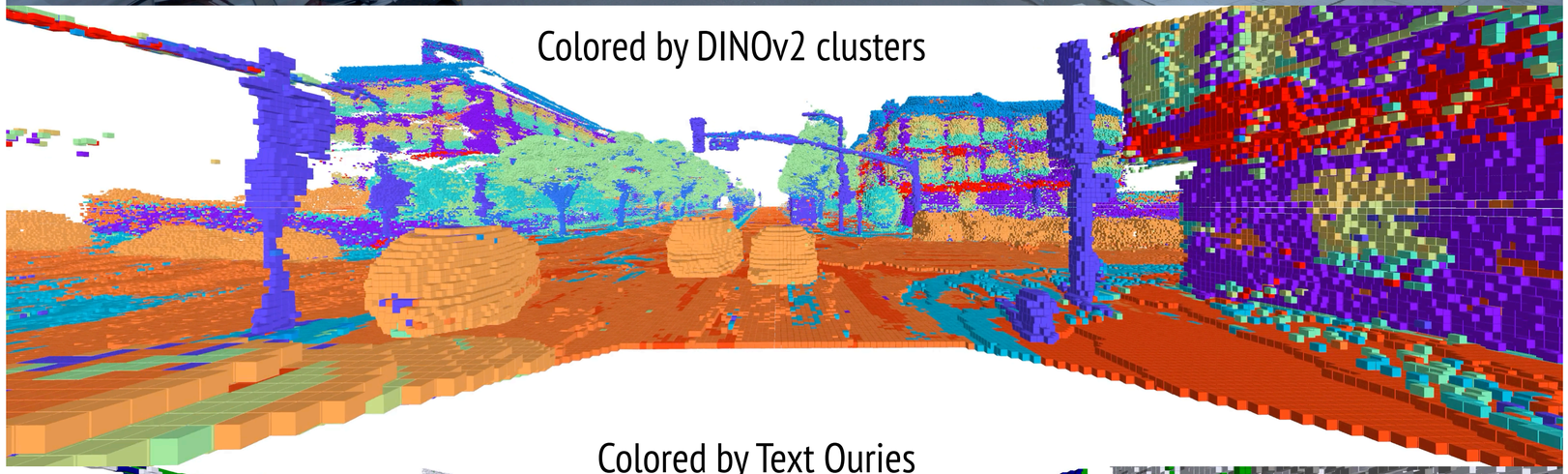
Why neural field scene representation?

EmerNeRF Capabilities

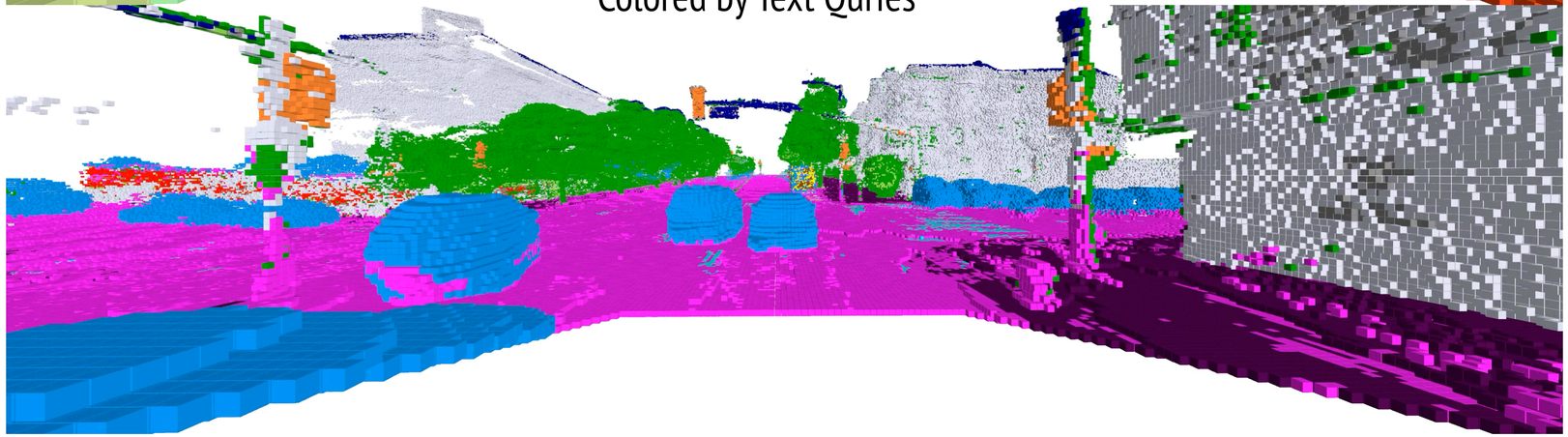
- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction

Camera Images

Colored by DINOv2 clusters



Colored by Text Quries



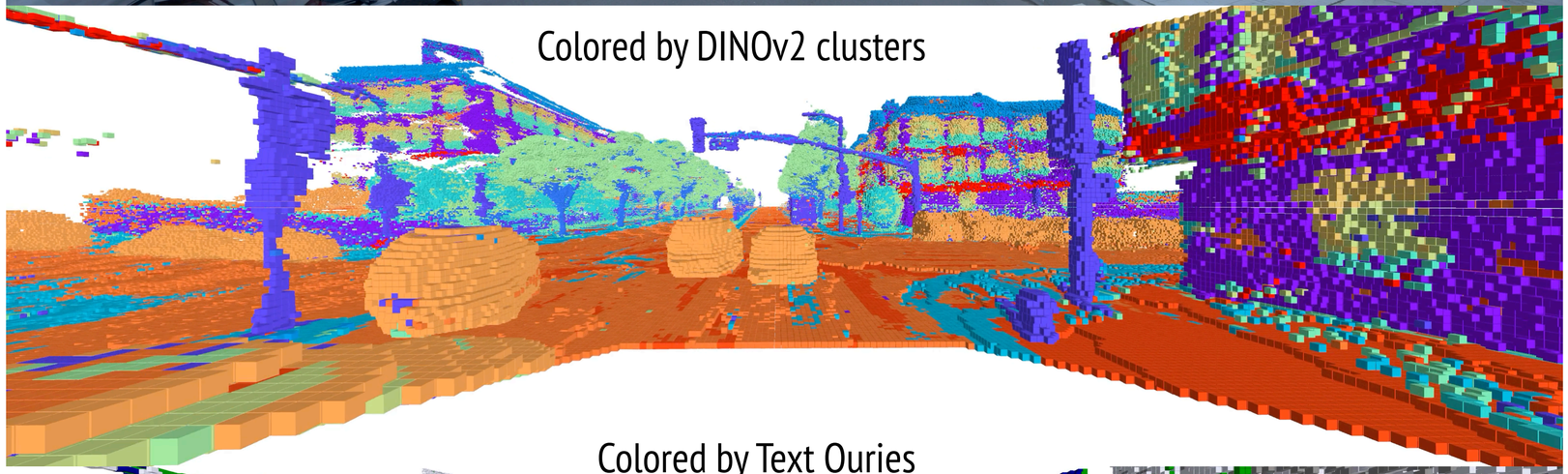
Why neural field scene representation?

EmerNeRF Capabilities

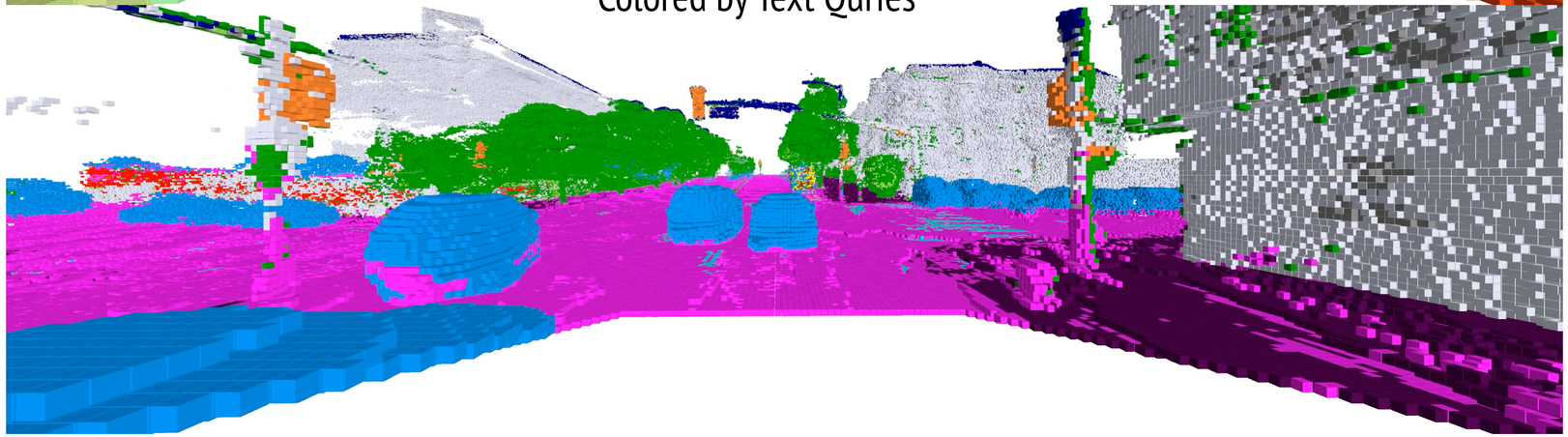
- With self-supervision, it can do
  - Log Replay
  - Static / Dynamic Decomposition
  - Motion Estimation
  - Semantics Understanding
  - Novel View Synthesis
  - Occupancy Reconstruction

Camera Images

Colored by DINOv2 clusters



Colored by Text Quries



EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision

- Self-supervised learning to reconstruct dynamic scenarios **at scale**.
- Through **self-supervision**, EmerNeRF learns:
  - **Static-dynamic** scene decomposition
  - Highly accurate 3D **scene flows**
  - **Artifacts-free** foundational models' features
- Please refer to our project page and **open-sourced** code for more details:
  - Project page: https://emernerf.github.io/
  - Code Page: https://github.com/NVlabs/EmerNeRF

Project Page

Thanks for watching!