DistillNeRF: Perceiving 3D Scenes from Single-Glance Images by Distilling Neural Fields and Foundation Model Features

NeurIPS 2024

Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic,

Steven Waslander, Yue Wang, Sanja Fidler, Marco Pavone, Peter Karkus

NVIDIA

UNIVERSITY OF TORONTO

USC University of Southern California

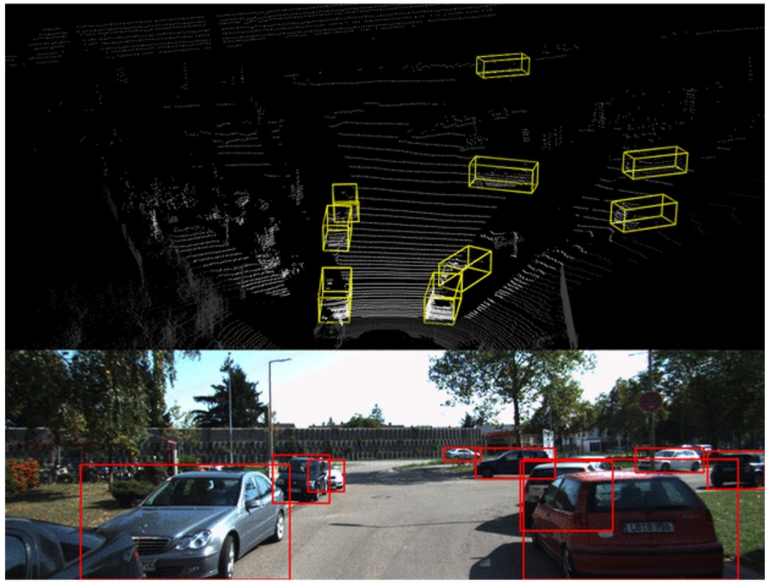
NUS National University of Singapore

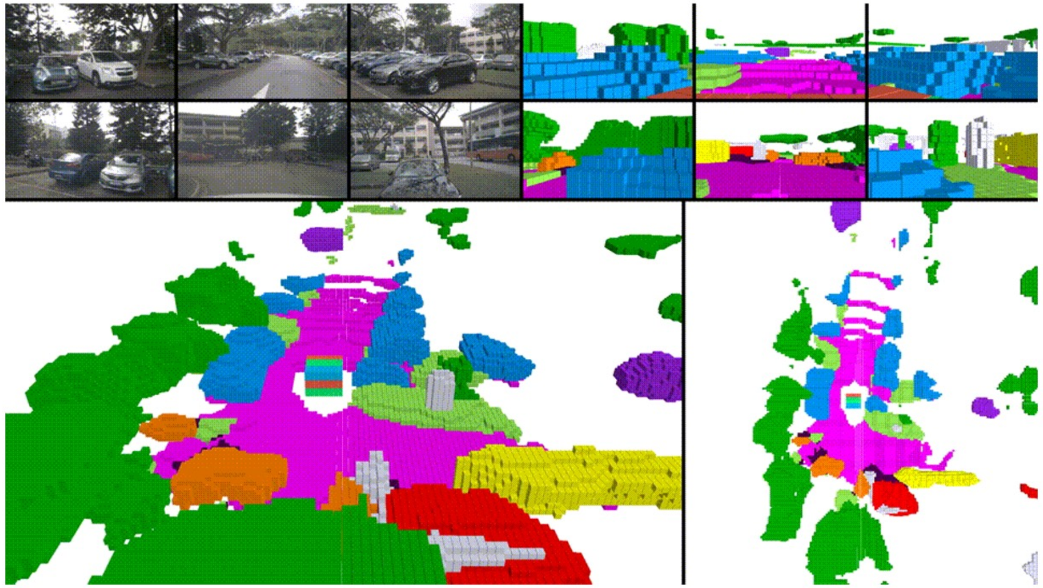
Stanford University

Perceiving the 3D World – Traditional Perception Tasks

- Various perception tasks have been proposed to perceive 3D world with 2D observations
  - Detection, tracking, segmentation, semantic occupancy…

- Hard to scale with a large of data
  - Industries have collected tons of data
  - But, the data annotation is painful: imagine you need to annotate each voxel for the scene…

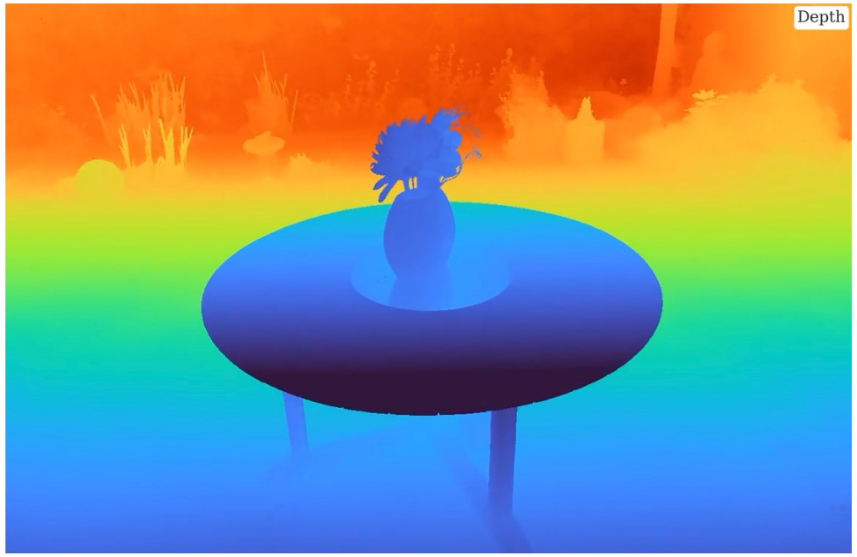


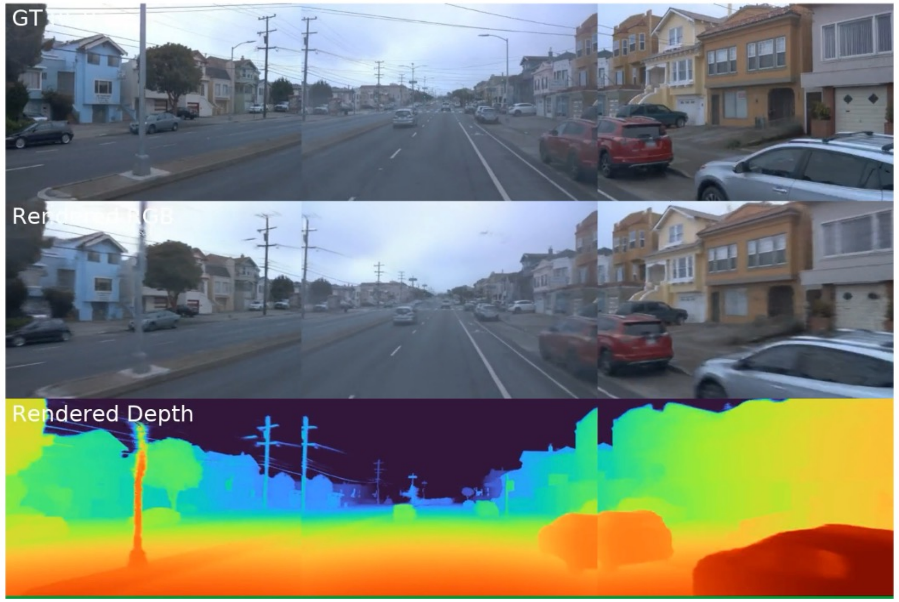


- Liang, M., Yang, B., Chen, Y., Hu, R. and Urtasun, R., 2019. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7345-7353).
- Huang, Y., Zheng, W., Zhang, Y., Zhou, J. and Lu, J., 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9223-9232).

Neural Radiance Fields (NeRF)

- Self-supervised learning: reconstruct the scene with RGB and optional LiDAR as inputs and supervision
- Very popular and fast-moving topic
- From simple objects, to unbounded scene, to autonomous driving scenes recently

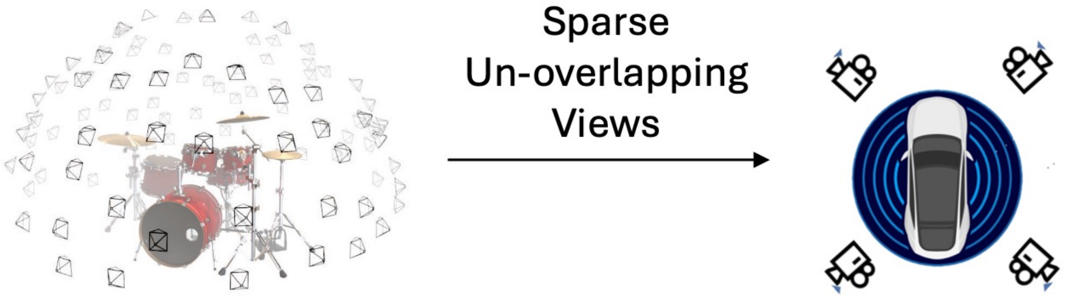




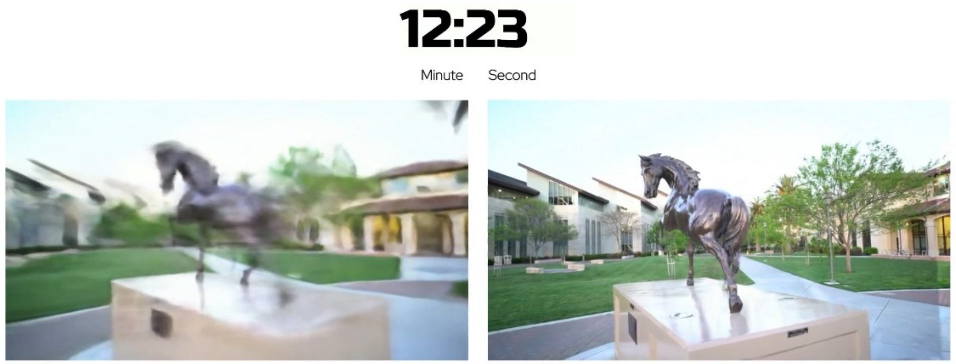
- Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P. and Hedman, P., 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5470-5479).
- Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M. and Wang, Y., 2023. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. arXiv preprint arXiv:2311.02077.

Neural Radiance Fields (NeRF)

- Sounds all good! What could be wrong?
  - Requires a large number of overlapping images
  - Need training for each scene at test time, which take hours/minutes/seconds
  - Only focus on view-synthesis tasks, lack rich semantics in learned 3D representations
- Issues for autonomous driving
  - Sparse cameras with limited overlaps on vehicle, usually 4/6 cameras
  - Need to run in real time for online driving - usually only ~0.1s latency is allowed for on-vehicle computation
  - Need models with capabilities in downstream tasks



Sparse Un-overlapping Views



NoPe-NeRF

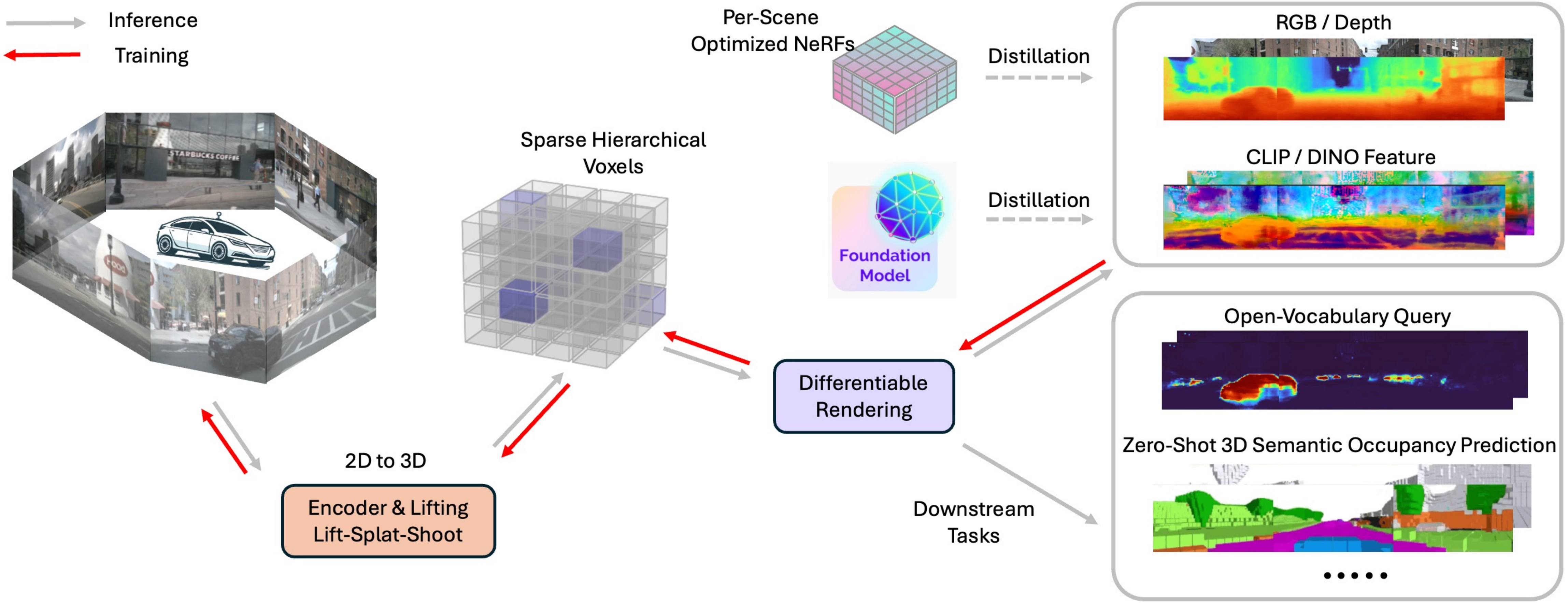
Ours (9s ✔)

Can we bring NeRF to handle sparse non-overlapping views, be online and generalizable to new scenes, and support downstream tasks?

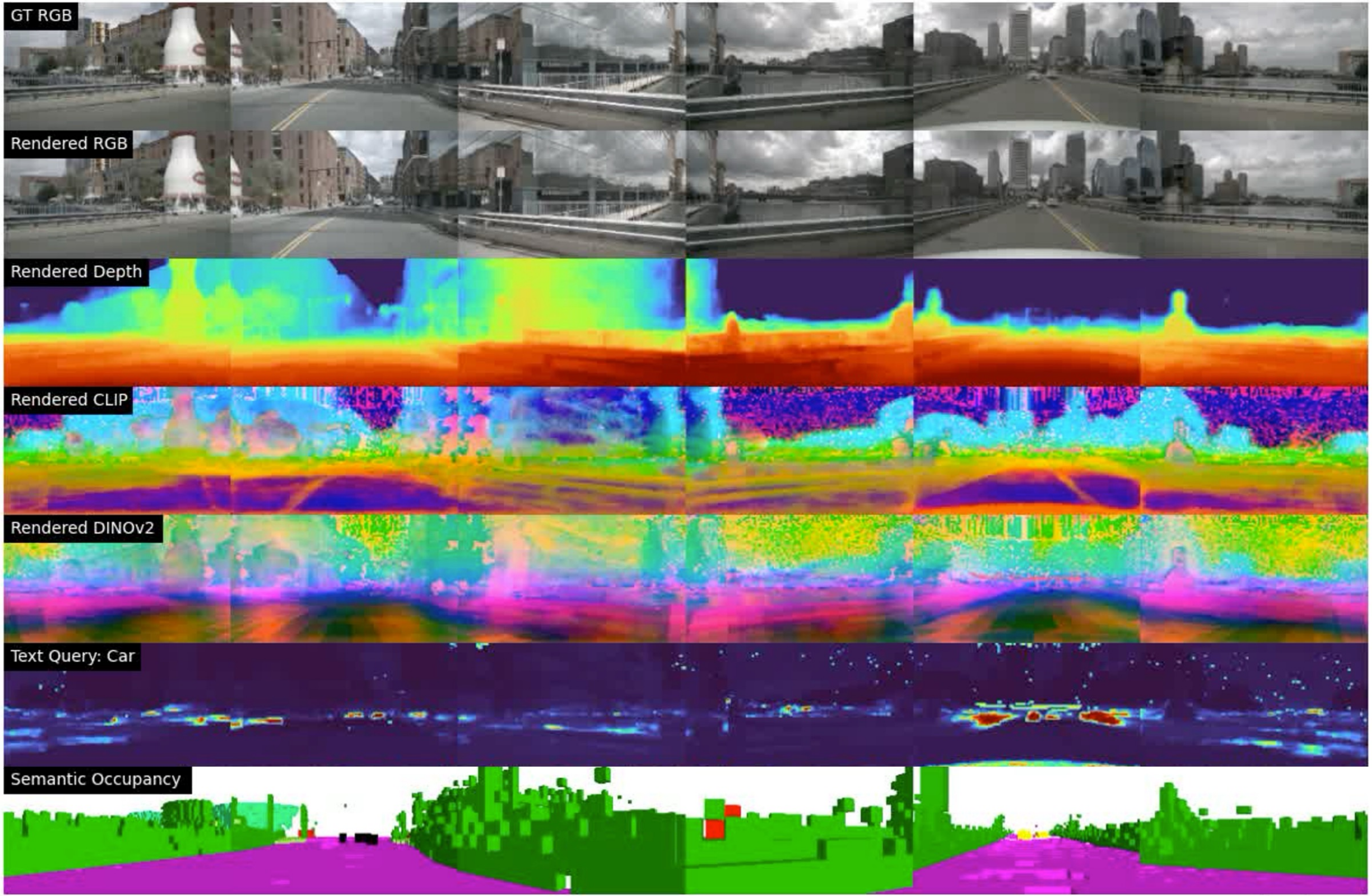
- Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G. and Wang, Z., 2024. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*.

Our Approach - DistillNeRF

- An online model that lifts 2D features into 3D, and can render RGB/depths, without test-time per-scene training
- Distill a bunch of per-scene optimized NeRFs into one online model, for enhanced 3D geometry
- Distill foundation model features into the online model, for enriched semantics
- Support downstream tasks: rendering, open-vocabulary query, zero-shot semantic occupancy prediction



Capabilities: Rendering without Test-Time Per-Scene Optimization, Enable Zero-Shot Downstream Tasks



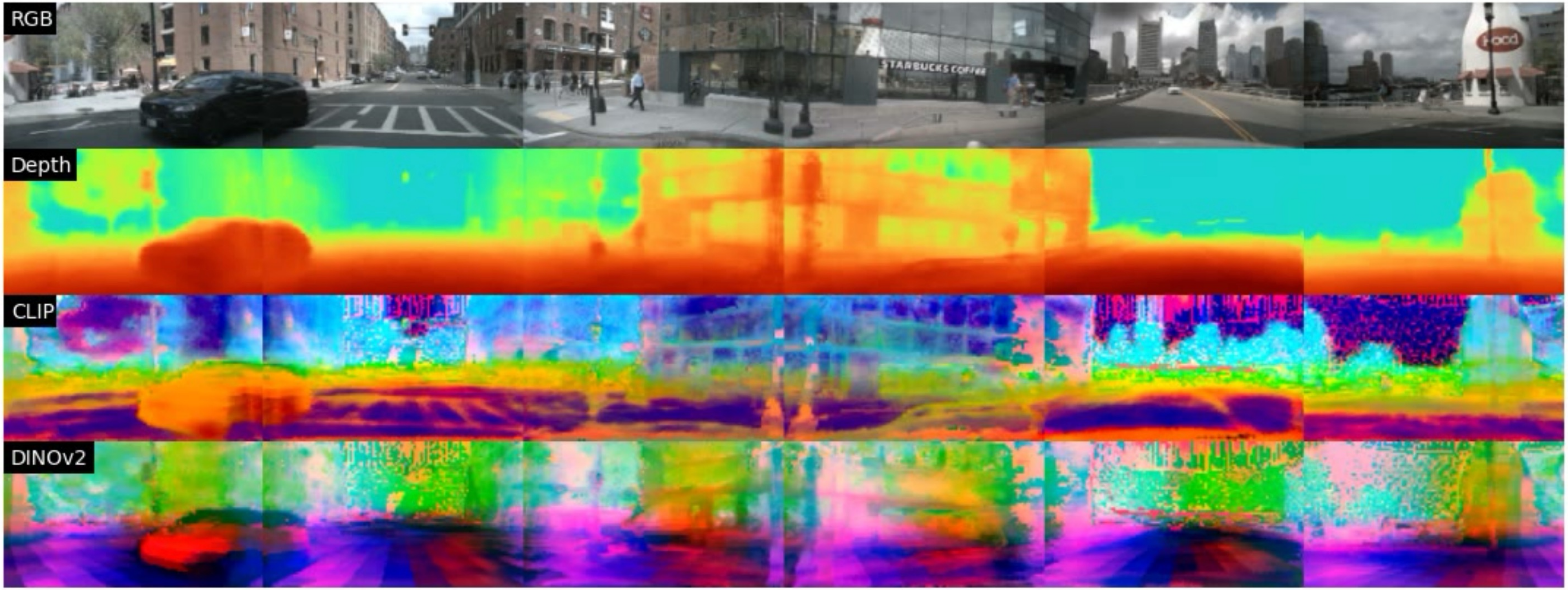
Single-Frame Input

Scene Reconstruction

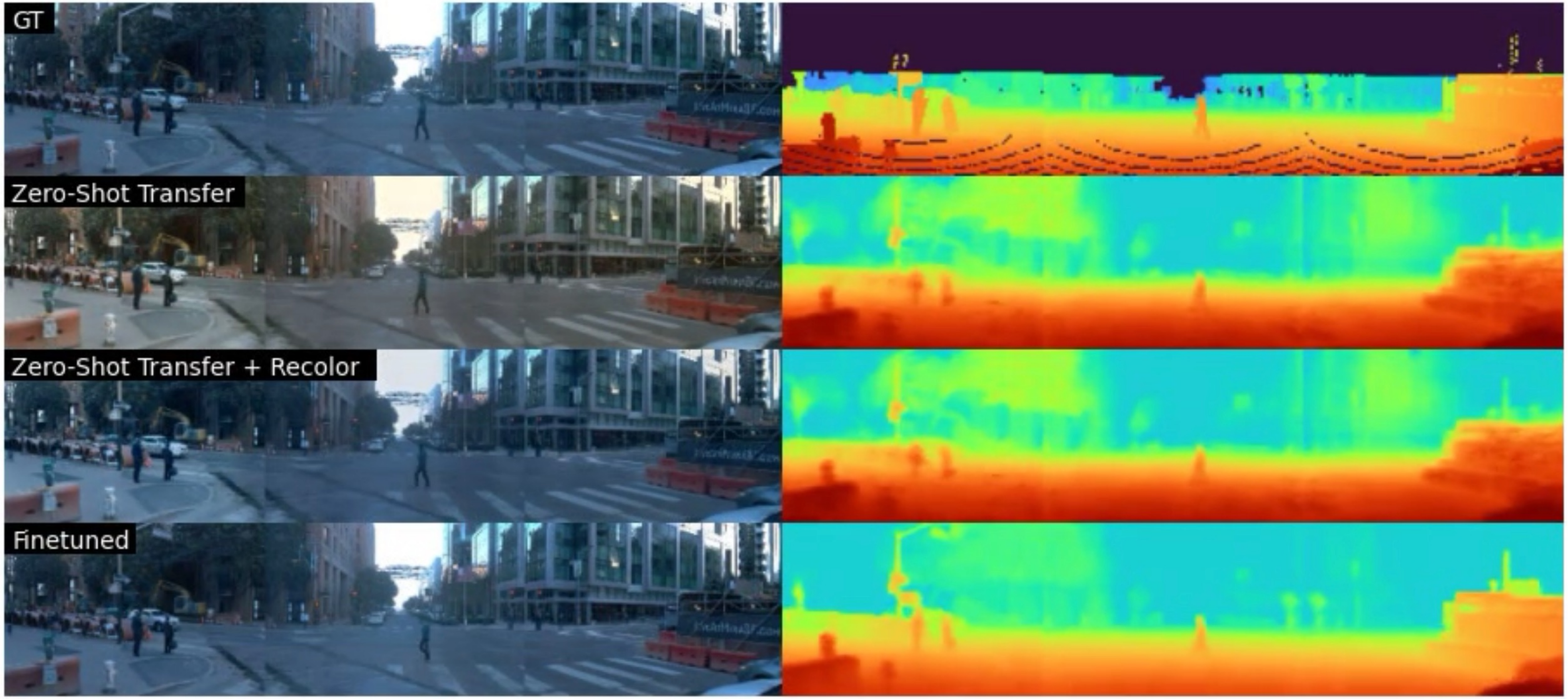
Foundation Model Feature Prediction

Downstream Tasks

Novel View Synthesis without test-time per-scene optimization, given single-frame images

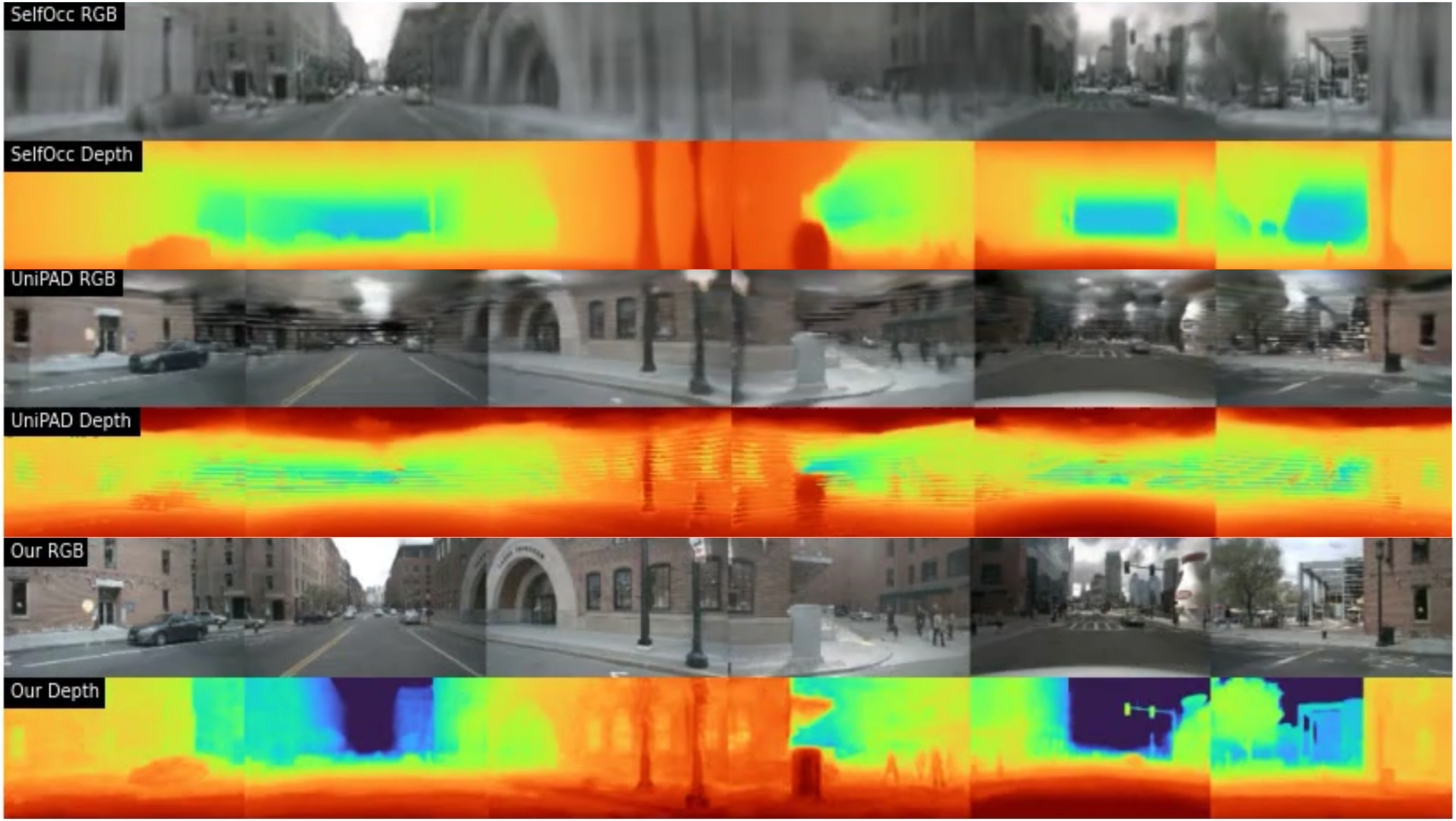


Generalization: Trained on nuScenes, strong generalized performance on the unseen Waymo dataset

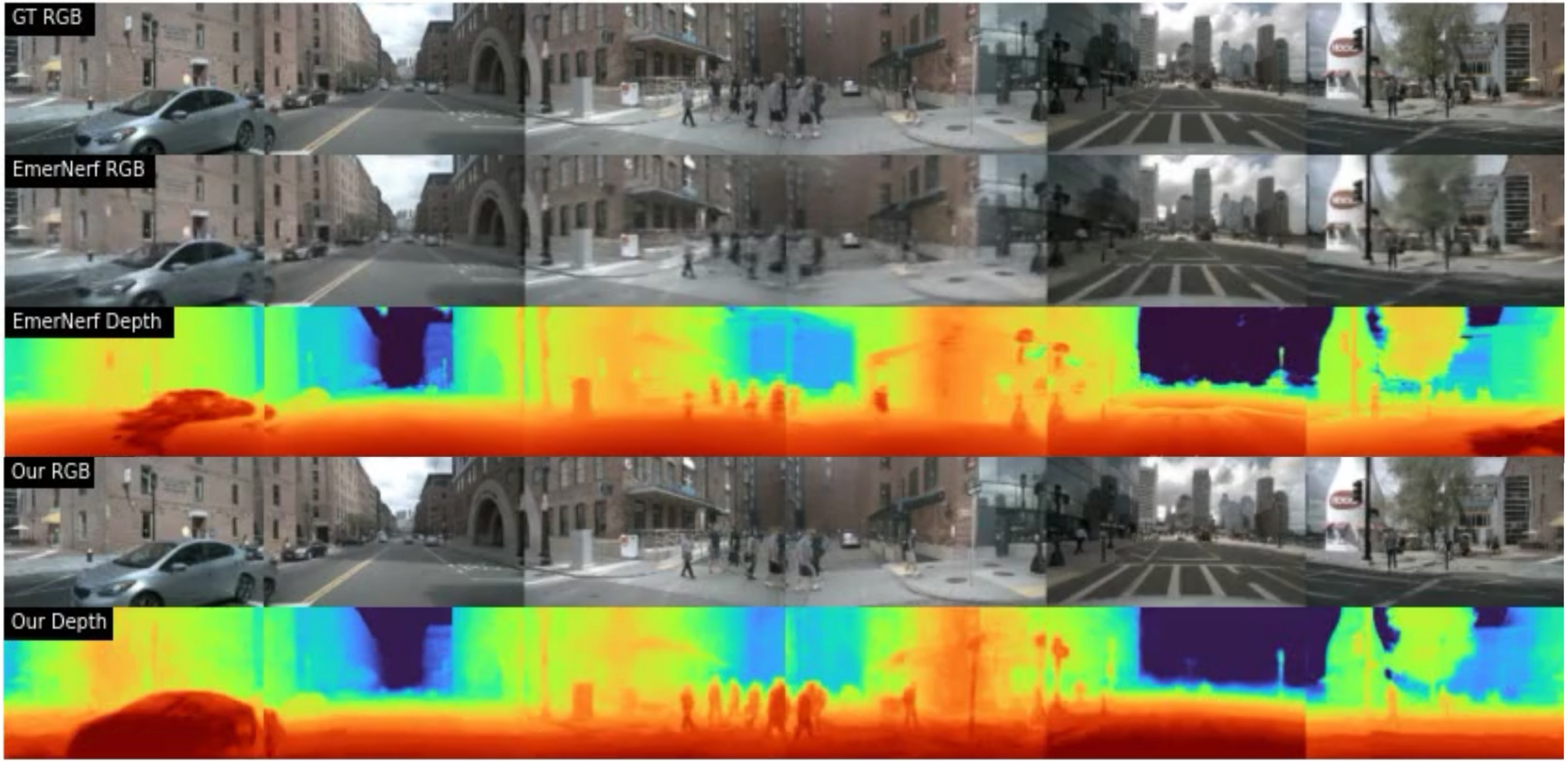


Row 2: zero-shot transfer with decent reconstruction quality
Row 3: enhanced quality via simple color alterations to account for camera-specific coloring effects
Row 4: after fine-tuning, our model surpasses the SOTA per-scene EmerNeRF in the reconstruction quality

Comparison: Significantly outperform SOTA generalizable NeRF methods in driving scenes



Comparison: On-par with SOTA per-scene optimized NeRF in driving scenes (EmerNeRF)



DistillNeRF: Perceiving 3D Scenes from Single-Glance Images by Distilling Neural Fields and Foundation Model Features

NeurIPS 2024

Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic,

Steven Waslander, Yue Wang, Sanja Fidler, Marco Pavone, Peter Karkus

NVIDIA

UNIVERSITY OF TORONTO

USC University of Southern California

NUS National University of Singapore

Stanford University