Controllable Neural Reconstruction for Autonomous Driving

Máté Tóth
aiMotive, Hungary
mate.toth@aimotive.com

Péter Kovács
aiMotive, Hungary
peter.kovacs2@aimotive.com

Zoltán Bendefy
aiMotive, Hungary
zoltan.bendefy@aimotive.com

Zoltán Hortsin
aiMotive, Hungary
zoltan.hortsin@aimotive.com

Tamás Matuszka
aiMotive, Hungary
tamas.matuszka@aimotive.com

ABSTRACT

Neural scene reconstruction is gaining importance in autonomous driving, especially for closed-loop simulation of real-world recordings. This paper introduces an automated pipeline for training neural reconstruction models, utilizing sensor streams captured by a data collection vehicle. Subsequently, these models are deployed to replicate a virtual counterpart of the actual world. Additionally, the scene can be replayed or manipulated in a controlled manner. To achieve this, our in-house simulator is employed to augment the recreated static environment with dynamic agents, managing occlusion and lighting. The simulator's versatility allows for various parameter adjustments, including dynamic agent behavior and weather conditions.

ACM Reference Format:
Máté Tóth, Péter Kovács, Zoltán Bendefy, Zoltán Hortsin, and Tamás Matuszka. 2024. Controllable Neural Reconstruction for Autonomous Driving. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH Posters '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3641234.3671082

1 INTRODUCTION

The development of robust autonomous driving systems relies heavily on diverse datasets for training and evaluation purposes. Conventional methodologies leverage recordings of real-world driving scenarios. However, these datasets [Caesar et al. 2020; Matuszka et al. 2023] often suffer from a lack of critical safety-related edge cases due to their inherent rarity. The high costs and logistical complexities associated with capturing such infrequent events necessitate the exploration of alternative approaches. This has led to a surge of interest in synthetic data generation techniques, which offer a promising solution to bridge the gap in safety-critical scenarios for autonomous vehicle development.

This work introduces an end-to-end learning framework for reconstructing extensible static 3D environments from real-world data. The reconstructed environment facilitates the virtual insertion of dynamic agents at arbitrary locations, environmental condition adjustments, and rendering from previously unseen camera viewpoints. Our solution departs from prior methods[Ljungbergh et al. 2024; Zhou et al. 2023] by integrating cutting-edge neural reconstruction techniques within a well-established rendering pipeline. This synergistic approach enables real-time generation of high-fidelity images in full 360° view directions of the desired scenarios. To train our 3D Gaussian Splatting (3DGS) and NeRF-based models, we leverage synchronized data collected by vehicles equipped with RGB cameras, precise GNSS devices, and LiDAR sensors.

SIGGRAPH Posters '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0516-8/24/07
https://doi.org/10.1145/3641234.3671082

2 METHOD

2.1 Neural model generation

Our pipeline uses four input data sources for the reconstruction: *images* from the high-resolution onboard cameras of the vehicle, *point clouds* generated by one or more onboard LiDARs, *egomotion* and *extrinsic and intrinsic calibration* of the cameras.

Since the vanilla implementations of both 3DGS and NERF struggle with temporarily inconsistent scenes, we remove the dynamic agents by masking them from the training images, only leaving the static scene to be learned. While segmenting the RGB images using an off-the-shelf segmentation tool, and masking out every object that potentially can move might be suitable, this approach would lead to serious artifacts behind stationary vehicles. Therefore our solution uses a custom dynamic object masking method. First, we track the segmentation masks of the relevant objects in image space using an optical flow-based frame-to-frame tracking method. Then, we generate 3D bounding boxes of potential dynamic objects in the scene by utilizing our in-house model. We determine the stationarity of the boxes in 3D world space, project them to image space, and match the bounding boxes with the segmentation masks, essentially creating a bounding box-based tracking of the segmentation masks with stationarity information. Then we filter and merge the image space and the bounding box-based tracks and drop the ones corresponding to stationary objects.

High-fidelity 3D scene reconstruction necessitates highly accurate intrinsic calibration and camera poses. We address this by refining camera poses and intrinsic parameters using a modified version of COLMAP.

While these steps would already result in sufficient quality for RGB renderings, the final depth-based compositing step also requires accurate depth information. Since both underlying reconstruction methods fail to estimate depth for weakly textured or blurred surfaces, we explicitly employ depth regularization based on LiDAR data. For this purpose, we use an egomotion-corrected, filtered, and aggregated point cloud. For each frame, we apply an adaptive voxel downsampling based on the distance from the ego vehicle to reduce the number of points to be processed. Then, we

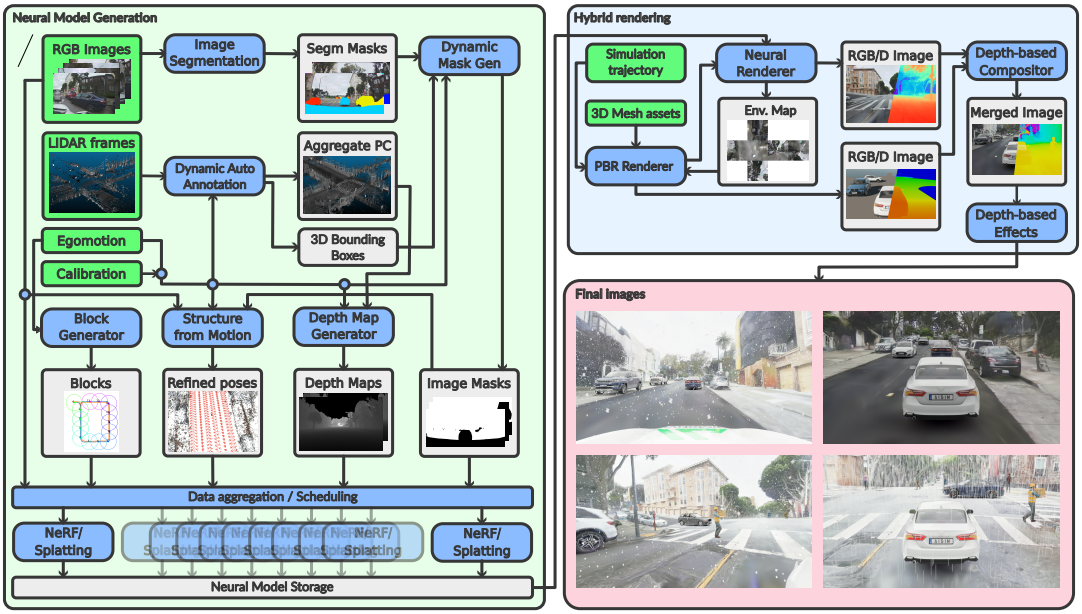

Figure 1: Schematic visualization of our simulation pipeline. Left: static scene reconstruction. Right: dynamic agent composition. The bright green, blue, and pink nodes correspond to the input, processing, and final output steps, respectively.

calculate the occluded point cloud using Open3D [Zhou et al. 2018] for each camera and generate a sparse depth map by projecting the points to the corresponding camera's image space. Finally, we apply a classical depth map completion algorithm [Ku et al. 2018] to acquire the dense depth map.

Autonomous driving tasks often need large scenes to be reconstructed in a highly scalable manner, however, neural reconstruction methods do not scale well to very large scenes. To tackle this problem, we opted to decompose the scene to multiple overlapping blocks in a simplified Block-NeRF [Tancik et al. 2022] like manner.

We use a customized version of Nerfstudio [Tancik et al. 2023] to train the individual models. A modified version of the Depth-Nerfacto model with depth loss scheduling, omnidirectional camera model support, and a learned affine color transform-based color correction is used for NeRF training, and a depth-supervised modification of the Splatfacto model is used as the 3DGS method.

2.2 Hybrid rendering pipeline

Our graphics pipeline is physically-based and responsible for rendering neural model-based static backgrounds and mesh-based dynamic objects. Image-based lighting (IBL) is also part of the pipeline so that the amount of light from the background is used to shade the dynamic objects, to ensure realistic lighting, reflections, and seamless integration. If tessellated ground surfaces are also available, an ambient occlusion (AO) method can be utilized for realistic contact shadows for dynamically placed objects. The depth output from the neural reconstruction model lets us use depth compositing to merge all elements and generate the final output image. This compositing step optionally allows us to add precipitation, like snow or rain, to the scene, which is also influenced by the IBL.

3 LIMITATIONS AND FUTURE WORK

Most of the limitations of our method stem from the characteristics of the automotive-grade cameras and the image signal processing pipelines we are using. These imaging systems typically employ rolling shutter sensors and utilize auto white balance, auto gain, and dynamic tone mapping. This can lead to artifacts, even when using appearance embeddings. Additional information about the limitations, ongoing works such as LiDAR simulation, and implementation details can be found in the supplementary material.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A multimodal dataset for autonomous driving. arXiv:1903.11027 [cs.LG]

Jason Ku, Ali Harakeh, and Steven L Waslander. 2018. In Defense of Classical Image Processing: Fast Depth Completion on the CPU. In *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 16–22.

William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. 2024. NeuroNCAP: Photorealistic Closed-loop Safety Testing for Autonomous Driving. https://doi.org/10.48550/ARXIV.2404.07762

Tamás Matuszka, Iván Barton, Ádám Butykai, Péter Hajas, Dávid Kiss, Domonkos Kovács, Sándor Kunsági-Máté, Péter Lengyel, Gábor Németh, Levente Pető, et al. 2023. aiMotive Dataset: A Multimodal Dataset for Robust Autonomous Driving with Long-Range Perception. In *ICLR 2023 Workshop on SR4AD*.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. 2022. Block-NeRF: Scalable Large Scene Neural View Synthesis. *arXiv* (2022).

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. 2023. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *ACM SIGGRAPH 2023 Conference Proceedings (SIGGRAPH '23)*.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A Modern Library for 3D Data Processing. *arXiv:1801.09847* (2018).

Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. 2023. DrivingGaussian: Composite Gaussian Splatting for Surrounding Dynamic Autonomous Driving Scenes. https://doi.org/10.48550/ARXIV.2312.07920