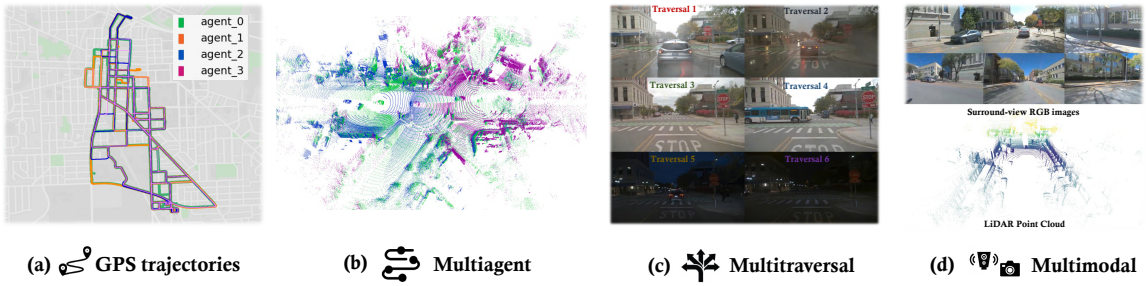
Multiagent Multitraversal Multimodal Self-Driving: Open MARS Dataset

Yiming Li Zhiheng Li Nuo Chen Moonjun Gong
Zonglin Lyu Zehong Wang Peili Jiang Chen Feng
New York University
yimingli@nyu.edu, cfeng@nyu.edu



(a) GPS trajectories (b) Multiagent (c) Multitraversal (d) Multimodal

Figure 1. **Overview of MARS. (a)** Within a geographical area, we operate four autonomous vehicles, displaying their GPS trajectories from a single day using different colors. **(b)** Vehicles occasionally come close together (visualized via distinct-colored point clouds), supporting research in multiagent systems. **(c)** We collect sensory data from repeated traversals of the same location under varying conditions, for learning and perception research with retrospective memory. **(d)** The dataset includes surround-view RGB images and LiDAR point clouds for cross-modal perception and learning. Note that our data is obtained from May Mobility : https://maymobility.com/

Abstract

Large-scale datasets have fueled recent advancements in AI-based autonomous vehicle research. However, these datasets are usually collected from a single vehicle's one-time pass of a certain location, lacking multiagent interactions or repeated traversals of the same place. Such information could lead to transformative enhancements in autonomous vehicles' perception, prediction, and planning capabilities. To bridge this gap, in collaboration with the self-driving company May Mobility*, we present* ***MARS*** *dataset which unifies scenarios that enable* ***M****ulti****A****gent, multitrave****RS****al, and multimodal autonomous vehicle research. More specifically, MARS is collected with a fleet of autonomous vehicles driving within a certain geographical area. Each vehicle has its own route and different vehicles may appear at nearby locations. Each vehicle is equipped with a LiDAR and surround-view RGB cameras. We curate two subsets in MARS: one facilitates collaborative driving with multiple vehicles simultaneously present at the same location, and the other enables memory retrospection through asynchronous traversals of the same location by multiple vehicles. We conduct experiments in place recognition and neural reconstruction. More importantly, MARS introduces new research opportunities and challenges such as multitraversal 3D reconstruction, multiagent perception, and unsupervised object discovery. Our data and codes can be found at https://ai4ce.github.io/MARS/.*

1. Introduction

Autonomous driving, which has the potential to fundamentally enhance road safety and traffic efficiency, has witnessed significant advancements through AI technologies in recent years. Large-scale, high-quality, real-world data is crucial for AI-powered autonomous vehicles (AVs) to enhance their perception and planning capabilities [1, 14]: AVs can not only learn to detect objects from annotated datasets [15] but also create safety-critical scenarios by generating digital twins based on past driving recordings [16].

The pioneering KITTI dataset [1] established the initial benchmark for tasks such as detection and tracking. Since its introduction, a number of datasets have been proposed to promote the development of self-driving; see Tab. 1. Two representative datasets are nuScenes [7] and Waymo Dataset [8] which introduce multimodal data collected from cameras and range sensors, covering a 360-degree field of view for panoramic scene understanding. These datasets have shifted the focus from KITTI's monocular cameras, receiving wide attention in the fields of vision and robotics.

Table 1. **Comparison of existing autonomous driving datasets with multimodal sensors.** C denotes the camera and L denotes LiDAR.

Datasets	Sensors	Camera view	Location	Source	Year	Multiagent	Multitraversal
KITTI [1]	C&L	Front	Germany	Academia	2012	✗	✗
Lyft Level 5 [2]	C&L	Surround	U.S.	Industry	2019	✗	✓
Argoverse [3]	C&L	Surround	U.S.	Industry	2019	✗	✗
ApolloScape [4]	C&L	Front	China	Industry	2019	✗	✗
A2D2 [5]	C&L	Surround	Germany	Industry	2020	✗	✗
A*3D [6]	C&L	Front	SG	Academia	2020	✗	✗
nuScenes [7]	C&L	Surround	U.S. & SG	Industry	2020	✗	✓
Waymo Open Dataset [8]	C&L	Surround	U.S.	Industry	2020	✗	✗
ONCE [9]	C&L	Surround	China	Industry	2021	✗	✗
KITTI-360 [10]	C&L	Surround	Germany	Academia	2022	✗	✗
Ithaca365 [11]	C&L	Front	U.S.	Academia	2022	✗	✓
V2V4Real [12]	C&L	Front&Back	U.S.	Academia	2023	✓	✗
Zenseact Open Dataset [13]	C&L	Front	Europe	Industry	2023	✗	✗
Open MARS Dataset (Ours)	C&L	Surround	U.S.	Industry	2024	✓	✓

Existing driving datasets generally focus on geographical and traffic diversity without considering two practical dimensions: multiagent (*collaborative*) and multitraversal (*retrospective*). The *collaborative* dimension highlights the synergy between multiple vehicles located in the same spatial region, facilitating their cooperative perception, prediction, and planning. The *retrospective* dimension enables vehicles to enhance their 3D scene understanding by drawing upon visual memories from prior visits to the same place. Embracing these dimensions can address challenges like limited sensing capability for online perception and sparse views for offline reconstruction. Nevertheless, existing datasets are typically collected by an individual vehicle during a one-time traversal of a specific geographical location. To advance autonomous vehicle research, especially in the *collaborative* and *retrospective* dimensions, the research community needs a more comprehensive dataset in real-world driving scenarios. To fill the gap, we introduce the Open **MARS** Dataset, which provides **M**ulti**A**gent, multitrave**RS**al, and multimodal recordings, as shown in Fig. 1. *All the recordings are obtained from May Mobility*[1]*'s autonomous vehicles operating in the real world.*

- **Multiagent.** We deploy a fleet of autonomous vehicles to navigate a designated geographical area. These vehicles can be in the same locations at the same time, allowing for collaborative 3D perception through vehicle-to-vehicle comminication.
- **Multitraversal.** We capture multiple traversals within the same spatial area under varying lighting, weather, and traffic conditions. Each traversal may follow a unique route, covering different driving directions or lanes, resulting in multiple trajectories that provide diverse visual observations of the 3D scene.
- **Multimodal.** We equip the autonomous vehicle with RGB cameras and LiDAR, both with a full 360-degree field of view. This comprehensive sensor suite can enable multimodal and panoramic scene understanding.

[1] https://maymobility.com/

We conduct quantitative and qualitative experiments in place recognition and neural reconstruction. More importantly, MARS introduces novel research challenges and opportunities for the vision and robotics community, including but not limited to *multiagent collaborative perception and learning, unsupervised perception under repeated traversals, continual learning, neural reconstruction and novel view synthesis with multiple agents or multiple traversals*.

2. Related Works

Autonomous driving datasets. High-quality datasets are crucial for advancing AI-powered autonomous driving research [7, 17, 18]. The seminal KITTI dataset significantly attracted research attention in robotic perception and mapping [1, 19–21]. Since then, a large number of datasets have been proposed, pushing the boundaries of the field by tackling challenges in multimodal fusion, multitasking learning, adverse weather, and dense traffic [6, 7, 10, 22–24]. In recent years, researchers have proposed multiagent collaboration to get rid of the limitations in single-agent perception, *e.g.*, frequent occlusion and long-range sparsity [25–31]. Previous efforts in curating multiagent datasets are usually limited by simulated environments [32, 33]. The recent V2V4Real [12] supports vehicle-to-vehicle cooperative object detection and tracking in the real world, yet the two-camera setup is insufficient for surround-view perception. Another relevant dataset, Ithaca365 [11], provides recordings from repeated traversals of the same route in different lighting and weather conditions, yet it only uses front-view cameras for data collection. To fill the gap, our MARS dataset provides multiagent, multitraversal, and multimodal driving recordings with a panoramic camera view; see Tab. 1. *Notably, the continuous and dynamic operation of May Mobility's fleet of vehicles makes our MARS dataset stand out in scale and diversity, featuring hundreds of traversals at a single location and enabling collaborative driving for up to four vehicles, thereby setting a record for both traversal and agent numbers.*

Table 2. **May Mobility sensor suite specification** of each vehicle.

Sensor	Details
1 × LiDAR	10Hz, 128 channel, horizontal FoV 360°, vertical FoV 40°
3 × RGB Camrea	10Hz, original resolution 1440 × 928, sampled to 720×464, Horizontal FoV 60°, Vertical FoV 40°
3 × Fisheye Camrea	10Hz, original resolution 1240 × 728, sampled to 620×364, horizontal FoV 140°, vertical FoV 88°
1 × IMU	10Hz, velocity, angular velocity, acceleration
1 × GPS	10Hz, longitude, latitude, elevation

Visual place recognition. In the field of computer vision and robotics, visual place recognition (VPR) holds significant importance, enabling the recognition of specific places based on visual inputs [34]. Specifically, VPR systems function by comparing a given query data, usually an image, to an existing reference database and retrieving the most similar instances to the query. This functionality is essential for vision-based robots operating in GPS-unreliable environments. VPR techniques generally fall into two categories: traditional methods and learning-based methods. Traditional methods leverage handcrafted features [35, 36] to generate global descriptors [37]. However, in practice, *appearance variation* and *limited viewpoints* can degrade VPR performance. To address the challenge of *appearance variation*, learning-based methods utilize deep feature representations [38–40]. In addition to image-based VPR, video-based VPR approaches [41–43] are proposed to achieve better robustness, mitigating the *limited viewpoints* with video clips. Moreover, CoVPR [44] introduces collaborative representation learning for VPR, bridging the gap between multiagent collaboration and place recognition, and addressing *limited viewpoints* by leveraging information from collaborators. Beyond 2D image inputs, PointNetVLAD [45] explores point-cloud-based VPR, offering a unique perspective on place recognition. In this paper, we evaluate both single-agent VPR and collaborative VPR.

NeRF for autonomous driving. Neural radiance fields (NeRF) [46] in unbounded driving scenes has recently received a lot of attention, as it not only facilitates the development of high-fidelity neural simulators [16] but also enables high-resolution neural reconstruction of the environment [47]. Regarding novel view synthesis (NVS), researchers have addressed the challenges such as scalable neural representations with local blocks [48, 49], dynamic urban scene parsing with compositional fields [50, 51], and panoptic scene understanding with object-aware fields [52, 53]. Regarding neural reconstruction, researchers have realized decent surface reconstruction based on LiDAR point cloud and image input [54, 55]. Meanwhile, several efforts have been made in multi-view implicit surface reconstruction without relying on LiDAR [47]. Existing methods based on NeRF are constrained by limited visual observations, often relying on sparse camera views collected along a narrow trajectory. There is significant untapped potential in leveraging additional camera perspectives, whether from multiple agents or repeated traversals, to enrich the visual input and enhance the NVS or reconstruction performance.

Figure 2. **Sensor setup** of the vehicle platform for data collection.

3. Dataset Curation

3.1. Vehicle Setup

Sensor setup. *May Mobility*'s fleet of vehicles includes four Toyota Sienna, each mounted with one LiDAR, three narrow angle RGB cameras, three wide angle RGB fisheye cameras, one IMU, and one GPS. The sensors have various raw output frequency, but all sensor data are eventually sampled to 10Hz for synchronization. Camera images are down-sampled to save storage. Detailed specifications of these sensors are listed in Tab. 2. In general, the LiDAR is located at the front top of the vehicle. The three narrow angle cameras are located at the front, front left, and front right of the vehicle. Three fisheye cameras are on the back center, left side, and right side of the vehicle; see Fig. 2. The IMU and GPS are located at the center top of the vehicle. The explicit extrinsic of these sensors are expressed as rotations and translations that transform sensor data from its own sensor frame to the vehicle's ego frame. For each camera on each vehicle, we provide camera intrinsic parameters and distortion coefficients. The distortion parameters were inferred by the AprilCal calibration method [56].

Coordinate system. There are four coordinate systems: sensor frame, ego frame, local frame, and global frame. Sensor frame represents the coordinate system whose origin is defined at the center of an individual sensor. The ego frame represents the coordinate system whose origin is defined at the center of the rear axle of an ego vehicle. The local frame represents the coordinate system whose origin is defined at the start point of an ego vehicle's trajectory of the day. The global frame is the world coordinate system.

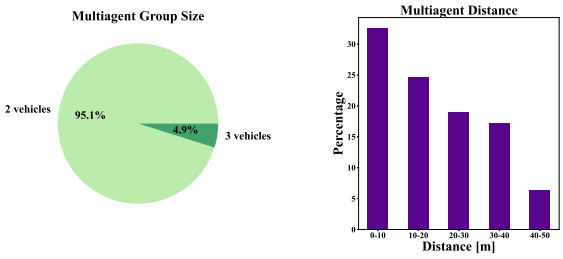


Figure 3. **Multiagent subset statistics.**

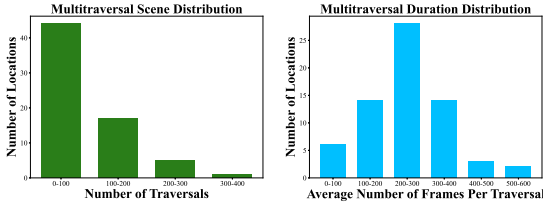


Figure 4. **Multitraversal subset statistics.**

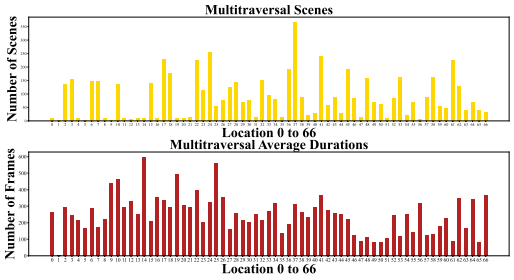


Figure 5. **Number of traversals and frames at each location.**

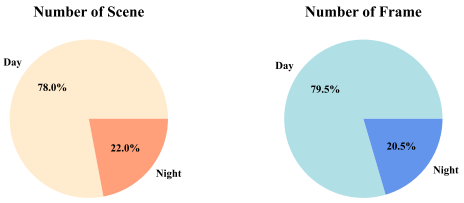


Figure 6. **Ratio of day and night scenes.**

3.2. Data Collection

May Mobility is currently focusing on micro service transportation, running shuttle vehicles on fixed routes in various orders and directions. The full route is over 20 kilometers long, encompassing residential, commercial, and university campus areas with diverse surroundings in terms of traffic, vegetation, buildings, and road marks. The fleet operates every day between 2 to 8 p.m., therefore covering various lighting and weather conditions. Altogether, May Mobility's unique mode of operation enabled us to collect multi-traversal and multiagent self-driving data.

Multitraversal data collection. We defined a total of 67 locations on the driving route, each spanning a circular area of a 50-meter radius. These locations cover different driving scenarios such as intersections, narrow streets, and long-straight roads with various traffic conditions. The traversals at each location take place from different directions at different times of each day, promising physically and chronologically comprehensive perceptions of the area. We determine via the vehicle's GPS location whether it is traveling through a target location, and data is collected for the full duration of the vehicle's presence within the 50-meter-radius area. Traversals are filtered such that each traversal is between 5 seconds to 100 seconds long.

Multiagent data collection. A highlight of our dataset is that we provide real-world synchronized multi-agent collaborative perception data that delivers extremely detailed spatial coverage. Determining from vehicles' GPS coordinates, we extract 30-second-long scenes where two or more ego vehicles have been less than 50 meters away from each other for more than 9 seconds, collectively providing overlapping perceptions of the same area at the same time but from different angles. For scenes where the encountering persisted less than a full 30 seconds, the encountering segment is placed at the center of the 30-second duration, with equal amount of non-encountering time filled before and after it (*e.g.* 20 seconds of encountering gets extended to a 30-second scene by adding 5 seconds before and 5 seconds after). Such encountering can take place anywhere around the map, constituting scenarios such as tailgating along a straight road and meeting at intersections, as shown in Fig. 7. Our method also ensures that at least one vehicle in the scene travels over 10 meters within 30 seconds.

3.3. Dataset Statistics

The multitarversal subset covers data from 26 different days between October 4th, 2023 and March 8th, 2024 , 4 of which were rainy. We collected a total of 5,757 traversals containing over 1.4 million frames of images for each camera and 360-degree LiDAR point clouds. Among the 67 locations, 48 have over 20 traversals, 23 over 100 traversals, and 6 over 200 traversals. Each traversal has 250 frames (25 seconds) on average, with the majority of traversals containing 100 to 400 frames (10 to 40 seconds). The specific distributions of traversals and frames across all locations are shown in Fig. 4 and Fig. 5. The muitlagent subset covers data from 20 different days between October 23rd, 2023 and March 8th, 2024. We collected 53 scenes of 30-second duration, stably involving 297 to 300 frames each scene, accounting for over 15,000 frames of images and LiDAR point clouds in total. Among the 53 scenes, 52 involves two vehicles, and 1 involves three vehicles. The distance between each pair of ego vehicles is analyzed for every frame. The distribution demonstrates that encountering takes place mostly with two vehicles being less than 50 meters away from each other, as shown in Fig. 3.

Figure 7. **Multiagent scene visualizations with three front cameras and LiDAR point clouds in bird's eye view (BEV).** Typical scenarios include straight road tailgating as well as meeting at intersection.

4. Benchmark Task and Model

4.1. Place Recognition

Problem definition. We consider a set of queries $\mathbf{Q}$ with M images and a reference database $\mathbf{D}$ with N images. In this task, the objective is to find $I_r \in \mathbf{D}$ given $I_q \in \mathbf{Q}$ such that I_q and I_r are captured at the same location.

Evaluation metric. We adopt recall at K as our evaluation metric for VPR. For a query image I_q, we select K reference images with Top-K cosine similarities between X_q and $\{X_r\}_{r=1}^{N}$. If at least one of the selected images is captured within S meters of I_q ($S = 20$ in this paper), then we count it as correct. The recall at K is computed as the ratio between the total number of correct counts and M.

Benchmark models. We adopt NetVLAD [38], PointNetVLAD [45], MixVPR [39], GeM [57], Plain ViT [58], and CoVPR [44] as benchmark models.

- **NetVLAD** consists of a CNN-based backbone and a NetVLAD pooling layer. NetVLAD replaces the hard assignment in VLAD [37] with a learnable soft assignment, taking features extracted by backbones as input and generating a global descriptor.
- **MixVPR** consists of a CNN-based backbone and a feature-mixer. The output of the backbone is flattened to $C \times H'W'$, fed to the feature-mixer with row-wise and column-wise MLPs, flattened to a single vector, and L^2-normalized.
- **PointNetVLAD** consists of a backbone, a NetVLAD pooling, and an MLP. We reduced the output dimension of the backbone from 1024 to 256 and omitted the last MLP layer for efficient computation.
- **GeM** consists of a CNN-based backbone and a GeM pooling. The GeM pooling is defined as $\frac{1}{N}(\sum_{i=1}^{N} X_i^p)^{\frac{1}{p}}$, where X_i is the patch feature, and we select p = 3 here.
- **Plain ViT** [58] consists of standard transformer encoder layers and a L^2 normalization over cls toekn.
- **CoVPR** [44] consists of a VPR model and a similarity-regularized fusion. The VPR model generates descriptors for the ego agent and collaborators, and the fusion module fuses them into a single descriptor.

4.2. Neural Reconstruction

Problem definition. Based on the number of available traversals, we divided the reconstruction task into two scenarios. The first is *single-traversal (dynamic scene recon-*

struction), where the input is a sequence of images $\mathcal{I} = \{I_1, I_2, \cdots I_k\}$ captured as one traversal video. And the goal is to reconstruct photorealistic scene views, including moving objects. The second is *multitraversal (environment reconstruction)*, where the input is a collection of image sequences $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n : \mathcal{I}_m = \{I_{m,1}, \cdots, I_{m,k_m}\}\}$ of the same scene. The objective in this task is to reconstruct the environment and remove dynamic objects.

Evaluation metrics. Building on the methods used in earlier works [59]. we use PSNR, SSIM and LPIPS metrics for our experiments of dynamic reconstruction. PSNR, defined as $PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$, assesses image quality by comparing maximum pixel value MAX_I and mean squared error MSE. SSIM, calculated by $SSIM(x, y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$, measures similarity between synthesized and ground truth images, factoring in mean, variance, and covariance. LPIPS, unlike the two metrics before, uses a pretrained neural network model to evaluate the perceptual similarity between two images.

Benchmark models. For the single-traversal task, we adopt EmerNeRF [60] and PVG [59] as benchmark models. Additionally, for comparison, we conduct experiments using iNGP [61] and 3DGS [62], which do not directly target this problem. Regarding multitraversal reconstruction, there are no algorithms specifically designed for this task. Therefore, we adopt iNGP as the basic model. Furthermore, to enhance the model's ability to remove dynamic objects, we also test RobustNeRF [63] and iNGP with Segformer [64].

- **Single-traversal: Dynamic scene reconstruction.**
  - **EmerNeRF.** Based on neural fields, EmerNeRF is a self-supervised method for effectively learning spatial-temporal representations of dynamic driving scenes. EmerNeRF builds a hybrid world representation by breaking scenes into static and dynamic fields. By utilizing an emergent flow field, temporal information can be further aggregated, enhancing the rendering precision of dynamic components. The 2D visual foundation model features are lifted into 4D space-time to augment EmerNeRF's semantic scene understanding.
  - **PVG.** Building upon 3DGS, PVG introduces periodic vibration into each Gaussian point to model the dynamic motion of these points. To handle the emergence and vanishing of objects, it also sets a time peak and a lifespan for each point. By learning all these parameters, along with the mean, covariance, and spherical harmonics of the Gaussians, PVG is able to reconstruct dynamic scenes in a memory-efficient way.
- **Multitraversal: Environment reconstruction.**
  - **RobustNeRF** RobustNeRF replaces the loss function of the original NeRF to ignore distractors, and we consider dynamic objects as distractors in our case. Additionally, RobustNeRF applies a box kernel in its loss estimator to prevent high-frequency details from being recognized as outliers.
  - **SegNeRF.** SegNeRF utilizes the pretrained semantic model SegFormer [64] to remove movable objects.

5. Experimental Results

5.1. Visual Place Recognition

Dataset details. We conduct experiments in VPR tasks with both multitraversal and multiagent data. In the multitraversal case, intersections numbered higher than or equal to 52 are used for testing. In the multiagent setting, scenes numbered higher than or equal to 50 are used for testing. Input images are resized to 400×224, and input point clouds are downsampled to 1024 points.

Implementation details. We evaluate our dataset on models mentioned in Sec. 4, where CoVPR [44] is evaluated with multiagent data, and all others are evaluated with multitraversal data. Backbones are pre-trained on ImageNet1K [65]. We use ResNet18 [66] as the backbone for NetVLAD and CoVPR, ResNet50 [66] for MixVPR and GeM, and PointNet [67] for PointNetVLAD. The number of clusters in NetVLAD-based methods is 32. Models are trained with Adam [68] optimizer with 1e-3 lr for PointNetVLAD, 1e-4 lr for others, and 1e-4 decay rate until convergence. The batch size is 20 for NetVLAD-based methods and 10 for others.

Result discussions. Quantitative results are shown in Tab. 3. Although GeM achieves lightweight characteristics in its pooling methods, it underperforms compared to NetVLAD with a smaller backbone. ViT demonstrates weaker performance in VPR without task-specific pooling methods, despite being a stronger backbone than ResNet. MixVPR achieves the best performance, as its feature-mixing mechanism provides richer features. PointNetVLAD, leveraging point clouds, attains better performance with smaller input sizes than NetVLAD. In the context of multiagent data, CoVPR consistently outperforms its single-agent counterparts. Qualitative results are depicted in Fig. 8. Our dataset encompasses both daytime and nighttime scenes, under various weather conditions such as sunny, cloudy, and rainy. Hard examples stem from nighttime scenarios and cameras affected by rain or backlighting.

5.2. Neural Reconstruction

Dataset details. In our single-traversal dynamic scene reconstruction experiments, we selected 10 different locations, each with one traversal, aiming to capture and represent complex urban environments. For our multitraversal environment reconstruction experiments, we selected a total of 50 traversals. This comprised 10 unique locations, with 5 traversals for each location, enabling us to capture variations in illuminating conditions and weather.

Figure 8. **Qualitative result of VPR.** We use MixVPR to obtain this qualitative result and mark incorrect results with red frames. Our dataset contains hard cases such as nighttime, back-lighting, and blurred cameras due to weather conditions.

Table 3. **Quantitative results of VPR.**

Data	Model	Recall @1	Recall @5	Recall @10
Multitraversal	NetVLAD [38]	63.51	69.60	72.42
	MixVPR [39]	**71.73**	**75.38**	**77.20**
	GeM [57]	61.00	68.47	71.73
	ViT [58]	53.33	58.79	62.37
	PointNetVLAD [45]	66.45	72.82	75.91
Multiagent	NetVLAD [38]	91.85	94.89	95.44
	CoVPR [44]	**92.27**	**95.30**	**95.86**

Implementation Details. Throughout all reconstruction experiments, we utilize 100 images from the three front cameras, along with LiDAR data, as input for each traversal. *Single-traversal experiments:* Both iNGP and EmerNeRF models undergo training for 10,000 iterations utilizing the Adam [68] optimizer with a learning rate of 0.01 and a weight decay rate of 0.00001. For EmerNeRF, we leverage the dino feature from the DINOv2 ViT-B/14 [69] foundation model. The estimator employed in this model is PropNet, incorporating linear disparity and uniform sampling. For 3DGS and PVG, we set the training iteration number to be 20000, with learning rate the same as in the original work [59]. We treat 3DGS as a special case of the PVG method, with a 0 periodic motion amplitude and an infinite lifespan, which we set to 10^6 in our experiments. *Multitraversal experiments:* Our NeRF model in this experiment is iNGP [61] with image embedding and DINO features. For RobustNeRF, we implement the robust loss and patch sample as described in the original paper [63]. In SegNeRF, we apply the SegFormer-B5 [64] model, trained on the Cityscapes [70] dataset. Among the 19 catogories in the SegFormer model, we identify 'person', 'rider', 'car', 'truck', 'bus', 'train', 'motorcycle' and 'bicycle' as dynamic classes and generate masks for them.

Result discussions. *Single-traversal experiments:* Based on the results presented in Tab. 4, PVG achieves higher SSIM scores and better LPIPS scores, indicating enhanced structural details. This superior performance by PVG is likely attributed to its flexible Gaussian points setup, which adeptly captures linear motions, and the emergence and disappearance of objects. EmerNeRF, on the other hand, excels in PSNR. This is likely due to its novel approach of dynamic-static decomposition. As shown in Fig. 9, EmerNeRF and PVG both demonstrate the ability to perfectly render dynamic objects like moving cars, whereas iNGP and 3DGS exhibit relatively poor performance in this regard. *Multitraversal experiments:* Thanks to image embedding, iNGP can render diversely illuminated scenes. However, it struggles with rendering dynamic objects accurately or removing them. As shown in Tab. 4, iNGP achieves the best similarity metrics since it preserves the most information about dynamic objects. RobustNeRF performs best in eliminating dynamic objects, albeit at the cost of rendering static objects with less detail. SegFormer, leveraging semantic information, achieves superior visual results compared to the other two methods. Yet the shadows of cars are not completely removed, likely due to the inadequate recognition of shadows by semantic segmentation models.

Table 4. **Quantitative results of Neural Reconstruction.** We compute average PSNR, SSIM and LPIPS of ten locations to assess the reconstructed appearance.

Task	Model	PSNR ↑	SSIM ↑	LPIPS ↓
Single-traversal	iNGP [61]	28.66	0.821	0.256
	3DGS [62]	27.77	0.867	0.235
	EmerNeRF [60]	**29.63**	0.839	0.237
	PVG [59]	29.28	**0.900**	**0.197**
Multitraversal	iNGP [61]	**26.04**	**0.759**	**0.346**
	RobustNeRF [63]	16.17	0.674	0.459
	SegNeRF [64]	24.44	0.748	0.358

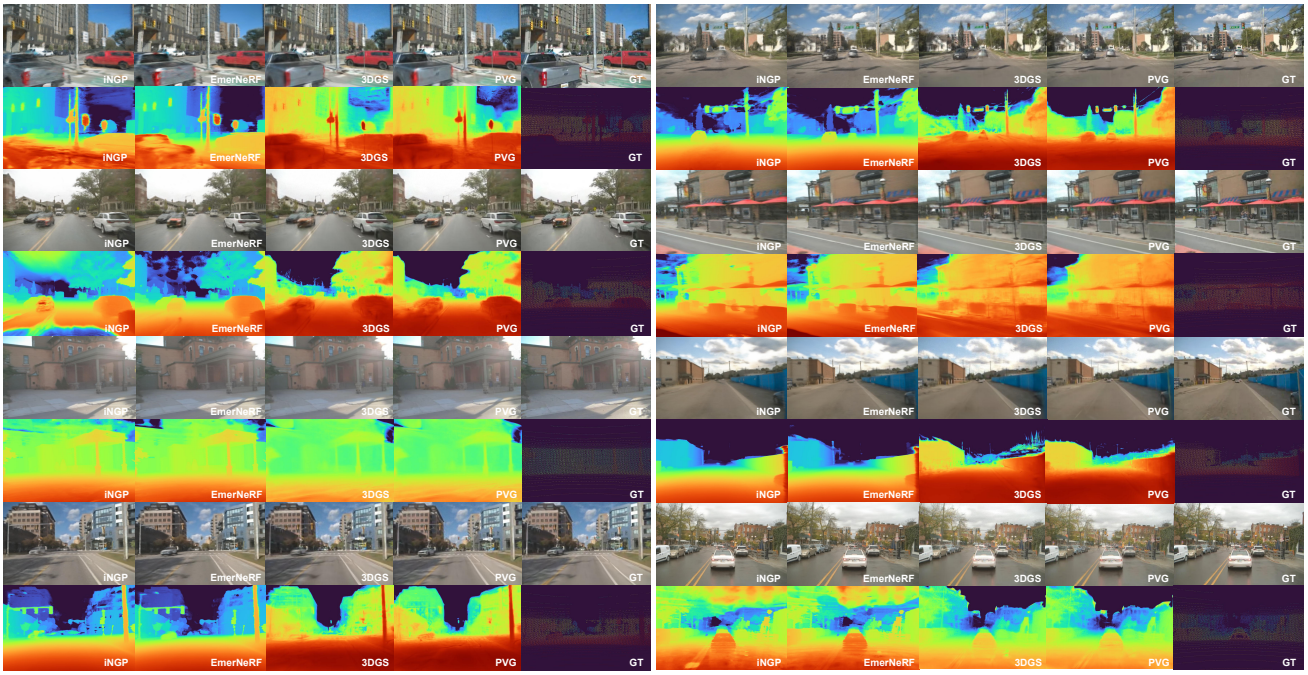


Figure 9. **Qualitative results of single-traversal reconstruction.** We stack the rendered image and the corresponding rendered depth vertically. Each column corresponds to one baseline method and the last column is the ground truth. The ground truth depth is obtained by projecting LiDAR points in the camera view.

6. Opportunities and Challenges

Our MARS dataset introduces novel research opportunities with multiagent driving recordings, as well as a large number of repeated traversals of the same location. We outline several promising research directions and their associated challenges, opening new avenues for future study.

3D reconstruction. Repeated traversals can yield numerous camera observations for a 3D scene, facilitating correspondence search and bundle adjustment in multiview reconstruction. Our dataset can be utilized to study camera-only multitraversal 3D reconstruction, which is crucial for autonomous mapping and localization. The main challenge is to handle appearance variations and dynamic objects across repeated traversals over time.

Neural simulation. Multiagent and multitraversal recordings are valuable for crafting neural simulators that can reconstruct and simulate scenes and sensor data. High-fidelity simulations are essential for developing perception and planning algorithms. The main challenge lies in replicating real-world dynamics and variability, such as modeling the behavior of dynamic objects, environmental conditions, and sensor anomalies, ensuring that the simulated data provides a comprehensive and realistic testbed.

Unsupervised perception. Exploiting scene priors in unsupervised 3D perception offers significant value, especially in multitraversal driving scenarios where abundant data from prior visits can enhance online perception. This approach not only facilitates a deeper understanding of the environment through the accumulation of knowledge over time but also enables unsupervised perception without the need for training with manual annotations.

7. Conclusion

Our MARS dataset represents a notable advancement in autonomous vehicle research, moving beyond traditional data collection methods by integrating multiagent, multitraversal, and multimodal dimensions. MARS opens new avenues for exploring 3D reconstruction and neural simulation, collaborative perception and learning, unsupervised perception with scene priors, *etc*. Future works include providing annotations for online perception tasks such as semantic occupancy prediction in scenarios of multiagent and multitraversal. We strongly believe MARS will establish a new benchmark in AI-powered autonomous vehicle research.

Acknowledgement

All the raw data is obtained from May Mobility. We sincerely thank Mounika Vaka, Oscar Diec, Ryan Kuhn, Shylan Ghoujeghi, Marc Javanshad, Supraja Morasa, Alessandro Norscia, Kamil Litman, John Wyman, Fiona Hua, and Dr. Edwin Olson at May Mobility for their strong support. This work is supported by NSF Grant 2238968 and in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 2

[2] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Woven planet perception dataset 2020. https://woven.toyota/en/perception-dataset, 2019. 2

[3] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 2

[4] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 2

[5] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 2

[6] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE, 2020. 2

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[8] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 2

[9] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, et al. One million scenes for autonomous driving: Once dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2

[10] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[11] Carlos A Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21383–21392, 2022. 2

[12] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 2

[13] Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk, Junsheng Fu, Jenny Widahl, and Christoffer Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20178–20188, 2023. 2

[14] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 1

[15] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 1

[16] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 1, 3

[17] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 2

[18] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 2

[19] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 2

[20] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023.

[21] Chao Chen, Xinhao Liu, Yiming Li, Li Ding, and Chen Feng. Deepmapping2: Self-supervised large-scale lidar map

optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9316, 2023. 2

[22] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 2

[23] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021.

[24] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021. 2

[25] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Advances in Neural Information Processing Systems*, volume 34, pages 29541–29552, 2021. 2

[26] Yiming Li, Juexiao Zhang, Dekun Ma, Yue Wang, and Chen Feng. Multi-robot scene completion: Towards task-agnostic collaborative perception. In *6th Annual Conference on Robot Learning*, 2022.

[27] Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 186–195, 2023.

[28] Sanbao Su, Yiming Li, Sihong He, Songyang Han, Chen Feng, Caiwen Ding, and Fei Miao. Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5588–5594. IEEE, 2023.

[29] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2023.

[30] Sanbao Su, Songyang Han, Yiming Li, Zhili Zhang, Chen Feng, Caiwen Ding, and Fei Miao. Collaborative multi-object tracking with conformal uncertainty propagation. *IEEE Robotics and Automation Letters*, 2024.

[31] Suozhi Huang, Juexiao Zhang, Yiming Li, and Chen Feng. Actformer: Scalable collaborative perception via active queries. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 2

[32] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 2

[33] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 2

[34] Stefan Schubert, Peer Neubert, Sourav Garg, Michael Milford, and Tobias Fischer. Visual place recognition: A tutorial, 2023. 3

[35] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 3

[36] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 2008. 3

[37] Relja Arandjelović and Andrew Zisserman. All about vlad. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013. 3, 5

[38] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 5, 7

[39] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2997–3006, 2023. 5, 7

[40] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[41] Sourav Garg and Michael Milford. Seqnet: Learning descriptors for sequence-based hierarchical place recognition. *IEEE Robotics and Automation Letters*, 2021. 3

[42] Sourav Garg, Madhu Vankadari, and Michael Milford. Seqmatchnet: Contrastive learning with sequence matching for place recognition & relocalization. In *Conference on Robot Learning*, pages 429–443. PMLR, 2022.

[43] Bruno Arcanjo, Bruno Ferrarini, Michael Milford, Klaus D. McDonald-Maier, and Shoaib Ehsan. A-music: An adaptive ensemble system for visual place recognition in changing environments, 2023. 3

[44] Yiming Li, Zonglin Lyu, Mingxuan Lu, Chao Chen, Michael Milford, and Chen Feng. Collaborative visual place recognition. *arXiv preprint arXiv:2310.05541*, 2023. 3, 5, 6, 7

[45] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 5, 7

[46] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[47] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 3

[48] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 3

[49] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12931, 2022. 3

[50] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 3

[51] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12375–12385, 2023. 3

[52] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 3

[53] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 3

[54] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 3

[55] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8370–8380, 2023. 3

[56] Andrew Richardson, Johannes Strom, and Edwin Olson. Aprilcal: Assisted and repeatable camera calibration. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1814–1821. IEEE, 2013. 3

[57] Filip Radenovic, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1655–1668, 2017. 5, 7

[58] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 5, 7

[59] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv:2311.18561*, 2023. 6, 7

[60] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In *International Conference on Learning Representations*, 2024. 6, 7

[61] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 6, 7

[62] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 6, 7

[63] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20626–20636, June 2023. 6, 7

[64] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021. 6, 7

[65] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009. 6

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[67] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 6

[68] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2014. 6, 7

[69] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 7

[70] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7