

RoMe: Towards Large Scale Road Surface Reconstruction via Mesh Representation

Ruohong Mei^{1*}, Wei Sui¹, Jiaxin Zhang^{1*}, Xue Qin², Gang Wang³, Tao Peng⁴ and Cong Yang^{4†}

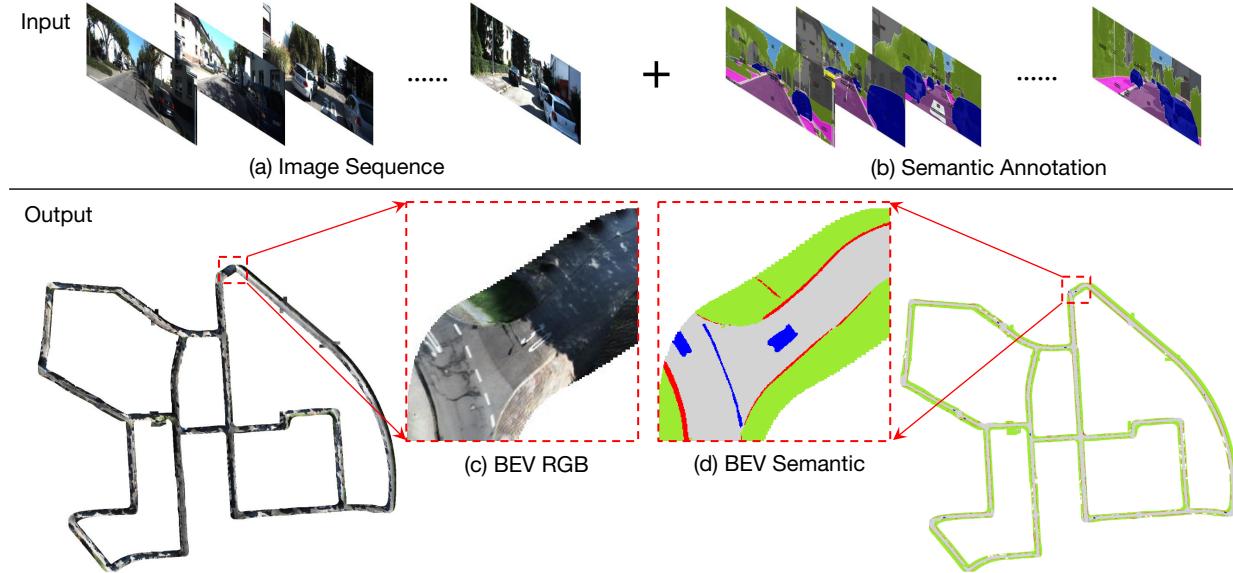


Fig. 1: Road surface reconstruction results (KITTI odometry sequence-00) using our proposed RoMe, covering an area of approximately 600 * 600 square meters. The first row displays the input image sequence with semantic annotations. The second row showcases the final results with close-up details highlighted in red rectangles: the reconstructed BEV RGB surface and its corresponding BEV semantics.

Abstract—In autonomous driving applications, accurate and efficient road surface reconstruction is paramount. This paper introduces RoMe, a novel framework designed for the robust reconstruction of large-scale road surfaces. Leveraging a unique mesh representation, RoMe ensures that the reconstructed road surfaces are accurate and seamlessly aligned with semantics. To address challenges in computational efficiency, we propose a waypoint sampling strategy, enabling RoMe to reconstruct vast environments by focusing on sub-areas and subsequently merging them. Furthermore, we incorporate an extrinsic optimization module to enhance the robustness against inaccuracies in extrinsic calibration. Our extensive evaluations of both public datasets and wild data underscore RoMe’s superiority in terms of speed, accuracy, and robustness. For instance, it costs only 2 GPU hours to recover a road surface of 600 * 600 square meters from thousands of images. Notably, RoMe’s capability extends beyond mere reconstruction, offering significant value for auto-labeling tasks in autonomous driving applications. All related data and code are available at [①](#).

Index Terms—Road Surface Reconstruction, Multilayer Perception Network, Waypoint Sampling, Extrinsic Optimization.

I. INTRODUCTION

In the realm of autonomous driving, bird-eye-view (BEV) perception has emerged as a pivotal tool, aligning seam-

¹ Ruohong Mei, Jiaxin Zhang and Wei Sui are with Horizon Robotics, Haidian District, Beijing, China. ^{*} Equal contribution.

³ Xue Qin is with Harbin Institute of Technology, Harbin, China.

³ Gang Wang is with Shandong University, Shandong, China.

⁴ Tao Peng and Cong Yang are with Soochow University, Suzhou, China.

† Corresponding (cong.yang@suda.edu.cn).

lessly with tasks such as planning and control. This underscores the significance of large-scale road surface reconstruction, especially when it comes to training and validating BEV perception tasks. Broadly, road surface reconstruction methodologies can be bifurcated into two primary categories: traditional methods [1], [2] and those anchored in neural radiance fields (NeRF) [3]–[6].

Traditional Multi-View Stereo (MVS) approaches often yield dense point reconstructions. While these are adept for surfaces with distinct textures, they tend to falter, producing noisy and incomplete results for more uniform road surfaces. Furthermore, their computational demands escalate for expansive reconstructions. Conversely, recent advancements have witnessed the adoption of implicit representation-based methodologies for photorealistic reconstruction, utilizing a curated set of posed images [4]–[6]. These leverage tools such as Multi-Layer Perceptions (MLP) to recreate intricate cityscapes. However, their extensive resource requirements often render them less feasible for large-scale applications.

In real-world scenarios, 3D road surfaces often exhibit discontinuities, suggesting they can be delineated as smooth meshes with nuanced elevations. Motivated by this, we conceived RoMe (Road Mesh), a methodical approach for large-scale road surface reconstruction, reliant solely on images. As delineated in Fig. 1, RoMe crafts a comprehensive 3D road mesh from a sequence of images, complemented by their semantic annotations. Each mesh vertex encapsulates details of elevation, color, and semantics. Fig. 2 presents

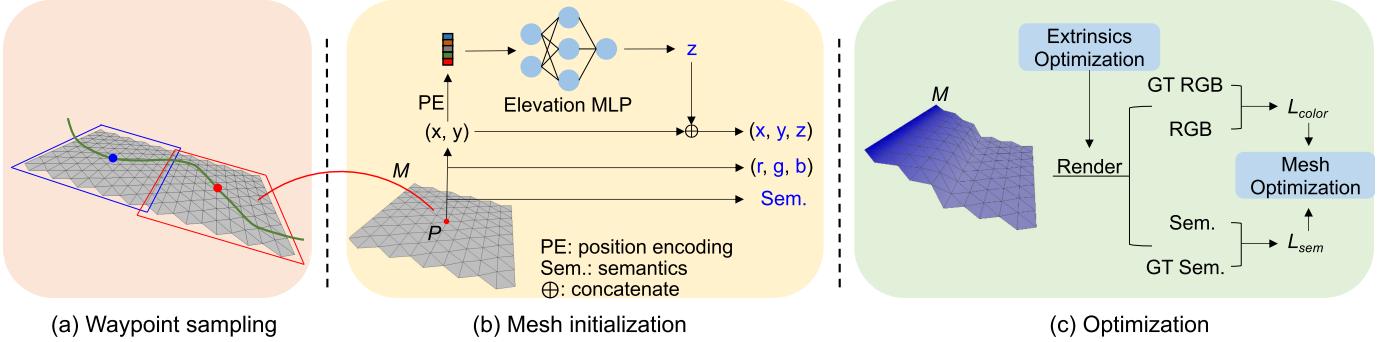


Fig. 2: Overview of RoMe. (a) Waypoint sampling: The green line depicts the camera’s path. Red and blue boxes indicate neighboring subareas, with corresponding red and blue dots representing waypoint samples, aiding in faster training. (b) Mesh initialization: Upon initializing mesh M , vertices are assigned a position (x, y, z) , color (r, g, b) , and semantic attributes. The elevation z of each vertex is fine-tuned using an elevation MLP network. (c) Optimization: The optimization targets, L_{color} and L_{sem} , enable rendering mesh M into RGB images with associated semantics. The parameters $(z, (r, g, b), \text{ and Sem.})$, highlighted in blue in (b)) are collectively adjusted to produce the final road mesh M . Best viewed in color.

the general idea of RoMe: (1) Waypoint sampling: aims to expedite the reconstruction process via a divide-and-conquer strategy: iteratively reconstruct subareas (only a tiny portion of the current view) rather than the whole surface. Herein, the green trajectory epitomizes the camera’s path, with the red and blue boxes demarcating adjacent subareas. (2) Mesh initialization: each vertex is encoded by position, color, and semantics. The elevation of each vertex is adeptly modeled via an MLP network. (3) Mesh optimization: focusing on color and semantics, facilitating the rendering of the mesh into RGB images with corresponding semantics. This intricate process ensures the joint optimization of parameters, culminating in the final reconstructed road mesh. To further boost the robustness of RoMe on cameras and environments, we also incorporate a mechanism to fine-tune the settings mentioned above during the reconstruction process.

In summary, our main contribution is the introduction of RoMe for large-scale road surface reconstruction. We also introduce a mesh representation with a waypoint sampling and optimization strategy to enhance the reconstruction efficiency and robustness of RoMe. Empirical evaluations on public datasets vouch for the precision and resilience of our proposed approach. Furthermore, upon efficient 3D road surface reconstruction, it paves the way for seamless labeling, potentially projecting these labels onto source images, underscoring its utility in automated labeling endeavors.

II. RELATED WORKS

Here, we briefly glanced through several existing multi-view stereo strategies, followed by a review of surface reconstruction methods. For a more detailed treatment of this topic in general, the recent compilation by [7]–[9] offer a sufficiently good review.

A. Multi-View Stereo

3D reconstruction is a process of deducing the three-dimensional structure of an object or scene using multiple images captured from varied camera positions. This domain has witnessed significant advancements over the years [10].

While effective in specific contexts, traditional Multi-View Stereo (MVS) methods often hinge on extracting and matching feature points. The performance is limited in textureless scenes (e.g., road surface) where feature points are sparse and unevenly distributed [1], [2]. Novel view synthesis, which produces photo-realistic images from previously unseen perspectives, shares a close affinity with MVS techniques. While some methods like [11]–[13] are tailored for road surface reconstruction, their scope is limited to smaller areas, making them unsuitable for expansive scenarios. Large-scale MVS methods, applicable even at city levels, have been proposed [14]. These typically involve extracting points from images, constructing sparse 3D points, and subsequently generating meshes. However, they primarily target building structures, often overlooking texture-less surfaces like roads.

Our RoMe approach stands distinct, capable of reconstructing entire road surfaces irrespective of texture variations. It excels in reconstructing expansive road surfaces while preserving essential features such as textures, semantics, and elevations.

B. Surface Reconstruction

Existing MVS methods are not computationally efficient for road surface reconstruction since they model whole scenes through dense point clouds. In practice, existing road surface reconstruction techniques can be broadly categorized into explicit and implicit methods. For the first one, Tong et al. [15] introduced a system that leverages cameras to construct large-scale semantic maps. However, these methods heavily depend on inverse perspective mapping (IPM) and may overlook elevation variations on road surfaces. Rendering-based techniques [16], [17] employ mesh representations with view-dependent appearances.

For the second one, implicit surface reconstruction has gained momentum with the advent of NeRF [3], which utilizes implicit representation and voxel rendering to achieve impressive novel view synthesis (NVS) results. Large-scale NeRF techniques aim to capture intricate details of city blocks or driving scenes. However, they often necessitate additional data acquisition tools, such as LiDAR and images from

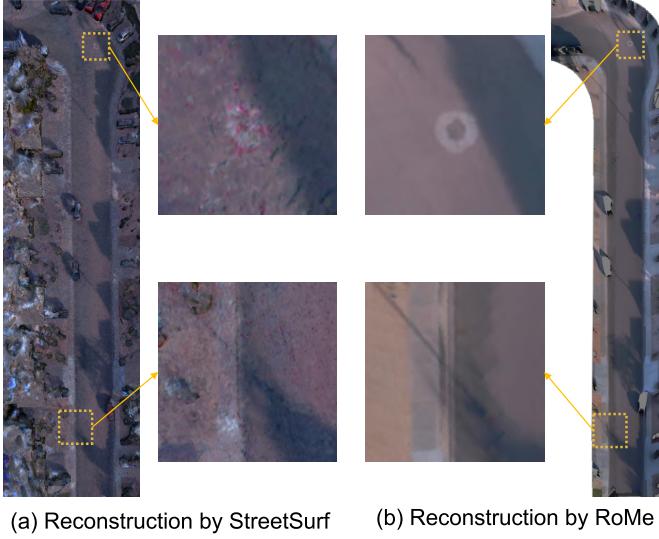


Fig. 3: Street reconstruction by StreetSurf [20] and RoMe.

varied angles [4]–[6]. In contrast, RoMe operates efficiently with a few vehicle-mounted cameras, making it compatible with platforms like nuScenes [18] and KITTI [19]. Besides, our proposed waypoint sampling approach can dramatically improve the reconstruction efficiency via a divide-and-conquer strategy, which is more friendly for parallel computing.

Noted that some pure-vision NeRF-based methods specifically target road surface reconstruction. For instance, Xie et al. [21] employs a voxelized neural radiance field to refine high-definition maps (HD-Maps). Wang et al. [22] introduces a plane regularization technique based on singular value decomposition (SVD) to optimize NeRF’s 3D structure. However, these methods are sensitive to camera pose variations [23]. RoMe, on the other hand, represents the road surface as a 3D mesh, optimizing it using multiple image supervisions, ensuring consistency and resilience to camera pose fluctuations. Though Guo et al. [20] segments the unbounded space into distinct sections (see Fig. 3 (a)), its road surface mesh is blurry and lacks semantics and textures. In contrast, our mesh is smooth, watertight, texture-rich, and properly preserves the original semantics (see Fig. 3 (b)).

III. APPROACHES

RoMe aims to reconstruct road surface textures and semantics using a sequence of images. As illustrated in Fig. 2, RoMe comprises three primary components: Waypoint Sampling, Mesh Initialization, and Optimization. For clarity in terminology, commonly used terms and expressions are defined:

- Ego: self-vehicle, usually same as the mounting position of Inertial Measurement Unit(IMU)/Global Navigation Satellite Systems(GNSS).
- Ego pose: self-vehicle transforms in world coordinate.
- Camera pose: camera transforms in world coordinate.
- Elevation: road surface elevation in world coordinate.
- Waypoints: points that divide road surface to sub-areas for faster reconstruction.

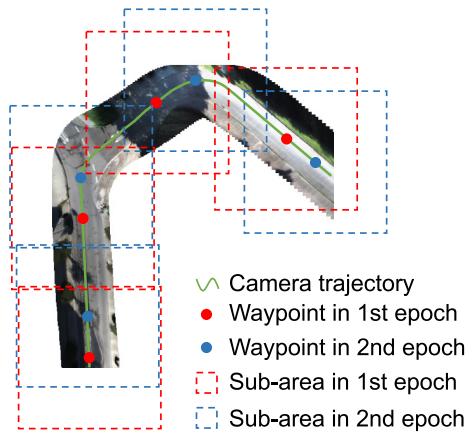


Fig. 4: Illustration of waypoint sampling. The camera trajectory is represented by the green line. Distinct colored dots and their associated boxes indicate sampled waypoints and their corresponding sub-areas across various epochs.

A. Mesh Initialization

Mesh initialization relies on camera poses estimated using ORB-SLAM2 [24] (or COLMAP [1]). ORB-SLAM2 is a real-time SLAM library for monocular, stereo, and RGB-D cameras that computes camera poses and a sparse 3D reconstruction. For instance, we use stereo cameras in KITTI for restoring camera poses. Then, the semantic segmentation method Mask2Former [25] is employed to generate semantics, including roads, curbs, sign lanes, vehicles, etc. Particularly, Mask2Former has robust and state-of-the-art performance on driving datasets like Cityscapes [26] and Mapillary Vistas [27]. These semantics are also used to mask out dynamic objects like vehicles and pedestrians, which could disrupt the consistency of the overall road structure.

We draw inspiration from [20] to achieve a more accurate mesh initialization. Specifically, we extend the ego poses horizontally to obtain semi-dense points. These points are then lowered by approximately equivalent to the ego height. This process yields points that are closer to the road surface. Pretraining the elevation MLP with these points aids in restoring elevation, especially in the areas with steep slopes. In Fig. 2, the initialized flat mesh, denoted by M , consists of equilateral triangles. Each face has three vertices, each vertex P possessing attributes including location (x, y, z) , color (r, g, b) , and semantics. The position encoding is applied to (x, y) , subsequently feeding them into the elevation MLP to predict elevation z as per Eq. 1. The rationale behind using $\text{MLP}(\cdot)$ is to control the smoothness of the road surface by adjusting the frequency of PE.

$$z = \text{MLP}(\text{PE}(x, y)) \quad (1)$$

B. Waypoint Sampling

To expedite the reconstruction of large areas (e.g., 600×600 square meters), we introduce a novel waypoint sampling approach to improve the efficiency of mesh initialization in Section III-A. As presented in Fig. 4, the core principle is divide-and-conquer. In other words, instead of reconstructing

Algorithm 1 Waypoint Sampling**Input:**

All camera poses, P ;
 All camera images, I ;
 Waypoint radius, R ;

Output:

N waypoints, p_1, p_2, \dots, p_N ;
 N corresponding subsets of areas ($A_{sub_1}, A_{sub_2}, \dots, A_{sub_N}$);
 subsets of camera poses ($P_{sub_1}, P_{sub_2}, \dots, P_{sub_N}$);
 subsets of images ($I_{sub_1}, I_{sub_2}, \dots, I_{sub_N}$);
for loop i in range(loops) **do**
 $p_1 \leftarrow$ random select(P);
 $p_2, p_3, \dots, p_N \leftarrow$ farthest point sampling(p_1, R);
for waypoint j in waypoints **do**
 $A_{sub_j} \leftarrow$ square(p_j, R);
 $P_{sub_j} \leftarrow P$ inside A_{sub_j} ;
 $I_{sub_j} \leftarrow I$ inside A_{sub_j} ;
 Cut off gradient backward outside A_{sub_j} ;
 Train and optimize;
end for
end for

the entire road surface in one go, RoMe divides the vast area into smaller, manageable sub-areas centered around waypoints. Each of these sub-areas is then reconstructed individually. Once all sub-areas are processed, they are seamlessly merged to form the complete road surface reconstruction. It enhances computational efficiency and ensures detailed representation across the entire area.

As detailed in Algorithm 1, camera pose positions are treated as a set of point clouds P . The first waypoint is randomly selected from P . Given the desired sampling radius R , the farthest point sampling algorithm selects waypoints (p_1, p_2, \dots, p_N). Subsequent steps involve gathering all camera poses P_{sub_j} and images I_{sub_j} within the radius for each waypoint p_j and traversing each sub-area A_{sub_j} for optimization. This process is iteratively applied until all sub-areas are adequately covered, resulting in the entire road surface being updated. In practice, the initial waypoint is randomly selected in each training epoch to ensure consistency at the boundaries between different sub-areas.

C. Optimization

Our optimization strategy has twofold: (1) Extrinsic optimization to improve the robustness of RoMe on various camera settings and (2) Mesh optimization during the training process on color and semantics.

1) *Extrinsic Optimization*: In the context of camera calibration, extrinsic refers to the parameters that define the position and orientation of the camera in a world coordinate system. They capture the relationship between the camera's local coordinate system and a global, fixed coordinate system. Accurate camera extrinsic is not always guaranteed. For instance, we observed that the extrinsic among nuScenes cameras are not always ideal in some scenes. Ego poses pertain to the

position and orientation of the autonomous vehicle (or ego vehicle) within its environment. It provides a reference frame from which other objects and landmarks can be localized. In our approach, we decouple camera poses into vehicle ego poses and camera extrinsic. Camera extrinsic describes the transformation between the vehicle coordinate system (often called the ego coordinate system) and the camera coordinate system. This transformation is crucial for aligning the visual data captured by the camera and other sensors on the vehicle.

In RoMe, camera extrinsic is expressed as a transform matrix $T = [R|t]$ in SE(3), where $R \in SO(3)$ and $t \in \mathbb{R}^3$ denote rotation and translation, respectively. Translation t can be easily optimized because it is defined in Euclidean space. Rotation R is expressed as the axis-angle: $\phi := \alpha\omega$, $\phi \in \mathbb{R}^3$, where α is a rotation angle and ω is a normalized rotation axis. It can be converted to R by Rodrigues' formula:

$$R = I + \frac{\sin(\alpha)}{\alpha} \phi^\wedge + \frac{1 - \cos(\alpha)}{\alpha^2} (\phi^\wedge)^2 \quad (2)$$

in which skew operator $(\cdot)^\wedge$ converts a vector ϕ to a skew matrix:

$$\phi^\wedge = \begin{pmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \end{pmatrix}^\wedge = \begin{pmatrix} 0 & -\phi_2 & \phi_1 \\ \phi_2 & 0 & -\phi_0 \\ -\phi_1 & \phi_0 & 0 \end{pmatrix} \quad (3)$$

In practice, we optimize relative camera extrinsic compared with calibrated extrinsic by α , ϕ , and translation t for faster and easier convergence.

2) *Mesh Optimization*: To derive training supervision, we first input the source mesh M into the differentiable renderer in [28]. Specifically, as shown in Eq. 4, we rasterize M to obtain rendering results of image views from the j -th camera pose π_j :

$$[C_j, S_j, D_j, Mask_j] = \text{Rasterize}(\pi_j, M) \quad (4)$$

Here, C_j , S_j , and D_j represent the j -th rendered RGB, semantic, and depth images, respectively. $Mask_j$ is the corresponding silhouette image indicating the area of supervision. $j = 1, \dots, N$. N is the maximum number of source images and corresponding poses. D_j could be supervised if sparse or dense depth is provided.

Building on Eq. 4, we define the color (aka. RGB) loss L_{color} and the semantics loss L_{sem} for training RGB images and semantics, respectively:

$$L_{color} = \frac{1}{N * \text{sum}(Mask_j)} \sum_{j=1}^N Mask_j * |C_j - \bar{C}_j| \quad (5)$$

$$L_{sem} = \frac{1}{N * \text{sum}(Mask_j)} \sum_{j=1}^N Mask_j * CE(S_j, \bar{S}_j) \quad (6)$$

where \bar{C}_N and \bar{S}_N denote the ground truth of RGB images and semantics, respectively. $CE(\cdot)$ refers to the cross-entropy loss. During training, each vertex is optimized by multiple images from different views. Once all of them (from thousands to millions depending on mesh resolution) are properly optimized, the final mesh (with elevation, colors, and semantics) is obtained to represent the whole road surface.

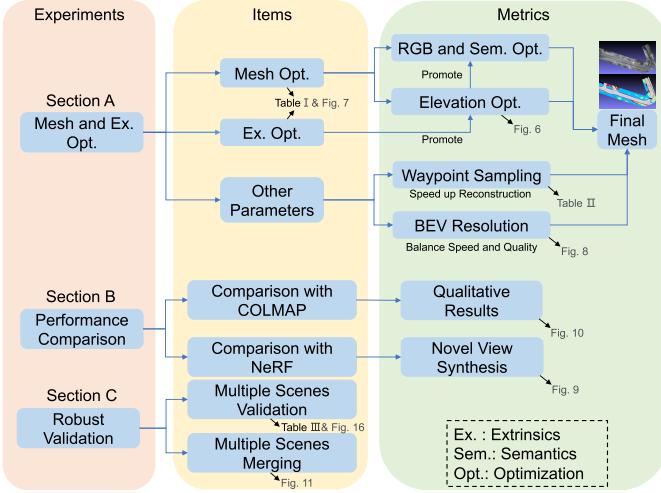


Fig. 5: Workflow of our experiments.

D. Implementation

RoMe initializes a road surface mesh based on [28], utilizing RGBs, semantics, and elevation. Adam optimizer [29] is used during the training. We set the learning rates for RGBs and semantics at 0.1 and for elevation at 0.001. Typically, running the model for seven epochs, with a halving of the learning rate at the 2nd and 4th epochs, yielded satisfactory results for most scenes. We set the BEV resolution at 0.1 meters per pixel. The Elevation MLP, a simple 8-layer network with a width of 128, was adapted from [30].

IV. EXPERIMENTS

In this section, we first introduce the experimental setting, including datasets and metrics. After that, as presented in Fig. 5, we conduct our experiments with three major parts:

- In Section IV-A, we subdivide experiments into mesh optimization, extrinsic optimization, and other parameters that affect final reconstruction results. Mainly, mesh optimization can be divided into RGB and semantics (with learnable parameters) and elevation (with MLP networks) optimization. Extrinsic optimization promotes elevation optimization, followed by RGB and semantics optimization to get final finer reconstruction results. Waypoint sampling speeds up the reconstruction, and BEV resolution can balance the speed and quality.
- In Section IV-B, we compare RoMe with COLMAP on quality and vanilla NeRF on novel view synthesis tasks in a single scene.
- In Section IV-C, we conduct experiments on 100 scenes chosen from nuScenes for multiple-scene validation to show RoMe’s robustness and efficiency in merging multiple scenes to reconstruct larger areas.

Datasets: We conducted our experiments on two renowned driving datasets: nuScenes [18] and KITTI [19]. The nuScenes dataset encompasses 1000 scenes, each being a 20-second video clip annotated at a frequency of 2 Hz. This dataset utilizes a camera rig with six cameras, providing a comprehensive 360-degree field of view. On the other hand, the

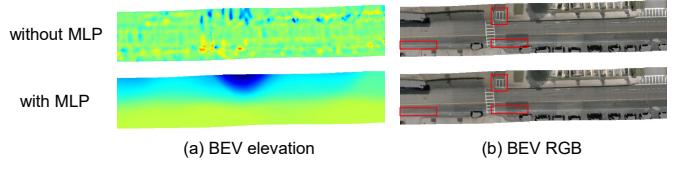


Fig. 6: Ablation study on BEV elevation learning methods. The colormap jet displays BEV elevation ranging from -0.2 meters to 0.2 meters. Utilizing MLP results in smoother elevation, enhancing reconstruction quality (highlighted in red boxes).

KITTI odometry benchmark comprises 22 sequences, split into 11 training sequences (00-10) and 11 test sequences (11-21). For our experiments, we exclusively used monocular images from KITTI’s left RGB camera. For semantic annotations, we employed the predictions from Mask2Former [25] with a Swin-L [31] backbone since it has a state-of-the-art performance on primary semantic segmentation datasets, such as the Mapillary Vistas [27]. For the nuScenes, we randomly selected 100 scenes, and for the KITTI, we chose sequence (00) for evaluation.

Metrics: Our experiments were executed on a Linux server with a single RTX-3090 GPU. We assessed the performance of all methods using standard NVS metrics: PSNR for image quality and mIoU for semantic segmentation accuracy. For 3D structure evaluation, we adopted the point cloud chamfer distance (CD) metric from [20]. It involves converting depth rendered from meshes and LiDAR depth into world coordinates to obtain point clouds:

$$CD(\hat{G}, G) = \frac{1}{\hat{G}} \sum_{x \in \hat{G}} \min_{y \in G} \|x - y\|_2^2 + \frac{1}{G} \sum_{y \in G} \min_{x \in \hat{G}} \|y - x\|_2^2 \quad (7)$$

where \hat{G} and G denote point clouds rendered from meshes and LiDAR depth respectively. Besides, x and y denote 3D points in corresponding point clouds. We restricted our evaluation to points with semantic classes that are expected to be flat. To filter out outlier observations in LiDAR points, we computed the closest 97% points in chamfer distance, following the approach in [20]. Additionally, we utilized the RMSE metric to gauge the discrepancy between LiDAR depth and depth rendered from meshes.

A. Mesh and Extrinsic Optimization

1) *Mesh Optimization:* Mesh optimization is composed of RGB, semantics, and elevation optimizations. RGB and semantics optimization use the presentation of learnable parameters due to their high-frequency details. In terms of elevation optimization, it should be smooth in most cases, so we initiate our experiments by exploring two methods for BEV elevation representation. The first method treats BEV elevation as independent optimizable parameters, similar to RGB and semantics. The alternative one utilizes an MLP representation. As depicted in Fig. 6, the latter approach yields superior results. Through our experiments, we observed that setting the position encoding frequency to 5 was adequate for most scenes.

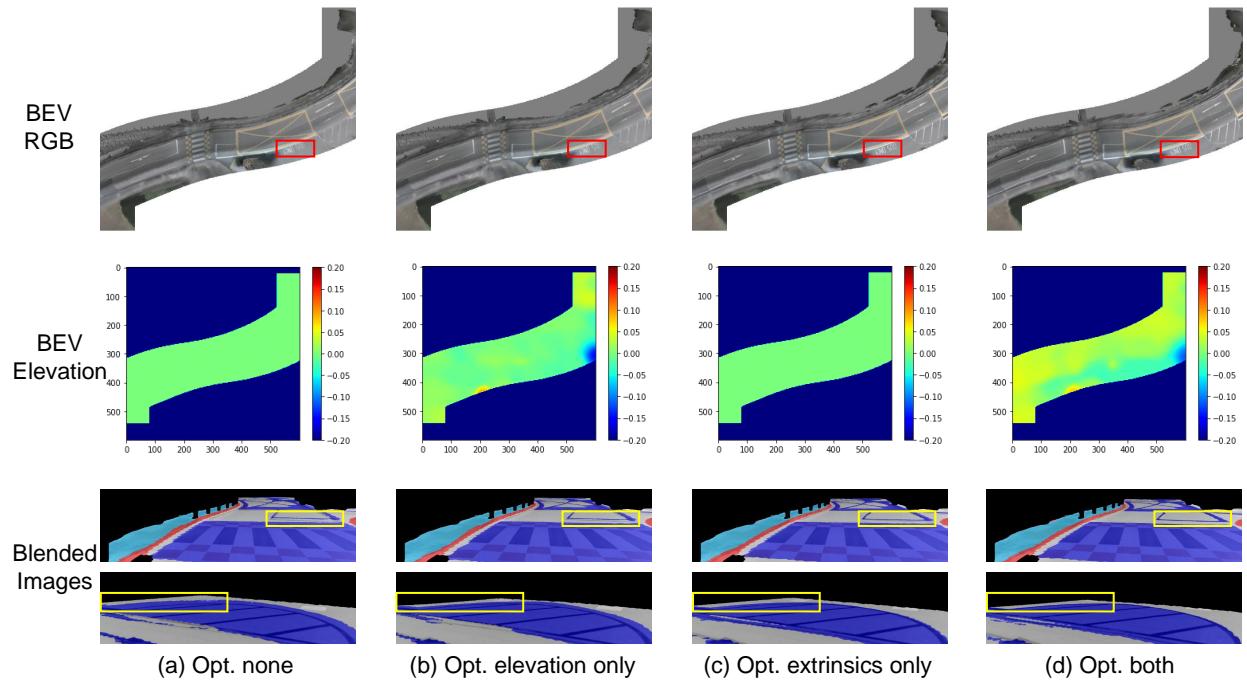


Fig. 7: Ablation study results for elevation and extrinsic. The top panel shows BEV RGB, the middle displays BEV Elevation, and the bottom presents Blended Images. Comprehensive results can be found in Table I. Enhanced elevation restoration and extrinsic optimization lead to improved alignment of RGB and semantics.

TABLE I: Ablation study on optimizing elevation and extrinsic. The best reconstruction results for both textures and semantics are achieved when both are optimized (highlighted in bold).

Opt. elevation	Opt. extrinsic	PSNR↑	mIoU (%)↑
×	×	25.5	71.4
✓	✗	25.9	73.9
✗	✓	26.0	79.0
✓	✓	26.7	83.0

2) *Extrinsic Optimization*: RoMe can restore road surface elevation and refine camera extrinsic, leading to a more precise reconstruction. For our ablation study, we select a short clip from the *scene-0865* of the nuScenes dataset. As detailed in Table I, implementing either elevation restoration or extrinsic optimization enhances the reconstruction results. Moreover, we observed that segmentation results were more sensitive to inaccuracies in extrinsic. For a more visual understanding, some results are illustrated in Fig. 7. The top row displays the BEV RGB. Without applying elevation estimation or extrinsic optimization, the results appear blurry, especially in areas highlighted by red boxes. The middle row visualizes the BEV elevation (in meters). Notably, the BEV elevation in Fig. 7 (d) exhibits more fluctuations than the others. The bottom row showcases blended images of the rendered semantics and the original images. A closer look at the yellow boxes reveals that without optimizing elevation and extrinsic, the rendered semantics do not align accurately with the source images.

3) *Other Parameters*: To assess the efficiency of our proposed waypoint sampling method, we constructed an area spanning 200 * 200 square meters from the KITTI odometry sequence-00. With waypoint sampling, we achieved a 2x speedup and reduced GPU memory consumption, all while

TABLE II: Waypoint sampling efficiency. Utilizing waypoint sampling, we achieve a 2x speed-up and reduced GPU resource consumption without compromising the results.

Waypoint Sampling	Time(min)	GPU(GB)	PSNR↑	mIoU(%)↑
✗	7	15	23.4	86.7
✓	3.5	11	23.4	86.7

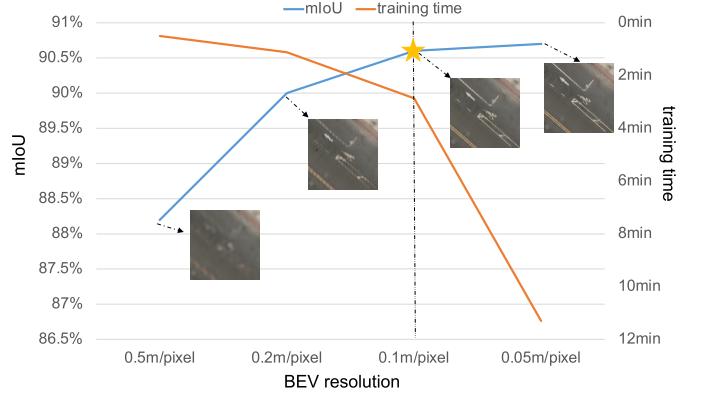


Fig. 8: Ablation study on BEV resolution. A resolution of 0.1m/pixel achieves realistic reconstruction with improved training speed.

maintaining the same reconstruction quality, as detailed in Table II. Additionally, we reconstructed the entire area using poses derived from ORB-SLAM2, as visualized in Fig. 1. This reconstruction of the entire road surface (covering 600 * 600 square meters) was completed in just two hours.

To strike a balance between training speed and reconstruction quality, we conducted experiments on BEV resolution using the *scene-0391* from the nuScenes dataset. The results are presented in Fig. 8. A BEV resolution greater than or equal to 0.2m/pixel led to blurry reconstructions.

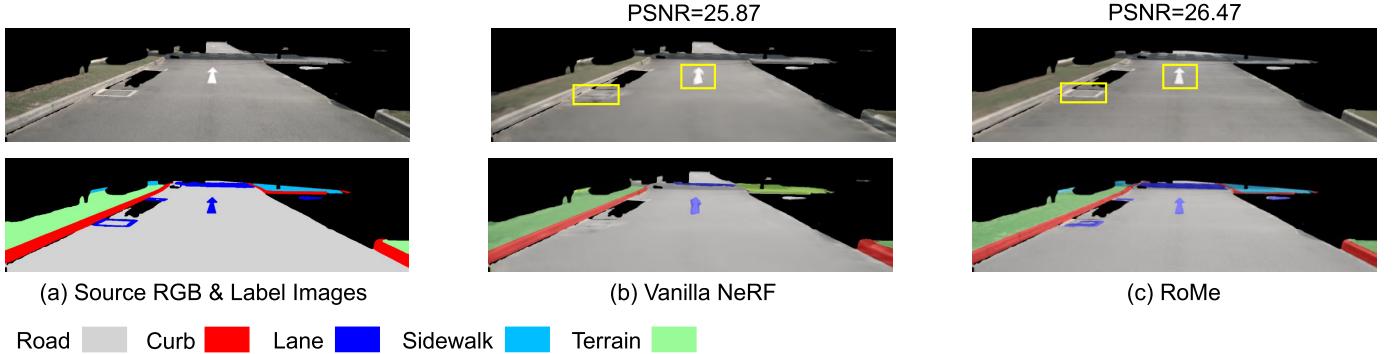


Fig. 9: RGB and semantic reconstruction comparison. A segment from the nuScenes dataset is chosen, with three frames set aside for testing. The rest serve as training data. The yellow boxes highlight that RoMe captures finer details than the standard NeRF [32].

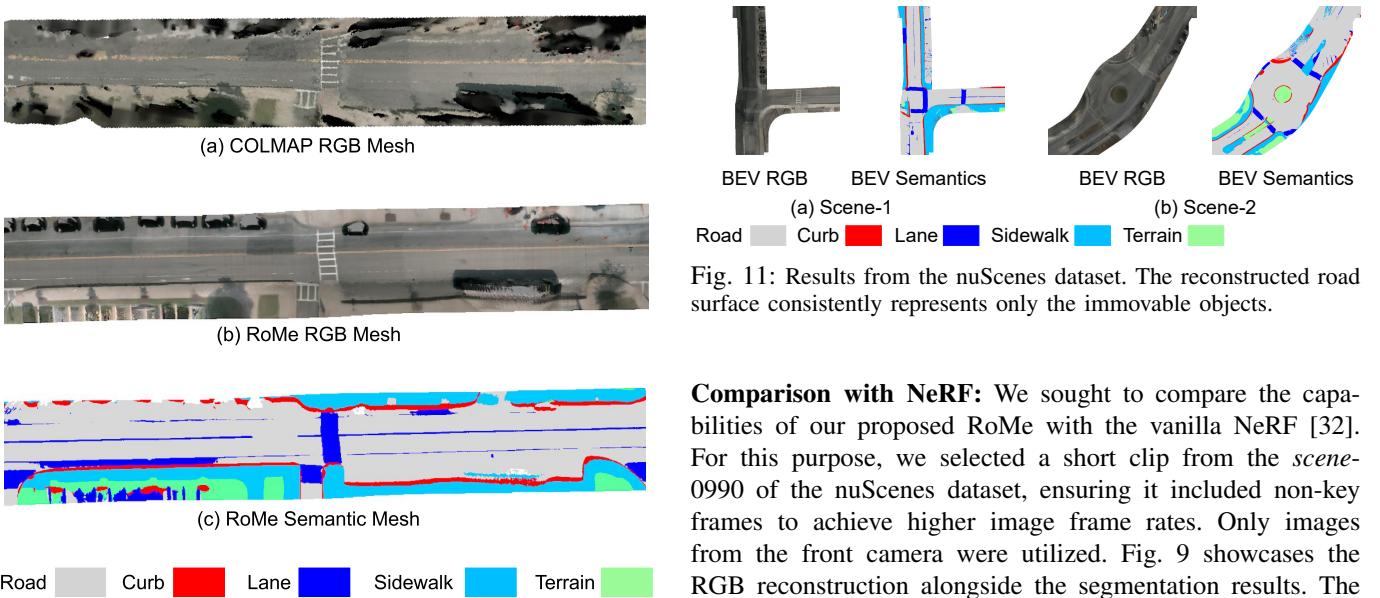


Fig. 10: Comparison with COLMAP. While COLMAP may produce holes in the presence of moving objects, RoME remains robust, reconstructing from unobstructed frames. Additionally, RoME simultaneously reconstructs semantics.

Conversely, resolutions less than or equal to 0.05m/pixel added unnecessary computational overhead. Thus, a resolution of 0.1m/pixel (highlighted with a star) provided the optimal trade-off between quality and speed.

B. Performance Comparision

Comparison with COLMAP: RoMe’s robustness to moving objects surpasses that of COLMAP [1], as illustrated in Fig. 10. We selected the *scene-0655* from the nuScenes dataset and masked all mobile obstacles. The BEV mesh generated by COLMAP (with the Poisson mesher) tends to produce holes when encountering moving objects. In contrast, RoMe consistently generates a complete road mesh, provided there’s at least one frame with a clear view of the road surface. Additionally, RoMe can simultaneously produce BEV semantics.

Comparison with NeRF: We sought to compare the capabilities of our proposed RoMe with the vanilla NeRF [32]. For this purpose, we selected a short clip from the *scene-0990* of the nuScenes dataset, ensuring it included non-key frames to achieve higher image frame rates. Only images from the front camera were utilized. Fig. 9 showcases the RGB reconstruction alongside the segmentation results. The first column displays the source RGB and label images. The second column presents RGB images reconstructed by the vanilla NeRF and semantics segmented by Mask2Former [25]. The third column features RGB images and semantics reconstructed using RoMe. Our method delivers more realistic RGB reconstructions and precise semantic results. The road elements, highlighted in the yellow boxes, are more distinct than those in the vanilla NeRF. In a region spanning 70×70 square meters, our method converged in approximately 8 minutes, whereas the vanilla NeRF required 20 hours. The original NeRF, due to its design, needs to restore depth across a broad range (e.g., $0 \sim 100$ meters) without depth supervision. In contrast, RoMe focuses on restoring elevations of less than 1 meter, which is more straightforward to optimize. The mesh representation inherently captures road surface features, which are predominantly flat but can exhibit significant changes at boundaries like curbs and slope edges.

C. Robustness Validation

Multiple Scenes Validation: We assessed the robustness of RoMe by conducting experiments on 100 scenes selected from the NuScenes dataset. Specifically, we chose scenes characterized by favorable daytime weather conditions and trajectories

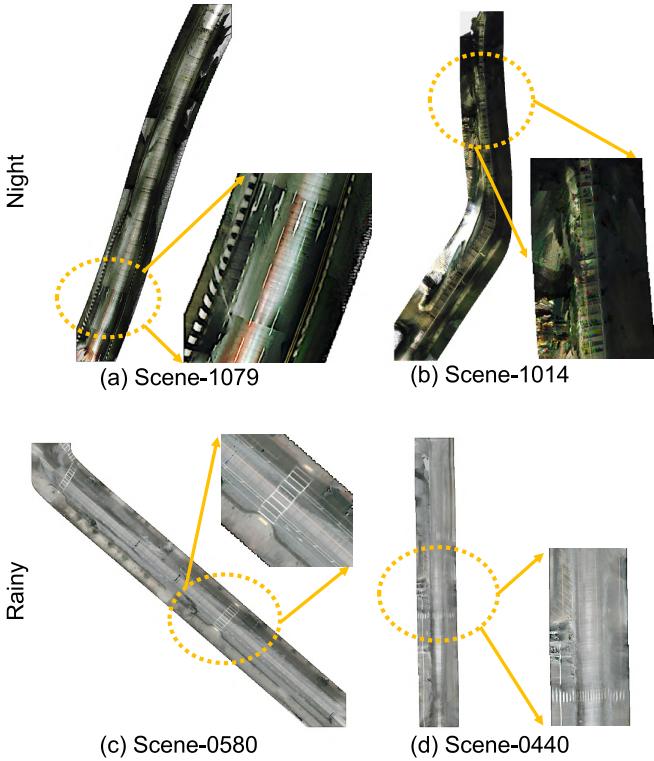


Fig. 12: Road surface reconstructions with RoMe during nighttime and rainy conditions. Exposures are slightly adjusted for a better view.

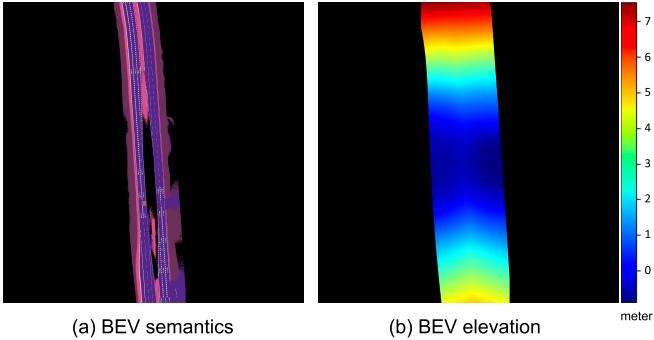


Fig. 13: Visualization of the reconstructed steep slope. The left side depicts BEV semantics, while the right side illustrates BEV elevation, ranging from -0.8 meters to over 7 meters. RoMe’s capability to accurately reconstruct such varied elevations is evident.

spanning more than 100 meters. The experiments on these 100 scenes mirrored the approach described in Table I. We utilized all images for reconstruction and evaluated every image within each scene. Additionally, we fine-tuned the learning rate and adjusted the rotation and translation ranges for optimizing extrinsic. The results are summarized in Table III.

It's worth noting that while a high PSNR indicates good reconstruction quality for RGB images, it doesn't directly reflect the accuracy of the 3D structure reconstruction. This discrepancy arises because networks might overfit to RGB images rather than learning an accurate 3D structure, especially when there's insufficient posed image data. This phenomenon is evident in the third row of Table III. By using a pose

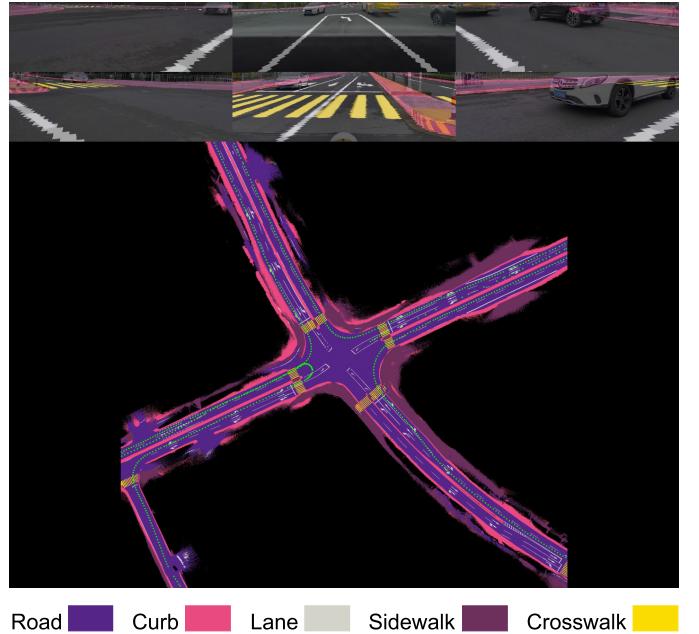


Fig. 14: Reconstruction visualization on wild data. The top two rows blend source RGB images with rendered semantics from the reconstructed road mesh. The bottom displays the reconstructed BEV semantics over a 300×300 square meter area, with green dots indicating trajectories. The alignment between rendered semantics and source RGB images is precise, thanks to accurate poses, BEV semantics, and elevation reconstruction.

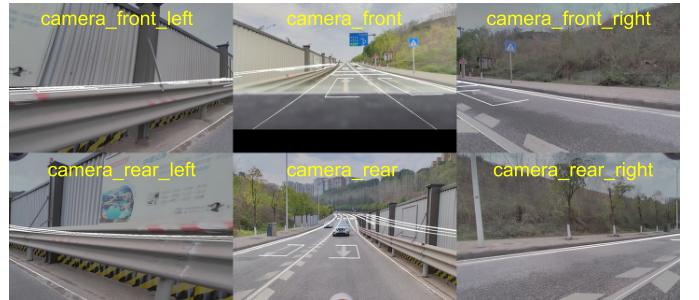


Fig. 15: Reprojection visualization on the reconstructed steep slope. Manually labeled lanes and arrows align seamlessly with road signs and lanes in the source images, validating the accuracy of our 3D road surface reconstruction.

network to refine extrinsic and setting a larger learning rate and rotation/translation ranges, we achieve a higher PSNR but at the cost of an inaccurate 3D structure. This misalignment also affects the mIoU metric, as an incorrect 3D structure disrupts the alignment between source images and rendered semantics.

Fig. 16 offers a qualitative comparison between RoMe meshes and LiDAR point clouds. For clarity, we visualized only the road points filtered by Mask2Former [25]. The RoMe meshes appear smooth and detailed, and they also provide semantic information, simplifying the labeling process.

Multiple Scenes Merging: RoMe can seamlessly integrate different scenes as long as they share common positions. Fig. 11 showcases the results of merging multiple scenes. Both *Scene-1* and *Scene-2* are composites of four individual

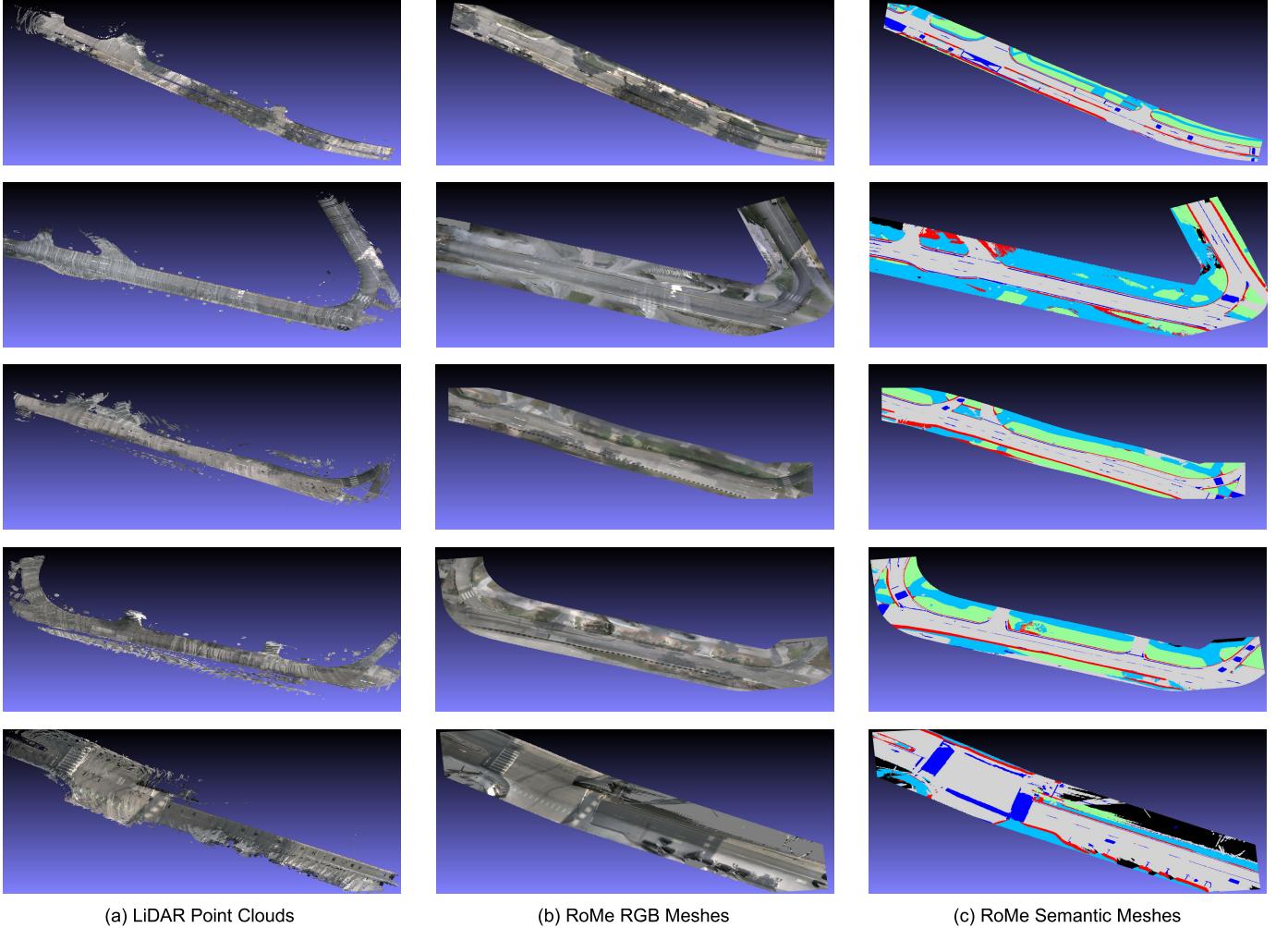


Fig. 16: Visualization of LiDAR point clouds and RoMe meshes.

TABLE III: Multiple scenes ablation study. The table details the impact of optimizing elevation and extrinsic on reconstruction quality across various metrics. The unit of rotation is in degrees and translation in meters. A larger extrinsic learning rate and rotation/translation ranges can lead to image overfitting, resulting in a higher PSNR but an inaccurate 3D structure.

Opt. elevation	Opt. extrinsic	extrinsic LR	rotation	translation	PSNR ↑	mIoU (%) ↑	CD ↓	RMSE ↓
×	×	—	—	—	23.90	67.00	1.25	1.09
✓	×	—	—	—	23.93	67.16	1.24	1.09
✓	✓	0.01	0.5	0.5	24.36	64.86	5.77	1.36
✓	✓	0.002	0.1	0.1	24.19	69.23	1.13	1.03

scenes. The smooth transitions between scenes are attributed to the precise camera poses and the ability to optimize the extrinsic. However, when there are significant differences in weather conditions, we prioritize semantics reconstruction, as semantics remain consistent across varying lighting conditions, unlike RGB images.

D. Limitations and Applications

Limitations: RoMe can reconstruct road surfaces on rainy days or at night, as shown in Fig. 12: (a) and (c) are promising since light conditions are tolerable. When encountering worse light conditions, it is challenging to reconstruct the road surface, as shown in (b) and (d). While RoMe demonstrates the capability to reconstruct road surfaces under various

conditions, including rainy days or nighttime scenarios, its performance can vary based on the severity of environmental conditions. However, in more adverse lighting conditions, the reconstruction quality can degrade, as evident. This limitation underscores the need for further enhancements to RoMe’s adaptability in diverse and challenging scenarios.

Applications: We applied RoMe to wild data, showcasing its versatility. Fig. 14 presents our reconstruction of an intersection. The alignment between the rendered semantics and the source RGB images is evident, demonstrating the precision of our method. This precision facilitates easy annotation of BEV lanes, curbs, arrows, crosswalks, and other static road elements directly on the road mesh.

To further illustrate the strength of learning BEV elevation, we selected a scene in Chong Qing characterized by a steep

slope. In addition to the method above, we utilized structure from motion (SfM) or MVS points generated by COLMAP for precise supervision. Fig. 13 displays our reconstruction. The left side represents BEV semantics, while the right side showcases BEV elevation, which varies from -0.8 meters to over 7 meters. Despite the significant elevation changes, RoMe provides a clear and accurate reconstruction. This accuracy is further demonstrated in Fig. 15, where manually labeled lanes and arrows align perfectly with road signs and lanes, even over an elevation range exceeding 8 meters.

V. CONCLUSION

Throughout this study, we have delved into the intricacies of road surface reconstruction, introducing RoMe as a ground-breaking solution tailored for expansive environments. RoMe stands out due to its unique approach, leveraging a mesh representation that ensures a robust reconstruction of road surfaces, seamlessly aligning with semantic data. This alignment is pivotal, especially when considering the challenges of large-scale reconstructions, where even minor misalignments can lead to significant inaccuracies.

Our evaluations, spanning areas as vast as 600 * 600 square meters and encompassing renowned datasets like nuScenes and KITTI, have consistently showcased RoMe's superiority in terms of accuracy, speed, and resilience, particularly when compared to existing methods like the vanilla NeRF. The waypoint sampling strategy, a hallmark of RoMe, not only accelerates the training process but also optimizes computational resources. By reconstructing road surfaces in segmented regions and then integrating them during training, RoMe demonstrates its adaptability to large-scale environments without compromising on precision.

Moreover, introducing the extrinsic optimization module addresses a critical challenge in road surface reconstruction: the potential inaccuracies stemming from extrinsic calibration. This module, combined with RoMe's inherent design, ensures that the framework remains robust even in diverse and challenging scenarios, as evidenced by our experiments on both public and wild data.

In the context of autonomous driving, precision is of utmost importance. RoMe emerges as a transformative solution. Its ability to provide accurate reconstructions paves the way for automating the labeling process, a crucial step toward the realization of fully autonomous vehicles. As we move forward, the innovations presented in this study underscore the potential of RoMe to revolutionize road surface reconstruction and its broader applications in autonomous driving.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant Number: 22KJB520008) and the Research Fund of Horizon Robotics (Grand Number: H230666).

REFERENCES

- [1] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixel-wise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.
- [3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.
- [4] K. Rematas, A. Liu, P. P. Srinivasan, J. T. Barron, A. Tagliasacchi, T. Funkhouser, and V. Ferrari, "Urban radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 12 932–12 942.
- [5] Z. Li, L. Li, Z. Ma, P. Zhang, J. Chen, and J. Zhu, "Read: Large-scale neural scene rendering for autonomous driving," *arXiv preprint arXiv:2205.05509*, 2022.
- [6] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8248–8258.
- [7] P. Musialski, P. Wonka, D. Aliaga, M. Wimmer, L. Van Gool, and W. Purgathofer, "A survey of urban reconstruction," *Computer Graphics Forum*, vol. 32, pp. 146–177, 09 2013.
- [8] X. Wang, C. Wang, B. Liu, X. Zhou, L. Zhang, J. Zheng, and X. Bai, "Multi-view stereo in the deep learning era: A comprehensive review," *Displays*, vol. 70, p. 102102, 10 2021.
- [9] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "Nerf: Neural radiance field in 3d vision, a comprehensive review," 2023.
- [10] O. Özışıl, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion*," *Acta Numerica*, vol. 26, pp. 305–364, 2017.
- [11] R. Fan, X. Ai, and N. Dahouni, "Road surface 3d reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.
- [12] S.-J. Yu, S. R. Sukumar, A. F. Koschan, D. L. Page, and M. A. Abidi, "3d reconstruction of road surfaces using an integrated multi-sensor approach," *Optics and lasers in engineering*, vol. 45, no. 7, pp. 808–818, 2007.
- [13] H. Brunken and C. Güthmann, "Road surface reconstruction by stereo vision," *PFG-Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 88, no. 6, pp. 433–448, 2020.
- [14] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.
- [15] T. Qin, Y. Zheng, T. Chen, Y. Chen, and Q. Su, "A light-weight semantic map for visual localization towards autonomous driving," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 248–11 254.
- [16] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 31–42.
- [17] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color! large-scale texturing of 3d reconstructions," in *European conference on computer vision*. Springer, 2014, pp. 836–850.
- [18] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Lioung, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.
- [20] J. Guo, N. Deng, X. Li, Y. Bai, B. Shi, C. Wang, C. Ding, D. Wang, and Y. Li, "Streetsurf: Extending multi-view implicit surface reconstruction to street views," 2023.
- [21] Z. Xie, Z. Pang, and Y.-X. Wang, "Mv-map: Offboard hd-map generation with multi-view consistency," *arXiv*, 2023.
- [22] F. Wang, A. Louys, N. Piasco, M. Bennehar, L. Roldão, and D. Tsishkou, "Planerf: Svd unsupervised 3d plane regularization for nerf large-scale scene reconstruction," 2023.
- [23] W. Bian, Z. Wang, K. Li, J.-W. Bian, and V. A. Prisacariu, "Nope-nerf: Optimising neural radiance field with no pose prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4160–4169.
- [24] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

- [25] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.
- [26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [27] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.
- [28] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, “Accelerating 3d deep learning with pytorch3d,” *arXiv preprint arXiv:2007.08501*, 2020.
- [29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, 2014.
- [30] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, “Nerf+: Neural radiance fields without known camera parameters,” *arXiv preprint arXiv:2102.07064*, 2021.
- [31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
- [32] C. Quei-An, “Nerf-pl: a pytorch-lightning implementation of nerf,” 2020. [Online]. Available: https://github.com/kwea123/nerf_pl/



Ruohong Mei is currently an Algorithm Engineer at Horizon Robotics in Beijing, China. He earned his B.S. Degree in Communication Engineering from Beijing University of Posts and Telecommunications in 2018, followed by an M.S. Degree in Information and Communication Engineering from Beijing University of Posts and Telecommunications in 2021. His primary research interests lie in 3D vision, and deep learning.



Wei Sui is a senior engineer at Horizon Robotics, leading the 3D Vision Team, providing mapping, localization, calibration, and 4D labeling solutions. His research interests include SFM, SLAM, Nerf, 3D Perception, etc. Dr. Wei received his B.Eng and Ph.D. degrees from Beihang University and NLPR (CASIA), Beijing, China, in 2011 and 2016 respectively. He led the computer vision team and successfully developed the 4D Labeling System and BEV perception for Super Drive on Journey 5. Dr. Wei Sui has published one research monograph and more than ten peer-reviewed papers in journals and conference proceedings, including elites like TIP, TVCG, ICRA, CVPR, etc. Dr. Wei received over 40 Chinese and 5 US patents.



Jiaxin Zhang is currently an Algorithm Engineer at Horizon Robotics in Beijing, China. He earned his B.S. degree in Applied Physics from the University of Science and Technology of China in 2018, followed by an M.S. degree in Electrical and Computer Engineering from Boston University in 2020. His primary research interests lie in SLAM, 3D vision, and deep learning.



Xue Qin, a Senior Engineer affiliated with Harbin Institute of Technology, boasts an extensive academic background in the realm of computer science and technology. He commenced his academic journey with a Bachelor's degree in Computer Science from the University of Melbourne, followed by a Master's degree in Network Computing from Monash University. Presently, Mr. Qin is deepening his research endeavors as a Ph.D. candidate in Computer Science and Technology at Harbin Institute of Technology. Beyond his technical pursuits, he has

also broadened his managerial acumen by completing an Executive Master of Business Administration from the prestigious Tsinghua University. His scholarly contributions predominantly revolve around artificial intelligence, computer vision, and pioneering anti-collision systems for autonomous vehicles.



Gang Wang, master degree in Vehicle engineering from Wuhan University of Technology, PhD in intelligent manufacturing from Shandong University. In 2015, he began to work as a senior professional manager and senior engineer in automobile manufacturing enterprises. His main research direction is the application and industrialization of intelligent manufacturing technology, 3D vision technology and SLAM laser navigation technology in the direction of unmanned logistics. He has published many high-level papers and software Copyrights in the direction of industrial big data and manufacturing digital transformation.



Tao Peng is an Associate Professor in Soochow University, China, since 2022. Before that, Dr. Peng received his Ph.D. degree in the Department of Computer Science and Technology at Soochow University in 2019. From 2020 to 2022, he was a postdoctoral researcher in the Department of Health Technology and Informatics at Hong Kong Polytechnic University, and Department of Radiation Oncology at University of Texas Southwestern Medical Center, Dallas, USA, successively. During this period, he obtained the “Research Talent” award

from Hong Kong government. He has published more than 40 peer-reviewed journal/conference papers, where the total impact factor (IF) of all the journal publications as the first author is IF \approx 98. He now serves as Guest Associate Editor of Medical Physics journal, a Co-Editor of Special Topic at Frontiers in Signal Processing journal, Program Committee of 20(th) PRICAI-2023 and iWOAR 2023 conference, and a reviewer of more than 20 high-quality journals/conferences. His main research interests include image processing, pattern recognition, machine learning, and their applications.



Cong Yang is an Associate Professor at Soochow University since 2022. Before that, he was a Postdoc researcher at the MAGRIT team in INRIA (France). Later, he worked scientifically and led the computer vision and machine learning teams in Clobotics and Horizon Robotics. His main research interests are computer vision, pattern recognition, and their interdisciplinary applications. Cong earned his Ph.D. degree in computer vision and pattern recognition from the University of Siegen (Germany) in 2016.