EMIE-MAP: Large-Scale Road Surface Reconstruction Based on Explicit Mesh and Implicit Encoding

Wenhua Wu[1], Qi Wang[2], Guangming Wang[3], Junping Wang[4], Tiankun Zhao[4], Yang Liu[4], Dongchao Gao[4], Zhe Liu[1]*, and Hesheng Wang[2]*

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] Department of Automation, Shanghai Jiao Tong University
[3] University of Cambridge
[4] Hozon New Energy Automobile Co., Ltd
{2142431167wwh, liuzhesjtu, wanghesheng}@sjtu.edu.cn

Abstract. Road surface reconstruction plays a vital role in autonomous driving systems, enabling road lane perception and high-precision mapping. Recently, neural implicit encoding has achieved remarkable results in scene representation, particularly in the realistic rendering of scene textures. However, it faces challenges in directly representing geometric information for large-scale scenes. To address this, we propose EMIE-MAP, a novel method for large-scale road surface reconstruction based on explicit mesh and implicit encoding. The road geometry is represented using explicit mesh, where each vertex stores implicit encoding representing the color and semantic information. To overcome the difficulty in optimizing road elevation, we introduce a trajectory-based elevation initialization and an elevation residual learning method. Additionally, by employing shared implicit encoding and multi-camera color decoding, we achieve separate modeling of scene physical properties and camera characteristics, allowing surround-view reconstruction compatible with different camera models. Our method achieves remarkable road surface reconstruction performance in open source datasets and a variety of real-world challenging scenarios.

Keywords: Road Surface Reconstruction · Surround View · Explicit Mesh · Implicit Encoding

1 Introduction

The development of autonomous driving systems has brought about a growing need for accurate road reconstruction, as it plays an essential part in enabling perception and high-precision mapping [13, 19, 20, 23, 44, 46]. In recent years, Bird's Eye View (BEV) perception has gained prominence in the field of autonomous driving due to its natural alignment with downstream tasks such as

* Corresponding author.

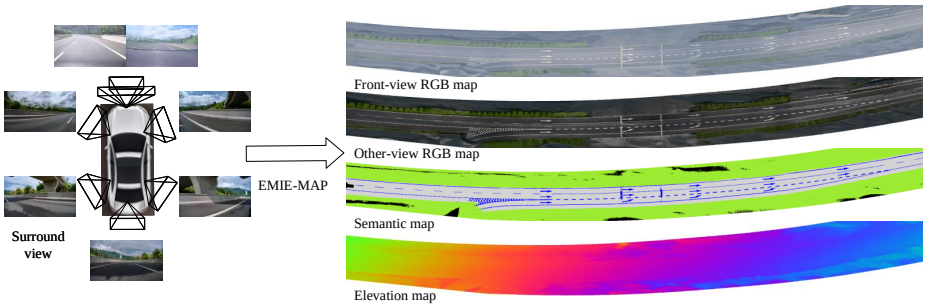

Fig. 1: We propose EMIE-MAP, a novel large-scale road surface reconstruction method based on **E**xplicit **M**esh and **I**mplicit **E**ncoding. By taking input surround-view video and localization information, EMIE-MAP is capable of reconstructing RGB maps, semantic maps, and elevation maps. The RGB maps are corresponding to different cameras. EMIE-MAP integrates the advantages of explicit and implicit representations, enabling accurate road surface reconstruction and rendering.

planning and control [47]. As a result, the importance of large-scale road reconstruction has been highlighted, especially its potential application in providing training and validation data for perception tasks of autonomous vehicles.

Previously, 3D reconstruction methods can be divided into traditional methods [28] and Neural Radiance Fields (NeRF)-based methods [7, 9, 26]. Although the traditional multi-view stereo methods can effectively reconstruct dense or semi-dense points, they often produce incomplete or noisy results when applied to untextured road. NeRF-based approaches show promise in photorealistic reconstruction. However, they faces challenges in directly representing geometric information and are limited by high resource consumption, especially in large-scale scenarios.

For large-scale road surface reconstruction, the work most similar to this paper is RoMe [23]. RoMe decomposes the 3D road into a triangular mesh and models the road elevation using a multi-layer perception (MLP). Each mesh vertex stores explicit color and semantic information to capture fine surface details. However, it faces difficulties in accomplishing road surface reconstruction for slope scenarios. Luminosity inconsistencies between surround views can cause convergence instability of RoMe. To address these challenges, we propose a large-scale road reconstruction method (EMIE-MAP) based on explicit grid and implicit encoding.

EMIE-MAP utilizes a combination of explicit mesh and implicit encoding. An explicit mesh representation is used to capture the road geometry, and each mesh vertex stores an implicit encoding that encapsulates color and explicit semantic information. To meet the challenge of road elevation optimization, we propose a trajectory-based elevation initialization method supplemented by an elevation residual learning method based on MLP. This combination allows us to achieve robust and accurate road reconstruction, especially in the case of large changes in road elevation. In addition, we combine shared implicit encoding and

multi-camera RGB implicit decoding, enabling independent modeling of scene physical properties and camera characteristics. This facilitates surround-view reconstruction compatible with a wide range of camera types, enhancing the versatility and applicability of our approach in a variety of autonomous driving systems.

The main contributions of our work can be summarized as follows:

- We propose EMIE-MAP, a large-scale road surface reconstruction method. The core of EMIE-MAP is a road surface representation based on explicit mesh and implicit encoding, facilitating the storage and optimization of road geometry, color, and semantic information.
- We introduce a trajectory-based elevation initialization and an MLP-based elevation residual prediction method to learn ground elevation, overcoming the difficulties in reconstructing road surfaces on slopes.
- We propose a color optimization method based on shared implicit color encoding and multi-camera color decoding, enabling the separation of scene physical attributes and camera characteristics. This allows the method to be applied to surround-view with different camera models.
- Experiments conducted in open source datasets and various real challenging scenes demonstrate that EMIE-MAP exhibits remarkable reconstruction results.

2 Related work

2.1 Explicit 3D Reconstruction

Explicit 3D reconstruction methods directly reconstruct 3D objects into point clouds or meshes. There has been a lot of accumulation of these algorithms [28]. Incremental methods of [27, 34, 35, 42] are a series of classical Structure from Motion (SfM) methods. COLMAP [30, 31] extracts feature points from image data to reconstruct 3D scenes, but it does not perform well in weak texture areas such as road surface, and the reconstructed points are too sparse. Sameer Agarwal *et al.* [1] use a large number of images on the Internet to complete the reconstruction of urban scale, but this method also faces the problems of sparse results and sensitivity to texture. Algorithms that directly target road reconstruction can produce dense outputs [3, 11, 12, 15, 45], but they are limited to the reconstruction of small road areas. Methods for representing a 3D scene in explicit meshes can automatically texture the model in large-scale 3D reconstruction [17, 40]. They can improve the visual effect of the reconstructed model and provide a new direction for 3D reconstruction.

2.2 Implicit 3D Reconstruction

NeRF [26] is one of the most fundamental works to pioneer implicit 3D reconstruction. It uses deep learning methods to learn an implicit scene representation from existing images from different perspectives, and can render the

scene to generate a simulated picture from a new perspective. The utilization of additional point cloud inputs can improve the effectiveness of NeRF [18, 29]. Block-NeRF [36] divides large scenes into blocks and trains the NeRF network separately. These networks are trained in parallel and joined together for inference. SUDS [39] applies NeRF to scalable urban dynamic scenarios [21, 22]. GM-NeRF [4] learns generalized model-based neural radiation fields from multi-view images. Andreas Meuleman *et al.* [25] propose asymptotically optimized local radiation fields for robust view synthesis. ABLE-NeRF [37] proposes a volume framework based on self-attention mechanism and introduces learnable embedded features to capture perspective-dependent effects in scenes. SPARF [38] solves the problem of sparse view and inaccurate pose in neural radiation field. Implicit scene representation is also widely used in simultaneous localization and mapping (SLAM) [6, 8, 10, 32, 43, 49]. Recently, there are multiple works using diffusion reconstruct an implicit 3D surface in high fidelity from a cloud of noise points [2, 24, 41, 48].

Explicit 3D reconstruction can directly represent the geometric information of a 3D scene, while implicit representation methods such as NeRF can model rich texture details. Combining both approaches can achieve better scene reconstruction [33]. RoMe [23] models the color and semantic information of the road surface into explicit meshes and uses MLP to model elevation. RoMe directly optimizes the explicit color, which makes it impossible to determine whether the scene is optimized in a brighter or darker direction when each camera's luminosity is inconsistent. There are some methods [29, 31, 40] proposing solutions for luminosity inconsistency, but they focus on the luminosity changes of the same camera at different times. Those methods ignore the inherent luminosity differences among various surround-view cameras carried by the autonomous vehicle. On the other hand, vehicle trajectory information can provide good prior for ground elevation prediction, and using MLP to predict elevation residual results can get more accurate elevation results. EMIE-MAP improves on both.

3 Method

The overview of the proposed EMIE-MAP is shown in Fig. 2. The left side presents a road surface representation based on explicit mesh and implicit encoding. We utilize a mesh composed of equilateral triangular faces to represent the road structure. Each vertex stores its initial coordinates (x, y, z_0), semantic information, and implicit color encoding. An elevation residual MLP predicts the elevation residual, while multiple RGB MLPs decode the implicit color features into observed colors for the corresponding camera. This yields a road surface with explicit information in the middle section of the framework. On the right side, optimization losses are employed. Through direct rendering, generated RGB and semantic maps are supervised by observed images. Additionally, Lidar point clouds are utilized to supervise the road surface elevation. In this section, we will provide a detailed description of EMIE-MAP.

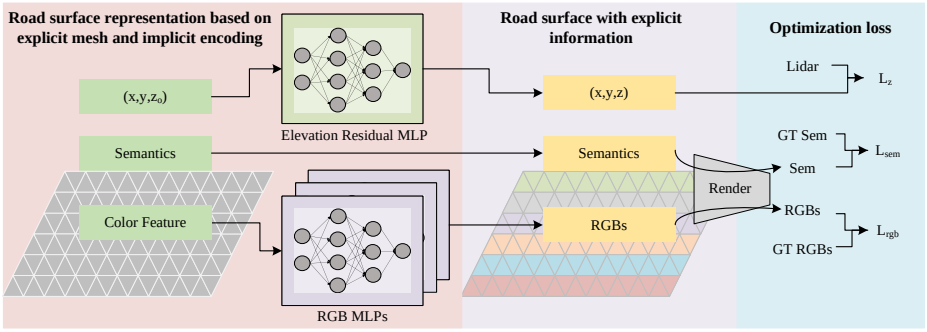

Fig. 2: Overview of EMIE-MAP. The left side presents the proposed road surface representation based on explicit mesh and implicit encoding. We utilize a mesh composed of equilateral triangular faces to represent the road structure. Each vertex stores its initial coordinates (x, y, z_0), semantic information, and implicit color encoding. An elevation residual network predicts the elevation residual at each vertex, while multiple RGB MLPs decode the implicit color features into observed colors for the corresponding camera. This yields a road surface with explicit information in the framework's middle section. The right side shows optimization losses. The generated RGB and semantic maps through direct rendering are supervised by observed images. Additionally, Lidar point clouds are utilized to supervise the road surface elevation.

3.1 Road Surface Representation based on Explicit mesh and Implicit Encoding

Traditional reconstruction algorithms use point clouds, meshes, or voxels for scene representation, directly displaying scene information. However, they face challenges such as holes, difficult optimization, and a lack of flexibility in representation. In contrast, NeRF [26] provide a purely implicit scene representation that uses MLP to model the mapping from spatial coordinates to spatial geometry and color information, resulting in a more flexible scene representation. However, on one hand, NeRF [26] requires substantial computational resources for image rendering through ray sampling integration. On the other hand, while the rendering results are realistic, it is challenging to obtain the explicit geometric information of the scene.

To address these issues, we propose a road surface representation method that combines an explicit mesh and implicit encoding. Specifically, we use a mesh composed of equilateral triangular faces to represent the scene structure. The length of each face is a. Each vertex stores two explicit attributes: position (x, y, z_0) and semantics sem, along with an implicit color encoding l_c. We employ a MLP to estimate elevation residuals, obtaining more accurate position information. Multiple RGB MLPs decode the implicit color features into observed colors for the corresponding camera. Next, we will look in detail at the three road surface attributes of elevation, color, and semantics.

Elevation. RoMe [23] directly uses an MLP to predict elevations, which is effective for relatively flat road surfaces. However, when the road surface has steep

slopes, simple MLPs struggle to predict drastic elevation changes. To address the difficulty, we propose an elevation prediction approach based on trajectory elevation initialization and residual prediction. Unlike RoMe [23], we no longer initialize the road surface as a horizontal plane. Based on the fact that the vehicle runs close to the ground along the lane, we use trajectory elevation to initialize the road surface elevation. Then, we employ an MLP to predict elevation residuals. Instead of directly predicting drastic elevation changes, the MLP only needs to predict low-frequency elevation residuals that are easier to learn. Specifically, the coordinate (x, y) is encoded and input to the residual prediction network MLP_{hr}. The final road surface elevation is the sum of the initial elevation and the residual:

$$z_r = MLP_{hr}(PE(x, y)), z_f = z_0 + z_r, \tag{1}$$

where $PE()$ refers to Positional Encoding, which employs a combination of sine and cosine functions to generate positional encoding vectors.

Color. In order to accommodate different perceptual needs, the surround-view cameras equipped in automobiles vary. Generally, the front view, serving as the primary perspective, utilizes cameras with a wider field of view and higher resolution. Different camera models result in variations in the RGB images when observing the same area, which are manifested in terms of brightness and saturation. In such cases, directly optimizing the explicit RGB values for RoMe [23] strategy leads to optimization conflicts. To address this challenge, we employ implicit color encoding l_c to replace the explicit RGB values. For different camera models, we utilize different decoders for color decoding:

$$rgb_i = MLP_{rgb\_i}(l_c). \tag{2}$$

The decoded colors are supervised against the observed colors from their respective cameras. It should be noted that different camera decoders share the same implicit color encoding. Implicit color encoding represents the physical attributes of the scene, while the decoders learn camera characteristics, thus achieving modeling of inconsistent observations from multiple cameras and ultimately obtaining consistent scene information.

Semantics. Unlike color, semantics are intrinsic properties of the scene and independent of camera. Therefore, we directly utilize explicit parameters to represent semantics.

3.2 Optimizing Strategies

Mesh initialization. By utilizing localization algorithm based on multiple sensors such as Inertial Measurement Unit (IMU), wheel encoders, Global Positioning System (GPS), and cameras, we can acquire high-precision vehicle trajectory information in both the horizontal and vertical directions. Under normal circumstances, the vehicle adheres to the road surface and travels along the lane. Therefore, we can extend the trajectory in the horizontal direction to obtain the road area. The extension distance is determined by the width of the road. After determining the road surface area, we construct a road surface model based on a

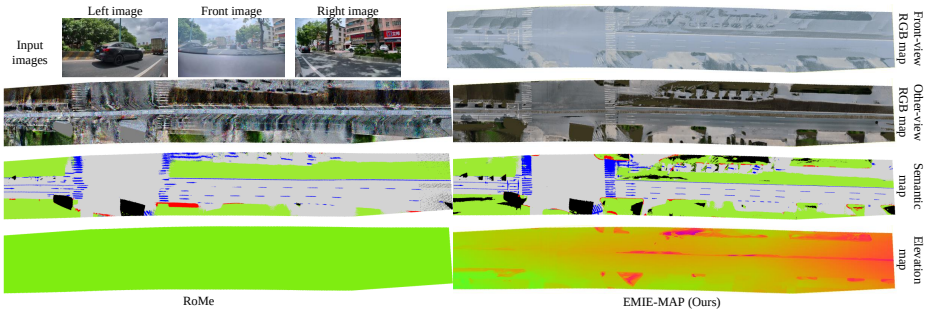

Fig. 3: Road surface reconstruction results in city street scene. From top to bottom are the input images, reconstructed road surface RGB images from different camera perspectives, semantic map, and elevation map. Despite the presence of dense traffic, the road surface in occluded areas is effectively filled in by images taken at different times.

designed representation. The elevation of the road z is initialized through interpolation of nearby trajectory points' elevations z_0. Semantics sem and implicit color encoding l_c are initialized randomly.

Data sampling and observation area sampling. To accelerate the training process, we sample a batch of B images for each training iteration. Since the observed road surface area varies at different positions, we employ a trajectory-based data sampling strategy to improve computational efficiency. The images within each batch are obtained from observations at adjacent trajectory points, resulting in these images having a similar observation area. Subsequently, we extract the road surface within an $80m$ distance before and after the trajectory point as the observation area for the batch of data. By employing this data sampling and observation area sampling strategy, each training iteration only requires explicit processing, rendering, and other operations on a small segment of the road surface, significantly speeding up the reconstruction process.

Road surface visualization. After determining the observation area for each training batch, we need to fully visualize the partially observed road surface. Specifically, a elevation residual network is used to predict the elevation residual at each vertex, which is then added to the initialized elevation to obtain the road surface elevation information. The implicit color encoding is decoded using the corresponding color decoder for each camera, resulting in explicit RGB values for visualization. The semantics are directly used from the stored semantic information. This process ultimately yields a fully visualized road surface, including coordinate, color, and semantic information.

Rendering. Unlike NeRF [26] that samples along the observation rays and integrates information at the sampled points, we employ a direct projection rendering method based on the pinhole camera principle. Given the road surface mesh, camera poses T, and camera intrinsic parameters K, the pixel coordinates $(u, v, 1)$ corresponding to each road surface point (x, y, z) can be obtained using

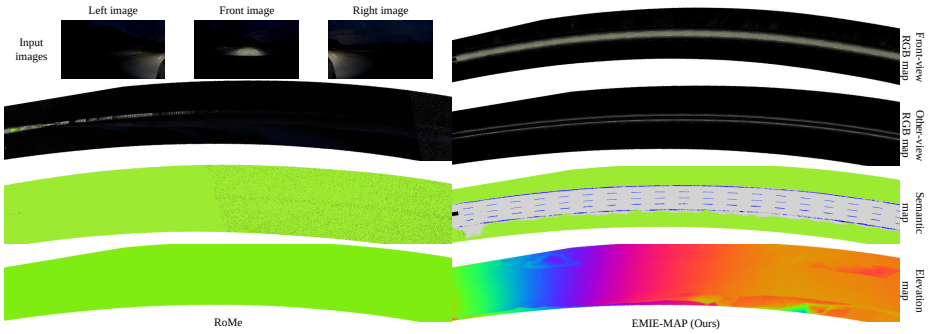

Fig. 4: Road surface reconstruction results in night scene. The only light source is the car headlights. In the input image with poor lighting quality, the RGB results are also dark. However, our method is still able to reconstruct the desired semantic map.

the following formula:

$$(u, v, 1)^T = K \cdot T \cdot (x, y, z, 1)^T. \tag{3}$$

The color and semantic information of the ground points are then projected onto the corresponding pixel, resulting in the rendered RGB image C and semantic image S. It is important to note that for the rendering of RGB images from each camera, we utilize the corresponding decoded RGB values. Compared to volume rendering, our rendering approach better simulates the process of acquiring real images and significantly reduces computational complexity. In reality, the rays emitted from the surface of an object, pass through the camera center and reach the pixel plane. Our rendering process simulates the process of capturing real images.

Training loss For color, we supervise the rendering of RGB images by using observed images from the cameras. For semantics, we utilize a pre-trained Mask2Former [5] for semantic segmentation of the observed images, which serves as the ground truth for segmentation. Then, we calculate the cross-entropy loss between the rendered semantics and the ground truth. Additionally, we construct a semantic-based road surface mask M to filter out irrelevant information such as pedestrians and vehicles. The loss formulas for color and semantics are as follows:

$$L_{rgb} = \frac{1}{|M|} \sum M|C - C_{gt}|, \tag{4}$$

$$L_{sem} = \frac{1}{|M|} \sum M \cdot CE(S, S_{gt}), \tag{5}$$

where C_{gt} represents the RGB ground truth and S_{gt} represents the semantic ground truth. $CE()$ denotes the cross-entropy loss.

To better optimize the road surface elevation, we employ Lidar points for elevation supervision. Specifically, we query Lidar points within a certain neighborhood range of ground points, and their elevations are used as the ground

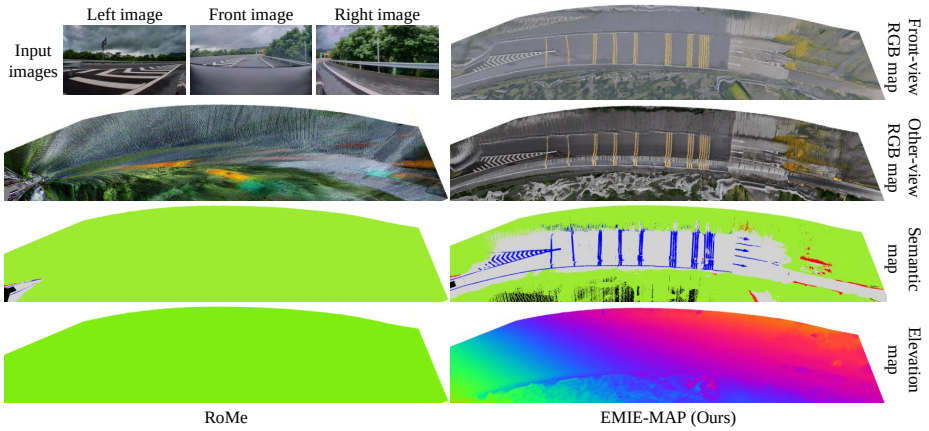

Fig. 5: Road surface reconstruction results in ramp scene. Our method remains remarkable performance in ramp scenes, capturing road surface color, semantics, and elevation.

truth z_{gt} for the road surface elevation. The formula for elevation supervision loss is as follows:

$$L_z = \frac{1}{|M|} \sum M|z - z_{gt}|. \tag{6}$$

Furthermore, based on the smooth of the ground, we add a Laplacian smooth loss:

$$L_{smooth} = \sum_{i=1}^{N} \sum_{j \in N(i)} |z_i - z_j|^2 , \tag{7}$$

where N represents the number of vertices in the mesh, N_i represents the set of vertices adjacent to the vertex i, and z_i and z_j represent the elevation values of the vertex i and j, respectively. For each vertex i, calculate the sum of the squares of its elevation difference from its neighbor j, and then sum all the vertices. This encourages the model to learn to generate smoother mesh elevation predictions.

The total loss function is the combination of the above four losses.

$$L_{total} = \lambda_{rgb} L_{rgb} + \lambda_{sem} L_{sem} + \lambda_z L_z + \lambda_{smooth} L_{smooth}, \tag{8}$$

where, λ_{rgb}, λ_{sem}, λ_z, and λ_{smooth} are the corresponding loss weights.

Parameters optimization and the final outputs. During the reconstruction process, the learnable parameters include the parameters of the elevation residual MLP, multiple RGB MLPs, and semantic and implicit color encoding. Once the training is completed, we predict the elevation of the road surface and decode the optimized implicit color feature using the trained color MLPs, resulting in a visual representation of the road surface that incorporates colors from different cameras and semantics.

4 Experiment

4.1 Experiment Setup

Datasets. We conducted experiments on the KITTI [14] dataset and a challenging custom dataset that includes diverse road surface scenes such as city streets, highways, slopes, tunnels, nighttime environments, and so on. Our data collection vehicle is equipped with a surround-view system consisting of seven cameras. Among them, there are two front-facing cameras, one wide-angle camera, and one telephoto camera, all with a resolution of 3840×2160. The other five cameras have the same model and a resolution of 1920×1080. They are positioned on the front-left, front-right, back-left, back-right, and back of the vehicle. Additionally, there is a 128-line Lidar sensor. The image capture frequency is 30Hz, while the Lidar operates at a frequency of 10Hz. Each data packet is approximately 30 seconds. Our dataset contains 34 data packets, totaling 30,379 frames of surround-view images.

Implementation details. The parameters to be optimized include the parameters of the elevation residual network and color decoders, as well as the semantic and color feature. We use the Adam [16] optimizer to optimize these parameters. The learning rate for the elevation residual MLP is set to 0.01, while the learning rate for the color MLPs is set to 0.005. The semantic learning rate is set to 0.1, and the learning rate for the color feature is set to 0.005. The loss weights are set as $\lambda_{rgb} = 1.0$,$\lambda_{sem} = 1.0$,$\lambda_z = 1.0$, and $\lambda_{smooth} = 1.0$.

The road mesh resolution a is set to 0.1m. The dimension of the color feature is 16. The elevation residual MLP consists of eight layers with width of 128. The color MLPs consist of two layers with width of 16. For each scene, a total of 5 epochs are trained with a batch size of 8. Due to the sufficient coverage of the road surface by the front-facing wide-angle camera and the front-left and front-right cameras, we only utilize images from these three perspectives for the reconstruction. All experiments are conducted on a server equipped with an NVIDIA A100 GPU. For more implementation details, please refer to the supplementary materials.

Metrics. For road surface color and semantics, we project the reconstructed road surface onto the perspective of each camera to obtain rendered images. We evaluate the results using the Peak Signal-to-Noise Ratio (PSNR) for color fidelity and mean Intersection over Union (mIoU) for semantic segmentation accuracy. For road surface elevation, we evaluate the performance by calculating the average distance between the Lidar ground points and the reconstructed road surface, defined as Elev-error.

Baseline. We compare our method with RoMe [23], which is a road surface reconstruction method based on explicit mesh.

4.2 Experimental Results

For the custom dataset, we selected three challenging scenes for presenting the experimental results: city street, night, and ramp. For additional experimental results on more scenes, please refer to the supplementary materials. Tab. 1

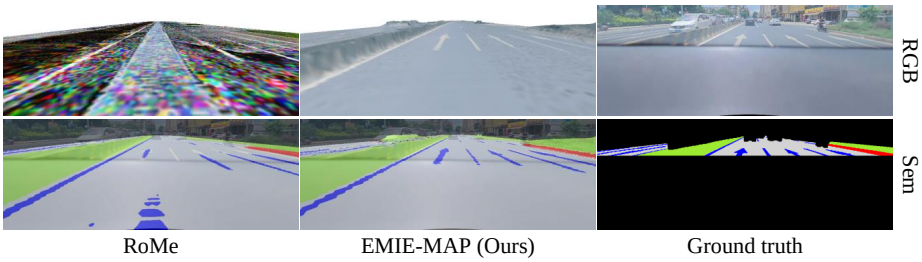

Fig. 6: Visualization of the rendered RGB and semantic images. It can be observed that the RGB map rendered by RoMe [23] appears chaotic, while our method is capable of rendering road surface colors accurately. Our method also achieves more accurate semantic reconstruction, as indicated by the higher consistency between the rendered semantic image and the ground truth RGB image overlay.

presents the road surface reconstruction evaluation results for the three scenes. Compared to RoMe [23], our method exhibits higher reconstruction accuracy and robustness in terms of road surface color, semantics, and elevation. Our method maintains remarkable reconstruction performance even under extreme lighting conditions and sharp elevation change, whereas RoMe [23] fails completely. Even without Lidar point cloud supervision, our method still significantly outperforms Rome [23].

Scenes	City street			Night			Ramp		
	PSNR ↑	mIoU (%) ↑	Elev-error (cm) ↓	PSNR ↑	mIoU (%) ↑	Elev-error (cm) ↓	PSNR ↑	mIoU (%) ↑	Elev-error (cm) ↓
RoMe [23]	17.31	94.87	45.08	0.99	4.91	331.23	3.93	3.95	496.24
EMIE-MAP w/o Lidar GT	24.44	95.12	17.08	22.12	94.14	27.34	19.20	85.07	93.12
EMIE-MAP w/ Lidar GT	**26.75**	**95.27**	**2.35**	**24.53**	**96.02**	**3.57**	**20.74**	**88.89**	**4.33**

Table 1: Road surface reconstruction evaluation results. We conducted evaluations in three challenging scenarios. For color and semantic information, we evaluate the performance using the Peak Signal-to-Noise Ratio (PSNR) and mean Intersection over Union (mIoU) calculated from rendered RGB and semantic images. For the elevation information, we measure the average distance between the Lidar road points and the reconstructed road surface, defined as the Elev-error. Compared to RoMe [23], our method demonstrates higher reconstruction accuracy and robustness. In the Night and Ramp scenarios, RoMe completely fails. The middle row shows the results without Lidar supervision, which are still significantly better than RoMe [23].

City street scene. Fig. 3 presents the results of road surface reconstruction in a city street scene. The input image reveals inconsistent luminance levels between the front view and other views. Our method is capable of reconstructing RGB maps that correspond to the camera's luminance levels, with the front view appearing brighter and the other view appearing darker. It is important to emphasize that both the front-view RGB map and other-view RGB map are decoded from the same optimized implicit color feature map. Different MLPs

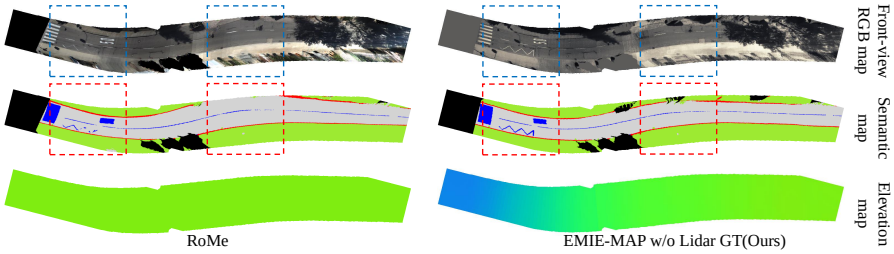

Fig. 7: Road surface reconstruction results in the KITTI-odometry dataset (sequence-09). EMIE-MAP is capable of reconstructing more accurate RGB, semantic, and elevation information than RoMe [23] in curved and uphill scenes.

Experiments	PSNR ↑	mIoU (%) ↑	Elev-error (cm) ↓	Runtime ↓	Memory (MB) ↓
RoMe [23]	20.87	94.57	107.98	**6'20"**	**3949**
EMIE-MAP w/o Lidar GT	**21.93**	**94.76**	**28.12**	14'20"	3969

Table 2: Quantitative results in the KITTI-odometry dataset (sequence-09).

are employed to decode the RGB features into their corresponding colors. The RGB features capture camera-independent scene essential information, while the different MLPs learn camera-specific attributes. Furthermore, despite the presence of heavy traffic on the street, our method is still able to reconstruct clear and complete road surfaces and lane markings. This is attributed to the consistency of scenes from different viewpoints, where occlusions can be mutually supplemented.

Fig. 6 presents the visualization of rendered RGB and semantic images. It can be observed that the RGB image rendered by RoMe [23] appears disordered due to the failure of optimization caused by inconsistent luminance between different cameras. In contrast, our method accurately renders road surface colors. From the overlay of the rendered semantic image and the ground truth RGB image, it is evident that our rendered semantic image exhibits higher consistency. This highlights the accuracy of our reconstruction method.

Night scene. Fig. 4 illustrates the road surface reconstruction results in a night scene. The scene is extremely dark, with only a small bright area visible in the image. In such circumstances, our method can only reconstruct RGB maps that include the visible portion of the road surface. However, our method is still able to achieve excellent results in terms of semantic mapping, which highlights the versatility of our method.

Ramp scene. Fig. 5 presents the road surface reconstruction results in a ramp scene. In this scene, there is a steep downhill slope on the road. Thanks to the designed trajectory-based elevation initialization and elevation residual prediction, our method accurately reconstructs the road surface elevation, thereby optimizing the scene's color and semantic information. It is important to note that if the road surface elevation optimization is inaccurate, there will be sig-

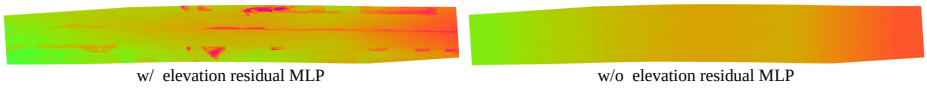

Fig. 8: Visualization of elevation maps in ablation experiments of elevation residual MLP. It is evident that using the elevation residual MLP results in more accurate elevation reconstruction.

nificant deviations in the RGB and semantic information based on projection optimization.

KITTI dataset. To demonstrate the robustness and accuracy of EMIE-MAP, frames 1140 to 1237 of sequence-09 were selected. In this test segment, the vehicle traversed a section of road that involved both an incline and a curve. Similar to RoMe [23], we utilize monocular images from KITTI's left RGB camera. To ensure a fair comparison with RoMe, we do not employ Lidar point cloud supervision for ground height. From Fig. 7, it can be seen that RoMe exhibits inconsistent road widths due to not considering changes in ground height. Even without Lidar point cloud supervision, EMIE-MAP can accurately model ground height and therefore achieve higher scores as shown in Tab 2. EMIE-MAP is slower than RoMe [23] because of more complex representation and optimization methods. It is acceptable for offline reconstruction.

4.3 Ablation Study

Effect of elevation residual MLP. Tab. 3 (a) demonstrates the effect of elevation residual MLP removal. Removing the elevation residual MLP results in a significant increase in elevation error. Fig. 8 illustrates that with the elevation residual MLP, the road surface elevation can be reconstructed more accurately with finer details.

Experiments	PSNR ↑	mIoU (%) ↑	Elev-error (cm) ↓
a. Ours w/o elevation residual MLP	26.63	94.35	12.78
b. Ours w/o RGB MLPs	15.60	95.16	2.49
c. Ours w/ embedding	25.83	95.06	2.49
d. Ours w/o sem	26.37	-	2.48
e. Full EMIE-MAP (Ours)	**26.75**	**95.27**	**2.35**

Table 3: Ablation study of our design choices on the city street. The results validate the effectiveness of each of our innovations.

Effect of the color representation. Tab. 3 (b) showcases the effectiveness of the proposed color representation. After removing multiple RGB MLPs and directly optimizing the color parameters, the PSNR significantly declines. It can be observed from Fig. 9 that the inconsistent luminance between different cameras leads to the failure of direct color optimization. However, our designed

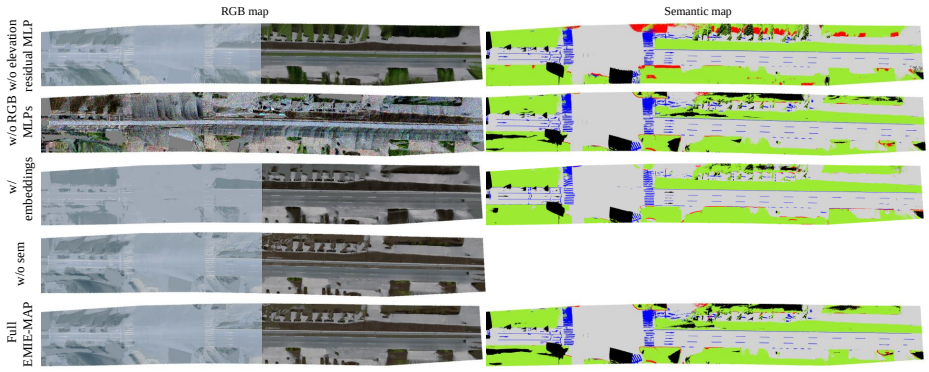

Fig. 9: Visualization of road surface reconstruction in ablation experiments. It is evident that using the elevation residual MLP results in more accurate elevation reconstruction. The inconsistent luminance between different cameras can cause explicit RGB optimization methods to fail. However, our color representation approach based on implicit color encoding and multiple RGB MLPs can address this challenge and achieve separate modeling for each camera.

representation, which utilizes shared implicit color encoding and multiple RGB MLPs, addresses this challenge and enables the mapping of observed colors from different cameras. Using rendering-related embedding can also help to deal with this. The results in Tab. 3 (c) demonstrate that different MLPs is superior to a single MLP with different embeddings, as the modeling capacity of MLPs is stronger.

Effect of semantic information. Tab. 3 (d) demonstrates the results lacking semantic information. The introduction of semantic can enhance the color and geometric reconstruction results, as the semantic consistency across multiple perspectives can facilitate reconstruction optimization.

5 Conclusion

We propose EMIE-MAP, a novel large-scale road surface reconstruction method that combines explicit mesh and implicit encoding. We introduce an elevation optimization approach based on trajectory-based initialization and elevation residual prediction, which enables high-precision elevation reconstruction. Additionally, we propose a color representation method based on implicit RGB encoding and multiple RGB MLPs, allowing for distinct modeling of scene characteristics and camera-specific characteristics. This addresses the challenge of inconsistent camera luminance across different camera models. Experimental results in various challenging scenarios demonstrate that EMIE-MAP achieves high accuracy and robustness in road surface reconstruction.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 62303307, 62225309, 62073222, 62361166632, and U21A20480.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)
2. Anciukevičius, T., Xu, Z., Fisher, M., Henderson, P., Bilen, H., Mitra, N.J., Guerrero, P.: Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12608–12618 (2023)
3. Brunken, H., Gühmann, C.: Road surface reconstruction by stereo vision. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science **88**(6), 433–448 (2020)
4. Chen, J., Yi, W., Ma, L., Jia, X., Lu, H.: Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20648–20658 (2023)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Deng, T., Chen, Y., Zhang, L., Yang, J., Yuan, S., Wang, D., Chen, W.: Compact 3d gaussian splatting for dense visual slam. arXiv preprint arXiv:2403.11247 (2024)
7. Deng, T., Liu, S., Wang, X., Liu, Y., Wang, D., Chen, W.: Prosgnerf: Progressive dynamic neural scene graph with frequency modulated auto-encoder in urban scenes. arXiv preprint arXiv:2312.09076 (2023)
8. Deng, T., Shen, G., Qin, T., Wang, J., Zhao, W., Wang, J., Wang, D., Chen, W.: Plgslam: Progressive neural scene represenation with local to global bundle adjustment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19657–19666 (June 2024)
9. Deng, T., Wang, N., Wang, C., Yuan, S., Wang, J., Wang, D., Chen, W.: Incremental joint learning of depth, pose and implicit scene representation on monocular camera in large-scale scenes. arXiv preprint arXiv:2404.06050 (2024)
10. Deng, T., Wang, Y., Xie, H., Wang, H., Wang, J., Wang, D., Chen, W.: Neslam: Neural implicit mapping and self-supervised feature tracking with depth completion and denoising. arXiv preprint arXiv:2403.20034 (2024)
11. Fan, R., Ai, X., Dahnoun, N.: Road surface 3d reconstruction based on dense subpixel disparity map estimation. IEEE Transactions on Image Processing **27**(6), 3025–3035 (2018)
12. Fan, R., Ozgunalp, U., Wang, Y., Liu, M., Pitas, I.: Rethinking road surface 3-d reconstruction and pothole detection: From perspective transformation to disparity map segmentation. IEEE Transactions on Cybernetics **52**(7), 5799–5808 (2021)
13. Feng, Z., Wu, W., Wang, H.: Rogs: Large scale road surface reconstruction based on 2d gaussian splatting. arXiv preprint arXiv:2405.14342 (2024)
14. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)

15. Guo, J., Tsai, M.J., Han, J.Y.: Automatic reconstruction of road surface features by using terrestrial mobile lidar. Automation in Construction **58**, 165–175 (2015)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Levoy, M., Hanrahan, P.: Light field rendering. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 441–452 (2023)
18. Li, Z., Li, L., Zhu, J.: Read: Large-scale neural scene rendering for autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1522–1529 (2023)
19. Liu, J., Wang, G., Jiang, C., Liu, Z., Wang, H.: Translo: A window-based masked point transformer framework for large-scale lidar odometry. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1683–1691 (2023)
20. Liu, J., Wang, G., Liu, Z., Jiang, C., Pollefeys, M., Wang, H.: Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8451–8460 (2023)
21. Liu, J., Wang, G., Ye, W., Jiang, C., Han, J., Liu, Z., Zhang, G., Du, D., Wang, H.: Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15109–15119 (2024)
22. Liu, J., Zhuo, D., Feng, Z., Zhu, S., Peng, C., Liu, Z., Wang, H.: Dvlo: Deep visual-lidar odometry with local-to-global feature fusion and bi-directional structure alignment. arXiv preprint arXiv:2403.18274 (2024)
23. Mei, R., Sui, W., Zhang, J., Qin, X., Wang, G., Peng, T., Chen, T., Yang, C.: Rome: Towards large scale road surface reconstruction via mesh representation. IEEE Transactions on Intelligent Vehicles (2024)
24. Melas-Kyriazi, L., Rupprecht, C., Vedaldi, A.: Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12923–12932 (2023)
25. Meuleman, A., Liu, Y.L., Gao, C., Huang, J.B., Kim, C., Kim, M.H., Kopf, J.: Progressively optimized local radiance fields for robust view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16539–16548 (2023)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
27. Moulon, P., Monasse, P., Marlet, R.: Adaptive structure from motion with a contrario model estimation. In: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV 11. pp. 257–270. Springer (2013)
28. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion*. Acta Numerica **26**, 305–364 (2017)
29. Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12932–12942 (2022)
30. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)

31. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 501–518. Springer (2016)
32. Shan, J., Li, Y., Xie, T., Wang, H.: Enerf-slam: A dense endoscopic slam with neural implicit representation. IEEE Transactions on Medical Robotics and Bionics (2024)
33. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)
34. Snavely, N.: Scene reconstruction and visualization from internet photo collections: A survey. IPSJ Transactions on Computer Vision and Applications **3**, 44–66 (2011)
35. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM siggraph 2006 papers, pp. 835–846 (2006)
36. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
37. Tang, Z.J., Cham, T.J., Zhao, H.: Able-nerf: Attention-based rendering with learnable embeddings for neural radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16559–16568 (2023)
38. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4190–4200 (2023)
39. Turki, H., Zhang, J.Y., Ferroni, F., Ramanan, D.: Suds: Scalable urban dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12375–12385 (2023)
40. Waechter, M., Moehrle, N., Goesele, M.: Let there be color! large-scale texturing of 3d reconstructions. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 836–850. Springer (2014)
41. Wang, Z., Zhou, S., Park, J.J., Paschalidou, D., You, S., Wetzstein, G., Guibas, L., Kadambi, A.: Alto: Alternating latent topologies for implicit 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 259–270 (2023)
42. Wu, C.: Towards linear-time incremental structure from motion. In: 2013 International Conference on 3D Vision-3DV 2013. pp. 127–134. IEEE (2013)
43. Wu, W., Wang, G., Deng, T., Aegidius, S., Shanks, S., Modugno, V., Kanoulas, D., Wang, H.: Dvn-slam: Dynamic visual neural slam based on local-global encoding. arXiv preprint arXiv:2403.11776 (2024)
44. Yang, L., Zhang, X., Yu, J., Li, J., Zhao, T., Wang, L., Huang, Y., Zhang, C., Wang, H., Li, Y.: Monogae: Roadside monocular 3d object detection with ground-aware embeddings. IEEE Transactions on Intelligent Transportation Systems (2024)
45. Yu, S.J., Sukumar, S.R., Koschan, A.F., Page, D.L., Abidi, M.A.: 3d reconstruction of road surfaces using an integrated multi-sensory approach. Optics and lasers in engineering **45**(7), 808–818 (2007)
46. Zhao, T., Xie, Y., Ding, M., Yang, L., Tomizuka, M., Wei, Y.: A road surface reconstruction dataset for autonomous driving. Scientific data **11**(1), 459 (2024)
47. Zhao, T., Yang, L., Xie, Y., Ding, M., Tomizuka, M., Wei, Y.: Roadbev: Road surface reconstruction in bird's eye view. arXiv preprint arXiv:2404.06605 (2024)

48. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12588–12597 (2023)
49. Zhu, S., Wang, G., Blum, H., Liu, J., Song, L., Pollefeys, M., Wang, H.: Sni-slam: Semantic neural implicit slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21167–21177 (2024)