

BEV-GS: Feed-forward Gaussian Splatting in Bird’s-Eye-View for Road Reconstruction

Wenhua Wu¹, Tong Zhao^{2†}, Chensheng Peng³, Lei Yang⁴, Zhe Liu¹, and Hesheng Wang^{1†}

¹ Shanghai Jiao Tong University, China

² Xpeng Motors, China

³ University of California, Berkeley, USA

⁴ Nanyang Technological University, Singapore

Abstract. Road surfaces serve as the only contact medium between vehicle wheels or robot feet and the physical world. Reconstructing the road surface is crucial for unmanned vehicles and mobile robots. Recent studies on Neural Radiance Fields (NeRF) and Gaussian Splatting (GS) have achieved remarkable results in scene reconstruction. However, they typically rely on multi-view image inputs and require prolonged optimization times. In this paper, we propose BEV-GS, a real-time single-frame road surface reconstruction method based on feed-forward Gaussian splatting. BEV-GS consists of a prediction module and a rendering module. The prediction module introduces separate geometry and texture networks following the Bird’s-Eye-View (BEV) paradigm. Geometric and texture parameters are directly estimated from a single frame, avoiding per-scene optimization. In the rendering module, we utilize grid Gaussian for road surface representation and novel view synthesis, which better aligns with road surface characteristics. Our method achieves state-of-the-art performance on the real-world dataset RSRD. The road elevation error is reduced to 1.73 cm, and the PSNR of novel view synthesis reaches 28.36 dB. The prediction and rendering FPS are 26 and 2061, respectively, enabling high-accuracy and real-time applications. Our code is available at <https://github.com/IRMVLab/BEV-GS>.

1 Introduction

Road surface plays a vital role in unmanned vehicles and robots, as it is the only contact medium between wheels and the physical world. Geometry and texture are the most significant road properties. Road geometry or elevation determines ride comfort and motion smoothness. Road texture and color reflect material, friction level, and unevenness, which are crucial clues for understanding road scene conditions.

Road surface reconstruction (RSR), aiming at recovering the properties above, significantly improves the driving safety and comfort of autonomous driving vehicles [37, 50]. Online road preview with camera sensors is a promising way that

[†] Corresponding Author.

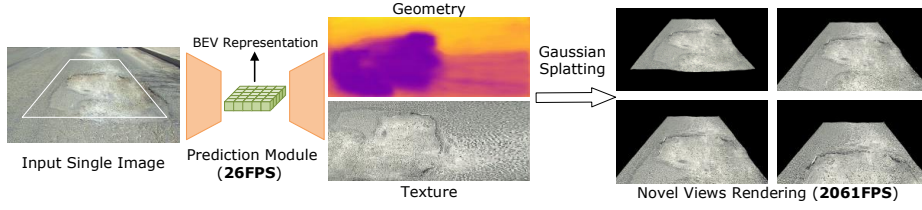


Fig. 1: BEV-GS is capable of real-time road surface reconstruction from a single image. Unlike existing per-pixel feed-forward Gaussian prediction methods, it predicts geometry and texture parameters of mesh grid Gaussians based on BEV features. BEV-GS achieves remarkable road surface reconstruction performance and can rapidly render novel views of the road. The images on the right side simulate the captured road as the car moves forward.

gains attention, where monocular visual perception has become the mainstream low-cost sensing solution for intelligent robot inspection and environmental reconstruction tasks [21]. Reconstruction methods based on monocular and multi-view images in perspective view (PV) have been extensively explored [9, 24]. They recover road point clouds from dense depth estimation associated with direct color projection. However, unstructured road representation inhibits downstream applications or fusing multi-modal data. RoadBEV [52] first proposes to recover road profiles in Bird’s Eye View (BEV) with both monocular and stereo settings. Nevertheless, they reconstruct only road geometry without considering color.

Recent studies on Neural Radiance Fields (NeRF) [25] and Gaussian Splatting (GS) [18] have achieved remarkable results in large-scale scene reconstruction, particularly in color rendering [6, 12, 42]. However, these methods typically rely on multi-view image inputs and require long optimization time [28, 38, 54]. EMIE-MAP [43] and RoGS [13] introduce implicit representations and GS into road surface reconstruction. They are offline and scene-specific, requiring iterative optimization over input image sequences. Additionally, they rely on structure from motion (SfM) or vehicle trajectory information for initializing road surface geometry.

The vanilla NeRF and GS methods memorize a certain scene from massive images. Despite their strong fitting capabilities, they do not generalize well to new scenarios since a universal prior is not learned. Feed-forward GS-based methods have been developed for single-frame scene reconstruction [34, 35]. Specifically, they use an image-to-image network to predict per-pixel GS parameters and then render novel-view images. With sufficient prior information learned during training, this paradigm is scene-independent and generalizable to unseen scenarios. However, existing methods rely on pre-trained depth estimation models. Per-pixel prediction is not suitable for scenes with large near-far distance spans like open streets. The perspective effect causes near points to be dense and far points to be sparse.

To resolve the aforementioned issues, we propose BEV-GS, a real-time road surface reconstruction method combining BEV paradigm and GS, without the need for multi-view inputs and iterative optimization, as shown in Figure 1. In contrast to existing pixel-wise feed-forward GS methods, we introduce a novel Gaussian prediction network based on BEV feature grids, where scene geometry and texture are decoupled. For road geometry, image features in PV are transformed into 3D road voxels of interest in BEV, which facilitates surface occupancy prediction for every grid. An independent backbone is introduced to extract texture features. The recovered road elevation contributes to more accurate texture feature query, based on which Gaussian-related color parameters are decoded for grids. Building upon the predictions, we construct a road surface representation based on grid Gaussian. Gaussians conditioned on the accurate surface profile naturally fit road scene, avoiding complex geometric regularization. Predictive and realistic novel view synthesis is achieved in real-time. Benefiting from the BEV representation and the decoupling of texture and geometry, the superiority of BEV-GS is sufficiently verified by experiments on a real-world dataset. Our main contributions are summarized as follows:

- We propose BEV-GS, a feed-forward Gaussian Splatting method based on BEV representation with application to road surface reconstruction. It addresses the issue of uneven information distribution in perspective view.
- We introduce decoupled geometry and texture prediction. It enables independent optimization, while the geometry branch promotes more accurate texture feature query.
- We achieve state-of-the-art performance on a real-world road surface reconstruction dataset, showcasing its capability in real-time road surface reconstruction and novel view synthesis.

2 Related Works

Road surface reconstruction. RSR is also an essential topic in pavement engineering and remote sensing [14, 53]. RSR for unmanned vehicles and robots concentrates more on surface profile, texture, and semantics recovery with affordable LiDAR [27] or camera [10, 16, 48], aiming at improving safety and smoothness. LiDAR directly scans unstructured road profile, while lacking texture information [39, 51]. RoadBEV reconstructs road geometry in BEV with both monocular and stereo cameras [52], achieving errors of 1.83 cm and 0.50 cm, respectively. RoME first utilizes road mesh representation and reconstructs large-scale roads with NeRF from sequential images, offering both road structure and color for 4D labeling [24]. Further, EMIE-MAP [43] predicts elevation and color with independent MLPs, achieving novel-view synthesis with direct projection. RoGS [13] uses 3GGS as the road representation, significantly enhancing the reconstruction speed. Nevertheless, road geometry suffers from poor accuracy because road profiles are unsupervised or supervised by sparse labels.

Reconstruction with Gaussian Splatting. GS with explicit scene representation is becoming popular for its impressive efficiency [3, 41]. Its improved forms

have been applied to various scenarios like street view [47], dynamic scene [54], and SLAM [23]. The Gaussians are generally initialized on points derived from SfM or depth estimation. Despite the density adjustment, geometry still suffers from artifacts such as floaters. Most recent works focus on geometric regularization for smooth and consistent surfaces [4, 8, 49]. GS2Mesh [40] extracts smooth mesh from noisy points with the aid of stereo images. Trim GS [8] proposes a trimming strategy that removes redundant or inaccurate Gaussians. Despite the efforts, Gaussian centers are still presented in an unstructured manner. Adaptive regularization strategies for universal scenes are challenging.

Feed-forward conditioned rendering. To tackle the challenge of scene extrapolation and generalization in vanilla NeRF and GS methods, feature-based rendering with scene encoding is introduced. Methods such as [19, 26, 33] treat grid-based features as parameters and optimize them by back-propagation, which still have the identified problems above. PixelNeRF [46] first propose a learning-based feed-forward framework for predicting NeRFs from single or several images. GSN [7] and RNRMap [20] map image features extracted by CNN to 2D floor-plan grids of local latent codes. Given a novel view, camera rays sample latent features by vertical projection and interpolation, based on which a feature map is rendered. Splatter Image [35] maps an input image to another image that holds the parameters of one coloured 3D Gaussian per pixel, achieving super fast single-view reconstruction. MVSplat [5], DepthSplat [44] and AnySplat [17] extend the feedforward Gaussian to multi-frame reconstruction.

3 Methods

3.1 Framework Overview

We propose BEV-GS, a single-frame and feed-forward GS model in BEV with application to road reconstruction. First, following RoadBEV [52], we predefine a horizontal region of interest in front of the vehicle with size $1.9\text{m} \times 5.0\text{m}$. The region of interest is defined based on the road the vehicle will traverse, which directly influences the vehicle’s response. A lateral width of 1.9m accommodates most passenger vehicles, ensuring usable surface data along both tire tracks. The 5m longitudinal length covers the travel distance for upcoming frames. The rectangular area is then discretized into grids, which sufficiently cover the road that ego-vehicle will pass through. Every grid in the mesh representation acts as the minimal unit for road geometry and texture reconstruction.

As illustrated in Figure 2, BEV-GS comprises a prediction module and a rendering module. The former follows the paradigm of BEV perception for representing road surface. To achieve few-shot scene generalization, we adopt the feed-forward scheme for recovering both road geometry and texture. Two independent backbones are introduced to encode road image and transform into grid or voxel features. The geometry decoder first estimates one-dimensional elevation values of the grids. The recovered 3D road profile facilitates view transformation and feature query of the texture branch. The texture decoder then predicts color-related parameters for every grid in BEV.

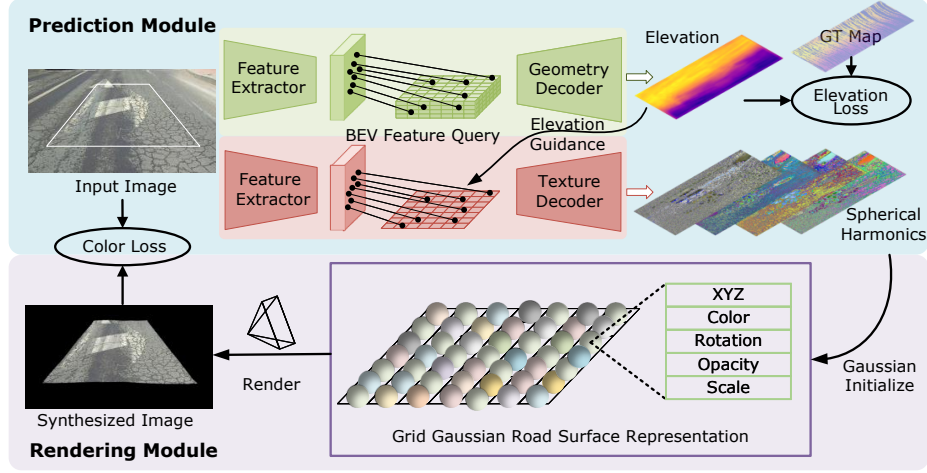


Fig. 2: Overview of BEV-GS. It consists of a prediction module and a rendering module. In the prediction module, independent geometry and texture branches based on BEV features predict road elevation and spherical harmonic parameters. In the rendering module, we represent the road surface with grid Gaussians, which are initialized by the predicted properties. Elevation and color losses individually supervise the feed-forward predictions.

In the rendering module, we leverage the grid Gaussians for reconstructing road texture and synthesizing RGB images. With parameters predicted by the texture decoder, 3D Gaussians conditioned on the recovered 3D meshes are splatted across the image plane with differentiable rendering. The texture branch can therefore be end-to-end optimized with the loss between synthesized and actual images. The feed-forward scheme enables predictive novel view synthesis with single-frame input.

Benefiting from the unified and structured BEV representation, the uneven distribution of information caused by the perspective effect is mitigated. The proposed framework facilitates separate training since the two branches interact only by geometry initialization. The decoupled design enables each branch to focus on its respective domain.

3.2 Road Geometry Reconstruction

Given a single-frame input $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$, the simplified EfficientNet [36] with pre-trained weights is adopted as the encoder E_g to extract road geometric features $\mathbf{F}_{pv}^g \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$. Instead of using transformer-based cross-attention to enlarge the receptive field during BEV query [22, 29], we adopt direct 3D-2D projection for view transformation since dense road voxels are already set for surface occupancy prediction. The projected reference points aggregate features of nearby pixels with bilinear weighting. A 4D voxel feature $\mathbf{F}_{vox} \in \mathbb{R}^{C \times N_x^g \times N_y^g \times N_z^g}$

is derived, where N_x^g , N_y^g , and N_z^g denote the number of anchor voxels in the lateral, longitudinal, and vertical directions, respectively.

RoadBEV [52] reduces the height dimension of voxel feature by direct channel concatenation. Instead, to reduce computation while preserving key features, we adopt weighted sum for fusing anchor features of grids. Stacked 3D convolution layers followed by a *Softmax* operation along the height dimension are introduced to evaluate the significance of every elevation anchor. The geometry BEV feature $\mathbf{F}_{\text{bev}}^g \in \mathbb{R}^{C \times N_x^g \times N_y^g}$ is derived as:

$$\mathbf{F}_{\text{bev}}^g(g) = \sum_{z=1}^{N_z^g} \mathbf{F}_{\text{vox}}(g, z) \cdot \text{Softmax}(\mathbf{F}'_{\text{vox}}(g, \cdot)), \quad (1)$$

where g indicates a horizontal grid at (x, y) , $\mathbf{F}'_{\text{vox}} \in \mathbb{R}^{N_x^g \times N_y^g \times N_z^g}$ is the single-channel voxel feature exported from the 3D convolution layers.

A geometry decoder D_g composed of basic blocks in EfficientNet [36] further estimates elevation based on $\mathbf{F}_{\text{bev}}^g$. Since the elevation reference plane is fixed below camera, actual road surface may have holistically deviation due to vehicle pitch vibration. Instead of direct regression, we separately estimate a global reference plane $\hat{H}_{\text{ref}}$ and an offset map $\hat{\mathbf{H}}_{\text{off}}$, which are all in the range of $[\frac{h_{\min}}{2}, \frac{h_{\max}}{2}]$, where $h_{\min}$ and $h_{\max}$ are the lower and upper elevation bounds, respectively. The final elevation map is their combination:

$$\hat{\mathbf{H}} = \hat{H}_{\text{ref}} + \hat{\mathbf{H}}_{\text{off}}, \quad (2)$$

The reference $\hat{H}_{\text{ref}}$ is derived by cascaded average pooling, MLP layers, and *Tanh* activation operations on the feature map. Since direct regression is ineffective for model supervision [52], the offset value for every grid is obtained by classification on offset bins and regression with weighted sum:

$$\hat{\mathbf{H}}_{\text{off}}(g) = \sum_{b=1}^{N_b} e_b \cdot \text{Softmax}(\mathbf{F}_{\text{bev}}^e(\cdot, g)), \quad (3)$$

where $\mathbf{F}_{\text{bev}}^e \in \mathbb{R}^{N_b \times N_x^g \times N_y^g}$ is the final feature map for elevation offset classification, e_b is the elevation value of corresponding bin. N_b is the number of offset bins.

Compared with existing GS methods that generally utilize COLMAP [31] for point cloud initialization, the proposed geometry branch offers structured points that are naturally suitable for describing road surface. The accurate scene geometry guarantees reliable feature query and texture decoding for predicting Gaussian-related properties.

3.3 Road Texture Representations

The architecture of the texture prediction branch is generally similar to the geometry branch. It also implements a BEV pipeline that predicts texture information for each Gaussian sphere conditioned on the grid. Unlike in [34], where a

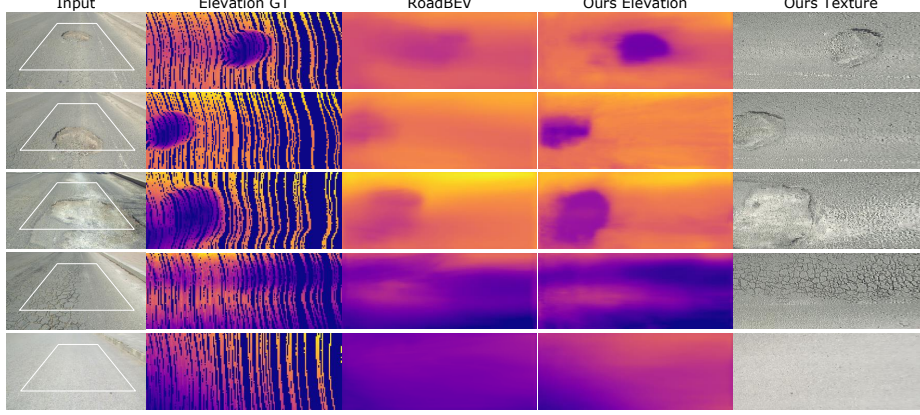


Fig. 3: Visualization of road reconstruction results under typical scenarios. Compared with RoadBEV [52], our method not only reconstructs more accurate road geometry but also the detailed road texture.

shared backbone is used for both geometry and texture prediction, we use a separate texture encoder E_t composed of simplified MobileNetV3 [15]. The texture feature map $\mathbf{F}_{pv}^t \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$ with $\frac{1}{4}$ resolution is obtained:

$$\mathbf{F}_{pv}^t = E_t(\mathbf{I}). \quad (4)$$

This decoupled design proves to be more effective as structure and texture recovery require distinct domain features, while shared features suffer from obscure representations. Similarly, we utilize 3D to 2D projection to query texture BEV features $\mathbf{F}_{bev}^t \in \mathbb{R}^{C \times N_x^t \times N_y^t}$. For more accurate feature queries, we leverage the elevation predicted by the geometry branch for projection guidance without the need for lifting anchors. Subsequently, a texture decoder D_t followed by an up-sampling layer is employed to predict dense spherical harmonics coefficients $\mathbf{C} \in \mathbb{R}^{3(L+1)^2 \times 2N_x^t \times 2N_y^t}$ of grids, where L is the order of the spherical harmonics. The up-sampling aims to obtain more fine-grained and high-resolution road texture information.

3.4 Feed-forward Gaussian Splatting

Grid Gaussian road surface representation. Building upon the BEV-based reconstruction of road geometry and texture, we establish a road surface representation with mesh grid Gaussian. Existing schemes typically maintain non-uniformly distributed 3D/2D Gaussian spheres, necessitating optimization of Gaussian positions and densities. Considering the physical characteristics of road surface, we arrange 3D Gaussian spheres on the uniformly partitioned road grid cells. The uniform distribution of the fixed Gaussians ensures coverage over all road surface areas without the need for density optimization. Each sphere comprises parameters of geometric center, scale, rotation, opacity, and color. The

Table 1: Comparison of road geometry reconstruction. Our method shows state-of-the-art performance. **Bold:** best. ‘PV’ denotes Perspective View.

Methods		AAE (cm) ↓	RMSE (cm) ↓	>5mm(%) ↓
PV	iDisc [30]	2.70	2.89	86.3
	LapDepth [32]	2.55	2.79	85.2
	PixelFormer [1]	2.47	2.66	84.9
	AdaBins [11]	2.36	2.55	84.0
	ZoeDepth [2]	2.41	2.59	82.5
	DepthAnything V2 [45]	2.30	2.48	83.2
BEV	RoadBEV [52]	1.93	2.16	83.4
	BEV-GS (Ours)	1.73	1.94	80.2

geometric centers μ_i are initialized by the horizontal coordinates (x, y) of grid centers and the estimated elevation $\hat{\mathbf{H}}(x, y)$. Colors are initialized by predicted spherical harmonic coefficients $\mathbf{C}(x, y) \in \mathbb{R}^{3(L+1)^2}$. Scales of Gaussians are all set to $\mathbf{S} \in \mathbb{R}^3$. Rotations $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ are treated as the identity matrix, and opacities are initialized as scalar σ , making each Gaussian sphere isotropic. The covariance of 3D Gaussians can be expressed as:

$$\boldsymbol{\Sigma} = \mathbf{R} \mathbf{S} \mathbf{S}^T \mathbf{R}^T. \quad (5)$$

The spatial influence at point p given by:

$$g_i(p) = e^{-\frac{1}{2}(p - \mu_i)^T \boldsymbol{\Sigma}^{-1}(p - \mu_i)}. \quad (6)$$

Gaussian Splatting. Given the camera pose $\mathbf{P}$, the grid Gaussians are projected onto the 2D image plane:

$$\boldsymbol{\Sigma}' = \mathbf{J} \mathbf{P} \boldsymbol{\Sigma} \mathbf{P}^T \mathbf{J}^T, \quad (7)$$

where $\mathbf{J}$ is the affine-approximated Jacobian matrix of the projection function. The Gaussian sphere distribution is sorted in depth order for rendering color of each pixel p :

$$\hat{\mathbf{I}}(p) = \sum_{i \in N} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \quad (8)$$

where c_i represents the color of the i -th Gaussian obtained from spherical harmonic coefficients, and α'_i represents the density computed from opacity and covariance.

3.5 Loss Functions

For the geometry branch, we use smoothed L1 loss to supervise elevation reconstruction. The elevation loss of a sample is the total loss of grids with valid label:

$$\mathcal{L}_{\text{ele}} = \sum_{g \in \mathcal{M}_{\text{ele}}} \left| \hat{\mathbf{H}}(g) - \mathbf{H}_{\text{gt}}(g) \right|, \quad (9)$$

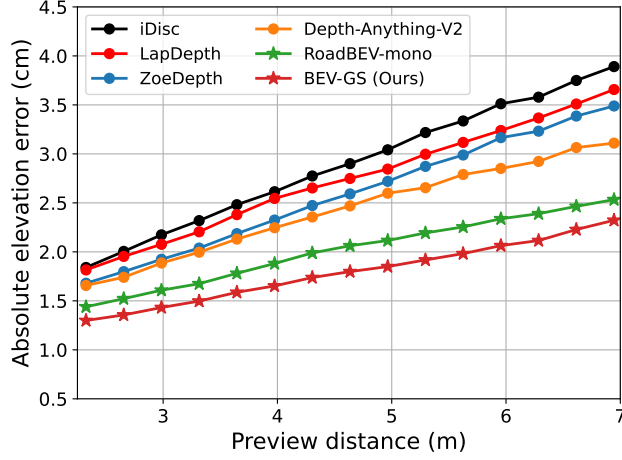


Fig. 4: Comparison of segment-wise elevation errors with the SOTA depth estimation and BEV models. Our BEV-GS outperforms all baselines in all segments, demonstrating robust fine-grained road geometry reconstruction for near and distant road surfaces.

where $\mathbf{H}_{\text{gt}}$ is the ground-truth (GT) elevation map, $\mathcal{M}_{\text{ele}}$ is the set of valid grids.

For the texture branch, we use both L1 loss and SSIM loss to supervise the synthesized image:

$$\mathcal{L}_{\text{rgb}} = \sum_{i \in \mathcal{I}} \lambda |\hat{\mathbf{I}}_i - \mathbf{I}_i| + (1 - \lambda)(1 - \text{SSIM}(\hat{\mathbf{I}}_i, \mathbf{I}_i)), \quad (10)$$

where $\mathcal{I}$ is the set of actual images under different poses. λ is the weight. Since there may be empty areas in the synthesized image, the loss only considers the splattered pixels.

4 Experiments

In this section, we demonstrate the excellent performance of our method through comprehensive experiments and highlight its superiority compared to both existing geometry and texture reconstruction methods.

4.1 Experimental Settings

Dataset. We conduct experiments on the RSRD dataset [51], which is a real-world dataset focusing on road surface reconstruction. It offers high-accuracy data under various road conditions, including stereo images, point clouds, and processed GT elevation maps. The scenarios cover common road types, including a significant number of potholes and cracked road surfaces. We partitioned the

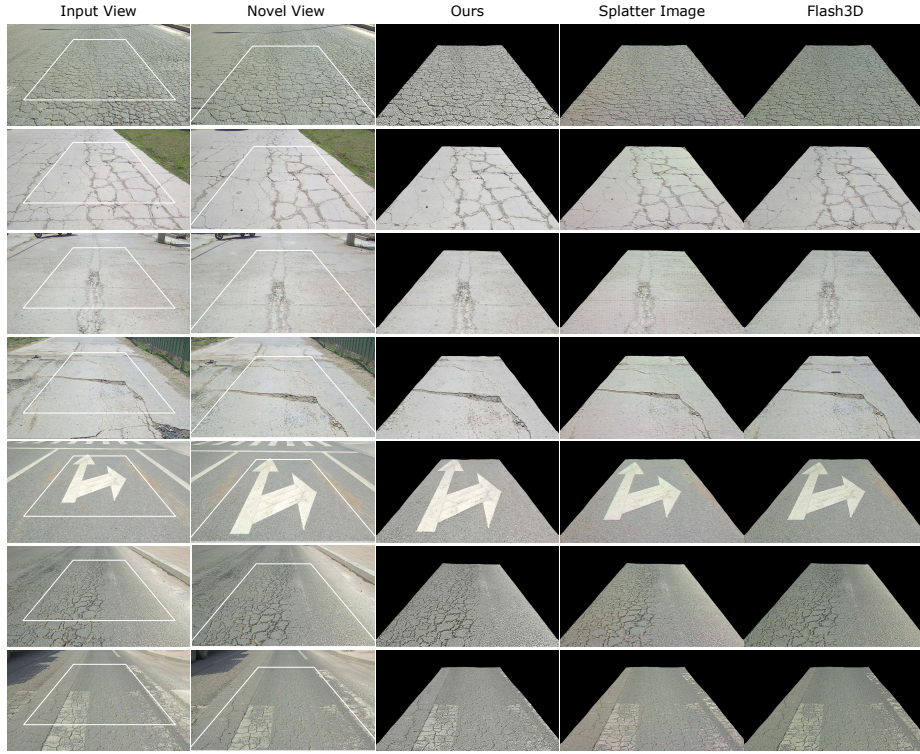


Fig. 5: Visualization of novel view synthesis, which is the next frame as the vehicle moves forward. For comparison purposes, the render results of Splatter Image [35] and Flash3D [34] are restricted to the same area as ours. Due to inaccurate geometric estimation in the perspective view, noticeable misalignment exists in their results.

RSRD dataset into a training set containing 926 samples and a test set containing 285 samples. The resolution of the input image is $H = 528$ and $W = 960$.

Evaluation Metrics. Following RoadBEV [52], we calculate the average absolute error (AEE), root mean square error (RMSE), and the percentage of grids with absolute error exceeding 5mm ($> 5\text{mm}$) between the predicted and GT elevation to evaluate geometry reconstruction performance. For texture metrics, due to the absence of GT colors of road surface in BEV, we evaluate texture reconstruction through perspective view rendering. We calculate the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and the learned perceptual image patch similarity (LPIPS) between the synthesized and observed images.

Implementation Details. The image feature maps have $C = 128$ channels. The grid interval of road area is set as 3 cm, resulting in grid size $N_x^g = 64$, and $N_y^g = 164$. The elevation range is 20 cm above or below the zero reference plane, i.e., $h_{\min} = -20$ cm, and $h_{\max} = 20$ cm. Each grid has $N_z^g = 20$ elevation

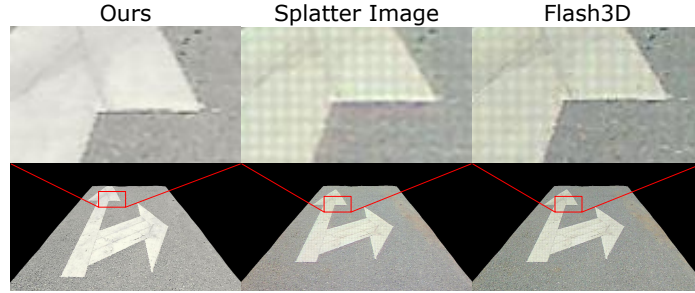


Fig. 6: Qualitative highlight of distant road surface novel view synthesis. Our BEV-GS retains fine texture and structural details for distant roads, outperforming baselines with per-pixel Gaussians.

Table 2: Quantitative comparison of novel view synthesis. **Bold:** best. Our BEV-GS outperforms baselines by a large margin on all metrics.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Splatter Image [35]	22.32	0.70	0.41
Flash3D [34]	23.27	0.71	0.41
BEV-GS (Ours)	28.36	0.77	0.14

anchors for geometry feature query. There are $N_b = 40$ bins for elevation offset classification.

For higher resolution and more realistic rendering, the grid size of texture representation is four times that of the geometry, i.e., $N_x^t = 256$, and $N_y^t = 656$. The order of spherical harmonics is $L=1$. The scale of Gaussians $\mathbf{S}$ is initialized to $[0.002, 0.002, 0.002]^T$, the opacity σ is initialized to 1.0. The weight λ in $\mathcal{L}_{\text{rgb}}$ is 0.5. In the training phase, we use stereo left and right images in RSRD for rendering supervision, i.e., $\mathcal{I} = \{\mathbf{I}_l, \mathbf{I}_r\}$. The prediction network is trained for 50 epochs with AdamW optimizer. Batch size is 8. The maximum learning rate is 5×10^{-4} with linear decay. All experiments are conducted on a server equipped with an NVIDIA A100 GPU.

Baselines. We compare our BEV-GS with state-of-the-art (SOTA) single-image reconstruction methods. For scene geometry, we compare against method in BEV like RoadBEV [52], and methods in perspective view like iDisc [30], ZoeDepth [2], and Depth Anything V2 [45]. We select the DPT-BEiT-L as the backbone for ZoeDepth, and Swin-T for iDisc. The models above are all re-trained for 50 epochs with their default configurations on the RSRD dataset [51]. The maximum depth is set as 13m for the metric depth estimation models. For texture reconstruction, we compare against Splatter Image [35] and Flash3D [34].

Table 3: Analysis of model runtime and complexity. ‘-’ denotes no rendering. BEV-GS achieves superior real-time performance with the least number of parameters while supporting both geometry reconstruction and novel view synthesis.

Methods		Infer. FPS	Rend. FPS	# Params
Geometry only	iDisc [30]	12.3	-	41M
	LapDepth [32]	83.2	-	16M
	PixelFormer [1]	43.0	-	89M
	AdaBins [11]	23.2	-	78M
	ZoeDepth [2]	15.2	-	345M
	Depth Anything V2 [45]	6.6	-	335M
	RoadBEV [52]	26.8	-	27M
Texture	Splatter Image [35]	18.9	412	56 M
	Flash3D [34]	6.3	96	400 M
	BEV-GS (Ours)	26.3	2061	8M

4.2 Results and Comparison

Geometry Reconstruction Results. Road elevation reconstruction results are shown in Table 1. For depth estimation methods in PV, we convert the recovered point clouds into BEV and generate elevation maps in the same style as building GT maps. Benefiting from direct elevation modeling, our BEV-GS achieves SOTA performance with a significant margin. The elevation AAE and RMSE reach 1.73cm and 1.94cm, reducing 10.4% and 10.2% than RoadBEV, respectively. Compared with monocular depth estimation, elevation in BEV is naturally more suitable for describing road geometry. Figure 3 visualizes the predicted and GT elevation maps. BEV-GS provides more accurate and uniform road profiles than RoadBEV [52], especially in areas with irregularities. The shape and edge of the potholes are all precisely captured. Recovered surface profiles on flat roads are also smooth without noise, avoiding wrong identifications or unnecessary actions in the subsequent tasks.

For more insights of reconstruction performance, we further visualize the segment-wise AAE along the longitudinal direction, as shown in Figure 4. The entire grid is divided into 15 segments from near to far, with each segment representing a 33-centimeter-long region. By averaging the AAE values within each segment, a penetrative performance assessment is achieved. Our BEV-GS thoroughly outperforms the others in all segments, verifying its capability to capture fine-grained road structure. The maximum error can be limited within 2.5cm. It is also indicated that the elevation error increases with distance, which is still caused by the loss of texture details in farther areas.

Texture Reconstruction Results. Our BEV-GS also recovers detailed road color and texture, as shown in the last column of Figure 3. Fine structures of road cracks and potholes are all clearly retained without distortion, which are important for downstream road perception tasks such as condition understanding. Table 2 presents the quantitative performance comparison of novel view synthe-

sis. Our BEV-GS outperforms the very recent SOTA models significantly across all three metrics. Figure 5 visualizes the comparison of actual and synthesized novel views in typical road scenarios. The accurate geometry and BEV-based grid Gaussian representation contribute to globally consistent and high-quality novel view synthesis. Compared with the per-pixel Gaussian prediction methods, our method synthesizes images with more accurate correspondence, more realistic colors, and finer details. Figure 6 highlights the rendering of distant road surfaces. Our method retains fine details, whereas others produce poor rendering images due to the sparsity of distant road Gaussians.

Runtime Analysis. Table 3 compares inference or rendering speed, as well as model complexity. Among the methods for road geometry recovery, BEV-GS achieves outstanding inference speed with much less parameters. For joint texture and geometry reconstruction, BEV-GS outperforms the others among all metrics. Benefiting from the BEV grid representation, our model achieves a feed-forward prediction speed of 26 FPS, which implies real-time capability for practical applications.

Test Time Optimization.

In addition to prediction, the scheme enables backward and iterative optimization during testing. The predicted geometry and texture properties are treated as initial parameters of GS optimizer, which further updates color, rotation, and opacity of Gaussians. The learning rates for zero and first order spherical harmonics, rotation, and opacity are 0.05, 0.001, 0.001, and 0.001, respectively. We only use the current frame for supervision with the color loss $\mathcal{L}_{\text{rgb}}$. Figure 7 illustrates the performance enhancement with further optimization. The synthesis

metrics are all improved, converging around 60 iterations. Test-time optimization significantly enhances texture reconstruction quality, albeit at the cost of increased processing time. In applications, the decision to employ test-time optimization and the number of iterations can be based on the trade-off between reconstruction quality and real-time requirements.

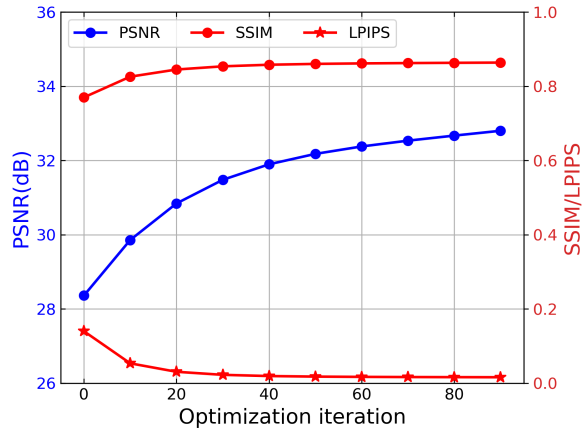


Fig. 7: Variation of synthesis metrics with the number of optimization iterations. Each iteration takes 11 ms. The metrics improve significantly with the increase of iterations and converge at around 60 iterations.

Table 4: Results for ablating different design configurations of our method. The results demonstrate that the full method outperforms all ablated configurations on both geometric and texture metrics, validating the effectiveness of our key design choices.

Configurations	a. w/ RoadBEV	b. w/ shared backbone	c. w/o elevation guidance	d. full method
AAE (cm) ↓	1.93	1.89	1.73	1.73
RMSE (cm) ↓	2.16	2.13	1.94	1.94
>5mm (%) ↓	83.4	82.1	80.2	80.2
PSNR (dB) ↑	26.75	24.90	25.00	28.36
SSIM↑	0.58	0.48	0.48	0.77
LPIPS ↓	0.20	0.28	0.26	0.14

4.3 Ablation Studies

Ablation of the Geometry Branch. Table 4 a. presents the results of replacing the geometry branch of BEV-GS with RoadBEV [52], where both the geometry and texture metrics exhibit a significant decline. Leveraging the weighted fusion strategy for BEV voxel features, along with the separate prediction of a global reference plane and an offset map, our geometry branch yields a more accurate road elevation map. This provides more reliable elevation guidance for the BEV feature query of the texture branch, effectively avoiding feature misalignment between the perspective view texture features and BEV grid positions.

Effectiveness of Geometry-Texture Separation. Table 4 b. presents the results of utilizing a shared feature extractor for geometry and texture. Both geometry and texture reconstruction metrics exhibit a noticeable decline. The geometry and texture exhibit distinct feature patterns. When a shared feature extractor is employed, geometry and texture features become coupled, leading to mutual interference. The proposed two-branch design decouples these aspects, facilitating improved learning of geometry and texture.

Effectiveness of Elevation Guidance. Table 4 c. demonstrates the results of BEV texture feature queries without elevation guidance. With hierarchical BEV feature queries, texture features from different elevations tend to overlap. In contrast, elevation-guided BEV feature queries not only retrieve more accurate texture features but also significantly reduce the computational burden, requiring only one query per grid feature. The experimental results underscore the enhancement in texture reconstruction results facilitated by elevation guidance.

5 Conclusion

We propose BEV-GS, a novel real-time single-frame road surface reconstruction method based on feed-forward Gaussian splatting. By introducing a feed-forward geometric and texture prediction network based on BEV representation, our method directly predicts road elevation and Gaussian parameters from a single image. Extensive experiments across various complex road surfaces demon-

strate that our method efficiently and accurately reconstructs road geometry and texture from a single image, while also providing high-quality novel-view synthesis. It has great potential for online applications like road condition preview and offline tasks such as autonomous driving testing. Despite the outstanding performance, the validation is constrained by the dataset, in which all experimental data were collected under sunny conditions with favorable illumination. In future work, we plan to extend our method to address extreme scenarios, including adverse weather and low-light nighttime conditions.

Acknowledgements

This work was supported in part by National Key R&D Program of China (Grant No.2024YFB4708900), the Natural Science Foundation of China under Grant 62225309, U24A20278, and 62361166632.

References

1. Agarwal, A., Arora, C.: Attention attention everywhere: Monocular depth prediction with skip attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5861–5870 (January 2023)
2. Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023)
3. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2024)
4. Chen, H., Li, C., Lee, G.H.: Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance. arXiv preprint arXiv:2312.00846 (2023)
5. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European Conference on Computer Vision. pp. 370–386. Springer (2024)
6. Deng, T., Wang, N., Wang, C., Yuan, S., Wang, J., Wang, H., Wang, D., Chen, W.: Incremental joint learning of depth, pose and implicit scene representation on monocular camera in large-scale scenes. IEEE Transactions on Automation Science and Engineering (2025)
7. DeVries, T., Bautista, M.A., Srivastava, N., Taylor, G.W., Susskind, J.M.: Unconstrained scene generation with locally conditioned radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14304–14313 (2021)
8. Fan, L., Yang, Y., Li, M., Li, H., Zhang, Z.: Trim 3d gaussian splatting for accurate geometry representation. arXiv preprint arXiv:2406.07499 (2024)
9. Fan, R., Jiao, J., Pan, J., Huang, H., Shen, S., Liu, M.: Real-time dense stereo embedded in a uav for road inspection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 535–543 (2019). <https://doi.org/10.1109/CVPRW.2019.00079>
10. 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 (2022)

11. Farooq Bhat, S., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4008–4017 (2021)
12. Feng, Y., Yu, C., Ran, F., Yang, Z., Liu, Y.: Exploregs: a vision-based low overhead framework for 3d scene reconstruction. In: 2025 IEEE International Conference on Real-time Computing and Robotics (RCAR). pp. 509–514. IEEE (2025)
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. 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)
15. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
16. Jia, B., Chen, J., Zhang, K.: Drivable road reconstruction for intelligent vehicles based on two-view geometry. *IEEE Transactions on Industrial Electronics* **64**(5), 3696–3706 (2017). <https://doi.org/10.1109/TIE.2016.2644601>
17. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. arXiv preprint arXiv:2505.23716 (2025)
18. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
19. Kulhanek, J., Sattler, T.: Tetra-nerf: Representing neural radiance fields using tetrahedra. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18458–18469 (2023)
20. Kwon, O., Park, J., Oh, S.: Renderable neural radiance map for visual navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9099–9108 (2023)
21. Li, X., Lu, Z., Zeng, C.: Robotic and intelligent technologies in composite material inspection: a review. *Robot Learning* **1**(1) (2024). <https://doi.org/10.55092/r120240005>
22. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022)
23. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18039–18048 (June 2024)
24. 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* pp. 1–13 (2024)
25. 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)
26. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* **41**(4), 1–15 (2022)
27. Ni, T., Li, W., Zhao, D., Kong, Z.: Road profile estimation using a 3d sensor and intelligent vehicle. *Sensors* **20**(13), 3676 (2020)
28. Ost, J., Mannan, F., Thurey, N., Knodt, J., Heide, F.: Neural scene graphs for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2856–2865 (2021)

29. Park, C., Seo, E., Lim, J.: Heightlane: Bev heightmap guided 3d lane detection. arXiv preprint arXiv:2408.08270 (2024)
30. Piccinelli, L., Sakaridis, C., Yu, F.: idisc: Internal discretization for monocular depth estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
31. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
32. Song, M., Lim, S., Kim, W.: Monocular depth estimation using laplacian pyramid-based depth residuals. IEEE Transactions on Circuits and Systems for Video Technology **31**(11), 4381–4393 (2021)
33. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2022)
34. Szymanowicz, S., Insaftudinov, E., Zheng, C., Campbell, D., Henriques, J.F., Rupprecht, C., Vedaldi, A.: Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. In: 2025 International Conference on 3D Vision (3DV). pp. 670–681. IEEE (2025)
35. Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter image: Ultra-fast single-view 3d reconstruction. Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
36. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114 (2019)
37. Theunissen, J., Tota, A., Gruber, P., Dhaens, M., Sorniotti, A.: Preview-based techniques for vehicle suspension control: a state-of-the-art review. Annual Reviews in Control **51**, 206–235 (2021)
38. 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)
39. Wang, Y., Wang, W., Liu, J., Chen, T., Wang, S., Yu, B., Qin, X.: Framework for geometric information extraction and digital modeling from lidar data of road scenarios. Remote Sensing **15**(3) (FEB 2023)
40. Wolf, Y., Bracha, A., Kimmel, R.: GS2Mesh: Surface reconstruction from Gaussian splatting via novel stereo views. In: European Conference on Computer Vision (ECCV) (2024)
41. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
42. Wu, W., Guan, H., Liu, Z., Wang, H.: Weathercity: Urban scene reconstruction with controllable multi-weather transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40949–40958 (2026)
43. Wu, W., Wang, Q., Wang, G., Wang, J., Zhao, T., Liu, Y., Gao, D., Liu, Z., Wang, H.: Emie-map: Large-scale road surface reconstruction based on explicit mesh and implicit encoding. In: European Conference on Computer Vision. pp. 370–386. Springer (2024)
44. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depth-splat: Connecting gaussian splatting and depth. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16453–16463 (2025)
45. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024)

46. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4578–4587 (2021)
47. Yu, Z., Wang, H., Yang, J., Wang, H., Xie, Z., Cai, Y., Cao, J., Ji, Z., Sun, M.: Sgd: Street view synthesis with gaussian splatting and diffusion prior. *arXiv preprint arXiv:2403.20079* (2024)
48. Yuan, H., Chen, T., Sui, W., Xie, J., Zhang, L., Li, Y., Zhang, Q.: Monocular road planar parallax estimation. *IEEE Transactions on Image Processing* **32**, 3690–3701 (2023)
49. Zhang, J., Zhan, F., Xu, M., Lu, S., Xing, E.: Fregs: 3d gaussian splatting with progressive frequency regularization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21424–21433 (2024)
50. Zhao, T., Guo, P., He, J., Wei, Y.: A hierarchical scheme of road unevenness perception with lidar for autonomous driving comfort. *IEEE Transactions on Intelligent Vehicles* **9**(1), 2439–2448 (2024)
51. 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)
52. Zhao, T., Yang, L., Xie, Y., Ding, M., Tomizuka, M., Wei, Y.: Roadbev: Road surface reconstruction in bird’s eye view. *IEEE Transactions on Intelligent Transportation Systems* **25**(11), 19088–19099 (2024)
53. Zhou, T., Sun, C., Fu, H.: Road information extraction from high-resolution remote sensing images based on road reconstruction. *Remote Sensing* **11**(1) (2019)
54. Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21634–21643 (2024)