EAGS: Error-Aware Gaussian Splatting with Dual-Confidence-Guided Modeling for Uncalibrated Driving Scenes

Zhigang Cen[1], Ningyan Guo[1]✉, Yulan Guo[2], Yifan Ge[1], and Zhiyong Feng[1]

[1] Beijing University of Posts and Telecommunications, Beijing 100876, China
{cenzhigang_0217,guoningyan, geyifan, fengzy}@bupt.edu.cn
[2] Sun Yat-sen University, Shenzhen 518707, China guoyulan@sysu.edu.cn

Abstract. Feed-forward 3D reconstruction for autonomous driving has demonstrated strong performance and promising generalization capabilities. However, pixel-level methods often suffer from multi-view inconsistencies and content duplication. While voxel-level approaches alleviate these issues by predicting 3D Gaussian primitives in voxel space, they remain limited by resolution and computational overhead. We introduce an efficient feed-forward framework for scene reconstruction from uncalibrated images. The key insight is the explicit decoupling of instantaneous generation necessity and long-term geometric persistence during Gaussian generation. Specifically, we introduce Generation Confidence driven by reprojection errors to enable demand-aware Gaussian expansion, and Reuse Confidence constructed from depth uncertainty and temporal decay to model the long-term stability. To balance fidelity and sparsity, we employ differentiable voxelization to aggregate neighboring features and suppress spatial redundancy, followed by confidence-aware soft-gating pruning. Gaussian distillation is further introduced to enable a compact set of backbone Gaussians to approximate the full scene representation. Extensive experiments on multiple datasets demonstrate that our method achieves strong reconstruction quality and robustness while maintaining a highly compact scene representation.

Keywords: Feed-forward 3DGS · Instantaneous Generation Necessity · Long-term Geometric Persistence

1 Introduction

In the cutting-edge fields of artificial intelligence and autonomous driving, the digital perception and modeling of 3D space [17, 18, 31, 37] has become foundational infrastructure that empowers machines with autonomous interaction, precise navigation, and environmental understanding. This capability is widely applied in robotics [15,47], virtual reality [8,14], gaming, film production [11,24],

✉ Corresponding author.

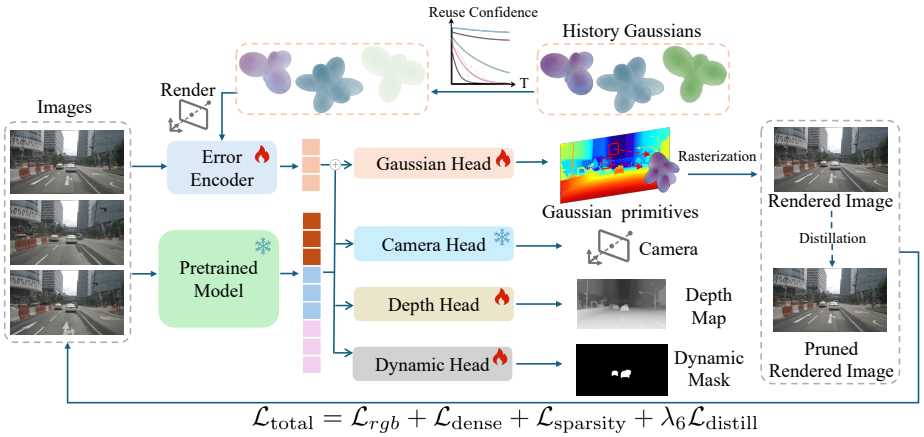

Fig. 1: Overview of EAGS. EAGS estimates poses, depth, and masks from uncalibrated images via a vision transformer. A Gaussian head predicts attributes and time-decayed reuse confidence. Following first-frame initialization, the model performs on-demand sparse generation in subsequent frames. This ensures that Gaussians are sparsely generated, each with a clear reconstruction gain. Finally, differentiable voxelization and soft gating prune redundancies to maintain a compact and stable scene representation.

and autonomous driving [7, 26]. In recent years, reconstruction methods represented by Neural Radiance Fields (NeRF) [22] and 3D Gaussian Splatting (3DGS) [16] have achieved remarkable rendering fidelity. However, these approaches typically necessitate per-scene optimization and rely heavily on calibrated camera poses, which significantly constrains their scalability.

Recent feedforward approaches [2, 5, 13, 41] leverage CNN and Transformer [33] architectures to predict 3D Gaussian primitives from 2D reference images, showing great potential. For instance, methods such as DGGT [4] and DrivingForward [31] employ a pixel-wise paradigm, mapping image features directly into Gaussian representations to reconstruct the scene. However, such approaches exhibit a "generate-everything" bias, repeatedly sampling overlapping regions across different time steps. This leads to a proliferation of floating artifacts and severe temporal inconsistencies during rendering, as the model lacks memory of previously reconstructed geometry. Furthermore, generating a dense set of Gaussians for every frame imposes a substantial computational overhead for both storage and downstream processing. To address these issues, VolSplat [35] and EVolSplat [21] aggregate information in a canonical voxel space for joint geometry–appearance reasoning. However, voxel-based methods are limited by resolution and incur high computational cost.

As illustrated in Fig. 1, we propose a Dual-Confidence-Guided Error-Aware Gaussian Splatting framework (EAGS) to explicitly decouple instantaneous generation necessity from long-term geometric persistence for efficient scene reconstruction. Following VGGT [34], our framework employs a transformer-based backbone built upon DINOv2 [23] and Alternating Attention to extract and

fuse spatiotemporal features from multi-frame image sequences. Based on shared features, several prediction branches estimate camera parameters, depth maps, dynamic masks, and Gaussian attributes. In addition, an error encoder projects historical backbone Gaussians onto the current view to capture photometric discrepancies, explicitly identifying under-reconstructed regions.

To decouple instantaneous generation necessity from long-term geometric persistence, we introduce a dual-confidence mechanism. First, Generation Confidence $\mathcal{C}_g \in \mathbb{R}_+^{H\times W}$ is driven by reprojection errors and determines which regions in the current frame require additional Gaussian primitives, enabling demand-aware expansion. Second, we derive the reuse confidence $\mathcal{C}_r^t \in \mathbb{R}_+^{H\times W}$ by integrating the generation confidence with depth uncertainty $\delta_d \in \mathbb{R}_+^{H\times W}$ and a temporal decay mechanism, thereby assessing the stability and long-term contribution of Gaussian primitives during cross-frame propagation. During reconstruction, $\mathcal{C}_g$ is binarized to achieve demand-aware Gaussian generation while suppressing redundant primitives. Meanwhile, primitives with lower $\mathcal{C}_r^t$ undergo faster temporal decay, enabling a clear separation between stable backbones and transient structures. For pruning, we adopt differentiated strategies. In static regions, differentiable voxelization and soft gating guided by $\mathcal{C}_r^t$ suppress spatial redundancy. In dynamic regions, considering the stochastic and sparse nature of dynamic objects, we apply voxelization only within a single frame to avoid cumulative drift caused by long-term tracking. Finally, a Gaussian Distillation Loss mitigates the performance gap before and after pruning, enabling a compact scene representation with a minimal set of backbone Gaussians. Extensive experiments on multiple benchmark datasets demonstrate that EAGS consistently achieves superior and robust reconstruction compared with state-of-the-art feed-forward and optimization-based approaches.

In summary, our main contributions are as follows:

- We introduce EAGS, a novel feed-forward framework that explicitly decouples instantaneous generation necessity from long-term geometric persistence for efficient scene reconstruction.
- We propose an error-driven, on-demand sparse generation mechanism integrated with a dual-confidence management strategy, ensuring multi-view consistency while maintaining a minimalist scene representation.
- Extensive experiments on multiple datasets demonstrate that EAGS achieves superior reconstruction performance and compactness, while exhibiting promising zero-shot generalization capabilities.

2 Related Work

2.1 Optimization-based Novel View Synthesis Methods

Traditional novel view synthesis methods [12, 16, 17, 22, 39] rely on precise geometric correspondences and densely sampled points, which limits their applicability in real-world scenarios. NeRF [22] learns a continuous, implicit representation of a scene and has achieved impressive results. However, NeRF-based

approaches [10, 19, 32, 43] suffer from extremely slow rendering speeds, making them unsuitable for real-time applications. To address this limitation, 3DGS represents scenes explicitly using millions of anisotropic Gaussians and performs differentiable rasterization to enable high-resolution photorealistic rendering in real time. Subsequent works [12, 17] further extend 3DGS to dynamic scenes by introducing deformation models. Despite these advances, most existing NeRF- and 3DGS-based methods assume access to accurate camera poses and rely on computationally expensive per-scene optimization, which restricts their scalability and generalization to unseen environments.

In contrast, the proposed method enables efficient feed-forward reconstruction without relying on costly scene-specific optimization, while maintaining real-time rendering performance and strong cross-scene generalization capability

2.2 Feed-Forward 3D Gaussian Splatting

Feedforward 3DGS approaches [3, 9, 44, 48] extract image features via neural networks, back-project them into 3D space using pixel depths, and decode corresponding pixel features into Gaussian attributes such as color, opacity, scale, and rotation. The paradigm allows 3D Gaussians to be predicted in a single forward pass through the network. Most existing feedforward Gaussian methods [4, 5, 13, 27, 31, 41] rely on pixel alignment as the fundamental mechanism to associate image features with Gaussians, maintaining a one-to-one correspondence between pixels and Gaussians. For instance, pixelSplat [2] uses an epipolar Transformer to estimate scene scale factors and combines them with depth to predict Gaussians. DrivingForward [31] jointly trains pose estimation, monocular depth prediction, and a Gaussian network to model driving scenes. MVSplat [5] introduces a cost volume fusion strategy to enhance view consistency. Flash3D [29] and DepthSplat [41] extend the reconstruction to full-scene levels, while NoPoSplat [46] reconstructs 3D scenes from sparse multi-view images without pose supervision. DGGT [4] reconstructs 4D dynamic scenes using lifecycle and motion interpolation. Although these methods rely on pixel-aligned strategies to predict Gaussian primitives, such designs often result in excessive Gaussian redundancy and inconsistency across viewpoints due to their pixel-level alignment nature. To address this, VolSplat [35] lifts 2D features into a 3D voxel grid to reduce spatial redundancy. AnySplat [13] performs pixel-wise predictions followed by voxelization-based pruning to eliminate redundancy. Long-lrm balances quality and efficiency through token merging and Gaussian pruning. EVolSplat [21] accumulates monocular depth into a point cloud to initialize the model and decodes it into Gaussian primitives using a 3D CNN.

However, these voxel- or point-based approaches incur high computational complexity due to 3D convolutions and are sensitive to scale variation, while often neglecting depth reliability. In contrast, the proposed method enables efficient and reliable scene reconstruction and real-time rendering directly from image sequences without requiring camera poses, by leveraging projection error and a dual-confidence mechanism.

3 Method

Our key idea is to address "when to generate" and "what to keep", instead of blindly increasing Gaussians. Built on the original VGGT backbone, the error encoder provides pixel-level gain signals for generation, the dual-confidence mechanism filters Gaussians to maintain efficiency, and Gaussian distillation concentrates effective information into a small number of key Gaussians. For "when to generate", the error-aware mechanism can be incorporated into other models as an interpretable generation trigger and extended to uncertainty-aware generation. For "what to keep", dual-confidence and Gaussian distillation can generalize to pruning methods for better reconstruction quality.

In the following sections, we begin by formalizing our problem setup in Sec. 3.1. Sec. 3.2 elaborates on the core components, including the feature extraction architecture, the prediction heads, and the Gaussian generation and management pipeline. Sec. 3.4 describes our training methodology.

3.1 Preliminary

Consider a single scene with a set of N uncalibrated input images $\mathcal{I} = \{I_i\}_{i=1}^N$, where $I_i \in \mathbb{R}^{H\times W\times 3}$. The proposed model extracts 2D features from these uncalibrated images and predicts their corresponding camera poses $\mathcal{P} = \{P_i\}_{i=1}^N$ and depth maps $\mathcal{D} = \{D_i\}_{i=1}^N$. These predictions are then used to jointly reconstruct the scene geometry and surface appearance by generating a set of anisotropic Gaussians $\mathcal{G} = \{(\mu_i, \alpha_i, \Sigma_i, c_i, \mathcal{C}_g, \mathcal{C}_r^t)\}_i^G$, with attributes: a mean position μ_i, an opacity α_i, a covariance matrix Σ_i, a generation confidence $\mathcal{C}_g$, a reuse confidence $\mathcal{C}_r^t$ and view-dependent color c_i (computed by spherical harmonics). During rendering, Gaussian primitives are first sorted in parallel according to their depth relative to the viewpoint. The final opacity α^t is then obtained by multiplying α and $\mathcal{C}_r^t$. Finally, real-time differentiable rendering is achieved through alpha blending:

$$c = \sum_{i\in G} c_i \alpha_i^t \prod_{j=1}^{i-1}(1-\alpha_j^t). \tag{1}$$

3.2 Geometry-Aware Scene Encoding

Feature Extraction. We adopt a Geometry Transformer [34] as the feature extraction network to encode a sequence of image frames into a shared latent space. Specifically, given N uncalibrated images, each image is transformed using DINOv2 [23] into a sequence of $\frac{HW}{P^2}$ tokens of dimension d, where H and W are the image height and width, and P is the patch size. To each image's token sequence $t_i^I \in \mathbb{R}^{L\times d}$, we additionally introduce a learnable camera token $t_i^g \in \mathbb{R}^{1\times d}$ and four register tokens $t_i^R \in \mathbb{R}^{4\times d}$. These multi-view features are then processed through an Alternating Attention mechanism to produce the final feature representation F_{backbone}. This mechanism captures both intra-frame

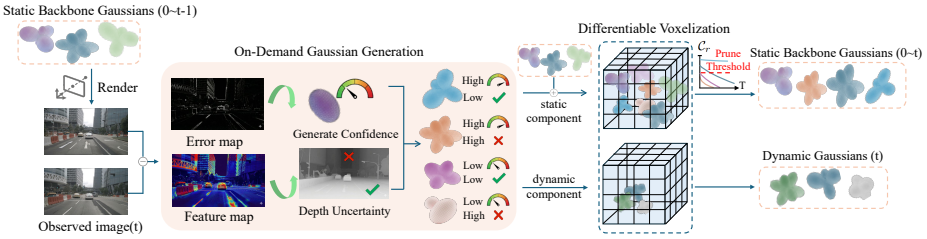

Fig. 2: On-demand Gaussian generation. Temporally decayed backbone Gaussians are reprojected into the current frame, where Gaussian attributes and generation confidence are predicted using 2D features and photometric residuals. Static components are voxelized and pruned via soft gating to update the scene backbone, while dynamic components are voxel-wise aggregated for efficient scene representation.

and inter-frame interactions, effectively integrating information across different views.

Camera Pose Estimation. Camera pose estimation is crucial for extending sequence length, enabling cross-dataset deployment, and supporting novel view geometry reconstruction. Following prior works [4, 13, 34], camera parameters are predicted from the refined camera token using four additional self-attention layers followed by a linear projection layer. The first camera coordinate system is set as the reference frame, and thus its extrinsic parameters are fixed as the identity transformation, i.e., the initial rotation quaternion is $q_1 = [0, 0, 0, 1]$, and the initial translation vector is $t_1 = [0, 0, 0]$.

Depth and Dynamic Dense Prediction Heads. We leverage the depth maps and estimated camera intrinsics and extrinsics to compute the spatial centers of Gaussian primitives via back-projection. To enable differentiated reconstruction of static and dynamic regions, we introduce a dynamic mask to decouple the scene. Specifically, $F_{backbone}$ is processed by a Dense Prediction Transformer (DPT) [25] layer to generate dense depth maps, a dynamic mask $\mathcal{M}_{dy}$ and a static mask $\mathcal{M}_{st} = 1 - \mathcal{M}_{dy}$. Based on this, the scene is decomposed into a static component $G_s = G \odot \mathcal{M}_{st}$ and a dynamic component $G_d = G \odot \mathcal{M}_{dy}$. Additionally, the depth network also predicts a depth confidence $C_{depth} \in \mathbb{R}_{+}^{H \times W}$ to characterize the uncertainty of depth estimation.

3.3 On-Demand Gaussian Generation and Confidence-Guided Pruning

Error-Aware Gaussian Generation. This component is jointly implemented by the depth map, error encoder and the Gaussian head, as illustrated in Fig. 2. The error encoder is designed to extract photometric discrepancies between the rendered outputs of existing Gaussians and the target observation at the current time step. These residuals are subsequently mapped into high-dimensional features to represent reconstruction deficiencies across different regions. The Gaussian head is designed to leverage the joint features of reprojection residuals,

DINO features, and F_{backbone} to adaptively infer whether new Gaussian primitives are needed in each region, and to regress their corresponding geometric and appearance attributes. Specifically, we render the previously constructed Gaussian primitives from the current viewpoint and compute the photometric residual between the observed $\hat{I}$ and rendered images I_{hist} using the L_1 norm and the Sobel gradient operator ∇_s, formulated as:

$$E = \|I_{\text{hist}} - \hat{I}\|_1 + \|\nabla_s I_{\text{hist}} - \nabla_s \hat{I}\|_1. \tag{2}$$

An error encoder with three 3×3 convolutions then transforms this error into a feature F_{error}. The DPT [25] layer integrates F_{error}, F_{backbone}, and image features to generate dense predictions, yielding both the Gaussian attribute maps and the associated generation confidence $\mathcal{C}_g$. To suppress redundant generation, we encourage the $\mathcal{C}_g$ to clearly differentiate between valid and invalid Gaussian primitives by driving its values toward a binary state. To achieve this, we introduce a sparsity penalty in the optimization objective, which enforces a bimodal distribution on the generation confidence map.

The validity mask is designed to help distinguish the first frame from subsequent frames during Gaussian generation. For the first frame, all regions are expected to generate corresponding Gaussian points to construct the initial scene. For the following frames, Gaussian primitives should be generated on demand based on the error features. To this end, we incorporate the validity mask into the last layer of the DPT module to guide the model in distinguishing the two patterns. Specifically, for the first frame, the generation confidence and validity mask are set to $\mathcal{C}_g = 1$ and $\mathcal{M}_{\text{valid}} = 0$, respectively. For all subsequent frames, we set $\mathcal{M}_{\text{valid}} = 1$.

Dual-confidence Mechanism. Existing feedforward reconstruction approaches typically retain all generated primitives without differentiation, which inherently conflates the objectives of transient appearance fitting and persistent geometric reconstruction. To address this, a dual-confidence mechanism is introduced to decouple these competing objectives.

Specifically, the $\mathcal{C}_g$ quantifies the necessity of capturing the current appearance based on the magnitude of reprojection residuals. It is inversely related to the ability of existing Gaussians to represent the current region. To extract a stable persistent scene backbone from transient compensations and eliminate redundant or outlier noise, we propose a reuse confidence $\mathcal{C}_r^t$ that integrates both geometric and temporal cues. First, $\mathcal{C}_g$ is coupled with depth confidence to enable geometry-aware modulation, formulated as $\mathcal{C}_g \cdot (\gamma + (1-\gamma) \cdot \mathcal{C}_{depth})$. This design aims to separate stable backbone primitives from temporary ones used for short-term fitting. Over time, Gaussians with high reuse confidence decay slowly and preserve consistent scene structures. In contrast, those with low reuse confidence decay rapidly to prevent visual artifacts across different views. Based on this principle, the final reuse confidence is defined as:

$$\begin{gathered} \sigma = \sigma_{min} + (1 - \mathcal{C}_r^{t_0}) \cdot (\sigma_{max} - \sigma_{min}), \\ \mathcal{C}_r^t = e^{-\sigma \cdot (t - t_0)}, \end{gathered} \tag{3}$$

where $\mathcal{C}_r^{t_0}$ denotes the initial reuse confidence, t_0 and t represent the timestamp of Gaussian creation and the current frame, respectively. The parameters $\sigma_{\min}$ and $\sigma_{\max}$ control the lifespan of a Gaussian, i.e., $\sigma \in [\sigma_{\min}, \sigma_{\max}]$.

Accordingly, Gaussian primitives are categorized into four groups: stable backbone primitives (High $\mathcal{C}_g$, Low δ_d), transient primitives for error compensation (High $\mathcal{C}_g$, High δ_d), redundant primitives in well-reconstructed regions (Low $\mathcal{C}_g$, Low δ_d), and noise-related outliers (Low $\mathcal{C}_g$, High δ_d). Backbone primitives decay slowly across multiple frames to maintain a stable scene representation, while transient primitives undergo rapid decay and are gradually replaced by their corresponding backbone primitives. Meanwhile, redundant primitives and noisy points are removed. We decay the opacity of Gaussian primitives using time-varying confidence scores to constrain their scope of influence. Compared to DGGT [4] and STORM [42], the proposed method more reliably and efficiently removes floating artifacts and preserves the scene backbone, thereby providing a multi-view consistent scene representation.

Gaussian Pruning. To fuse multi-granular descriptions of the same object across temporal sequences and suppress spatial redundancy, Gaussian merging is implemented via differentiable voxelization. To further remove invalid regions and improve robustness against resolution changes, we apply a soft gating strategy for refined pruning. In this process, Gaussian primitives are organized into different local clusters V. For each cluster, we compute a normalized weight w based on the $\mathcal{C}_r^t$, allowing the model to identify and retain only the most reliable primitives. Finally, the aggregated new Gaussian features (before activation) are obtained as a weighted sum:

$$F_{g_i}^{'} = \sum_{g_i \in V} \frac{\exp(\mathcal{C}_{g_i \to r}^t)}{\sum_{g_j \in V} \exp(\mathcal{C}_{g_j \to r}^t)} F_{g_i}, \tag{4}$$

where $g_i \to r$ denotes the reuse confidence of the i-th Gaussian within the local cluster V.

To enable end-to-end optimization of the pruning process and bypass the instability associated with non-differentiable hard thresholding, we employ a steep sigmoid function as a differentiable soft gating mechanism. Formally, the pruning mask is defined as $\mathcal{M}_{\text{prune}} = \text{Sigmoid}\left(s \cdot (\mathcal{C}_r^t - \tau)\right)$, where $s = 5$ is a scaling factor controlling the steepness of the transition, and $\tau = 0.7$ is the pruning threshold. As s increases, this function approximates the Heaviside step function while maintaining differentiability. To guide the network toward representing the scene with a sparse set of high-confidence, multi-view consistent Gaussians, we introduce both a distillation loss and a sparsity regularization term. During inference, this step is replaced with hard pruning based on a fixed threshold.

3.4 Training

Reconstruction Loss. For the reconstruction objective, we employ an L_1 loss and the LPIPS perceptual loss [49] computed between the rendered image I_{full}

and the observed image $\hat{I}$.The RGB reconstruction loss can be formulated as

$$\mathcal{L}_{\text{rgb}} = \mathcal{L}_{l_1} + \lambda_1 \mathcal{L}_{\text{LPIPS}}. \tag{5}$$

Dense Prediction Loss. For the dynamic head, we supervise the prediction of dynamic maps using a binary cross-entropy loss $\mathcal{L}_{\text{dynamic}} = \text{BCE}(M_{dy}, \hat{M}_{dy})$, where M_{dy} denotes the predicted dynamic map and $\hat{M}_{dy}$ represents the corresponding ground-truth (GT). Following VGGT [34], we fine-tune the depth head using the regression loss $\mathcal{L}_{\text{reg}}$. The dense prediction loss can be formulated as:

$$\mathcal{L}_{dp} = \lambda_2 \mathcal{L}_{dynamic} + \lambda_3 \mathcal{L}_{reg}. \tag{6}$$

Sparsity Loss. For the generation confidence $\mathcal{C}_g$, we introduce a binarization constraint to prevent the network from converging to ambiguous intermediate states, and apply L_1 regularization to avoid degeneration into dense pixel-level Gaussian generation. Furthermore, we introduce a geometry-aware pruning regularizer, defined as $L_{gp} = \mathbb{E}[\mathcal{C}_g(i,j) \cdot \mathcal{C}_{depth}(i,j)]$. This term guides the network to prioritize the removal of redundant Gaussians in regions where depth estimates are more reliable. Notably, the gradients for depth confidence are truncated, ensuring that only the confidence values are updated during optimization.

$$\mathcal{L}_{\text{sparsity}} = \lambda_4 \mathbb{E}[\mathcal{C}_g(1 - \mathcal{C}_g)] + \lambda_5 \mathbb{E}[\mathcal{C}_g(i,j) \cdot \text{sg}(\mathcal{C}_{depth}(i,j))], \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator.

Gaussian Distillation Loss. To balance reconstruction quality and representation compactness, we introduce a Gaussian distillation loss. By enforcing the rendered result of the pruned salient subset I_{prune} to approximate that of the full set I_{full}, the model is encouraged to concentrate meaningful texture and geometric information into a small number of key Gaussians. This ensures that the pruned components contribute minimally to the final rendered image. This process can be formulated as:

$$\mathcal{L}_{\text{distill}} = \frac{\sum_{i,j} \mathcal{M}_{st}(i,j) \cdot \left\| \text{sg}(I_{\text{full}}(i,j)) - I_{\text{prune}}(i,j) \right\|_1}{3 \cdot \sum_{i,j} \mathcal{M}_{st}(i,j) + \epsilon}, \tag{8}$$

where ϵ is a small constant for numerical stability. The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{rgb} + \mathcal{L}_{\text{dp}} + \mathcal{L}_{\text{sparsity}} + \lambda_6 \mathcal{L}_{\text{distill}}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Datasets. We train our model on the Waymo Open Dataset [28] and additionally evaluate its zero-shot generalization capability on the nuScenes [1] and

Table 1: Quantitative comparisons. " PGS " stands for average number of per-view Gaussians; " N/A " indicates not applicable; " - "indicates unavailable data.

Method	Rendering Quality			Pos-free	PGS	Inference times(s)
	PSNR↑	SSIM↑	D-RMSE↓			
EmerNeRF [43]	24.51	0.738	33.99	✗	N/A	14min
3DGS [16]	25.13	0.741	19.68	✗	N/A	23min
PVG [6]	22.38	0.661	13.01	✗	N/A	27min
DeformableGS [45]	25.29	0.761	14.79	✗	N/A	29min
LGM [30]	18.53	0.447	9.07	✗	152292	**0.06s**
MVSplat [5]	20.56	0.697	10.13	✗	152292	0.08s
NoPoSplat [46]	24.31	0.751	9.08	✓	152292	23.22s
DepthSplat [41]	23.26	0.696	10.05	✗	152292	0.11s
STORM [42]	26.38	0.794	5.48	✗	152292	0.18s
EVolSplat [21]	23.34	0.68	-	✗	399074	0.10s
UniSplat [27]	27.82	0.83	-	✗	103984	0.18s
DGGT [4]	27.41	0.846	3.47	✓	152292	0.39s
VGGT++ [34]	22.50	0.749	3.80	✓	152292	0.24s
Ours	**28.02**	**0.849**	**3.44**	✓	**49556**	0.16s

Argoverse2 datasets [38]. The full Waymo Open Dataset comprises 1,000 driving log sequences, with the training split containing 798 scenes, each consisting of approximately 190 to 200 frames, and the test split including 202 scenes that cover challenging scenarios. To assess our method, we follow standard evaluation protocols using peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [36], and feature-level Learned Perceptual Image Patch Similarity (LPIPS) [49] for images, as well as root mean square error for depth estimation (D-RMSE). For rendering efficiency, we report frames per second (FPS) during inference on the same hardware to ensure fair comparison.

Implementation Details. EAGS is implemented in PyTorch and optimized using the AdamW [20] optimizer with a cosine learning rate schedule. The remaining heads are trained with a learning rate of 5×10^{-5}. The loss function is composed of four terms, weighted by $\lambda_1 = 0.05$, $\lambda_2 = 0.05$, $\lambda_3 = 0.0001$, $\lambda_4 = 0.002$, $\lambda_5 = 0.005$ and $\lambda_6 = 0.25$. The γ, σ_{min} and σ_{max} are set to 0.6, 0.05 and 1.5, respectively. During training, we randomly sample $L = 4$ input images, spaced by an interval of 5 frames. The ground-truth dynamic masks are refined using SegFormer [40]. We resize training and test images to 518×294. The model is trained for 4 epochs using two NVIDIA RTX 4090 GPUs. For evaluation, all results are reported on a single NVIDIA A100 GPU to ensure fairness.

Baselines. We benchmark EAGS against a comprehensive suite of state-of-the-art methods, categorized into three paradigms: (1) Per-scene optimization approaches, including EmerNeRF [43], 3DGS [16], PVG [6], and DeformableGS [45]; (2) Pixel-wise feed-forward models, such as LGM [30], MVSplat [5], No-

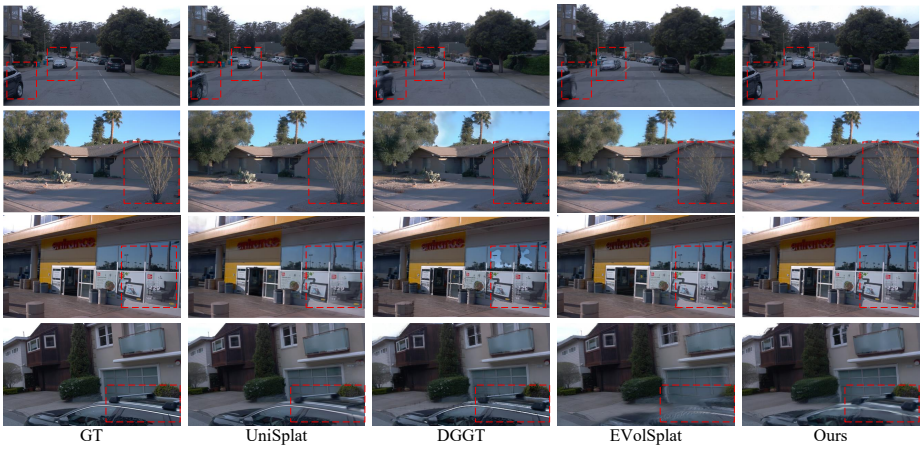

Fig. 3: Qualitative results on the Waymo [28] dataset. Compared with baseline methods that exhibit ghosting, inconsistencies, and blur, our approach maintains high-fidelity reconstruction quality across multiple scenes and viewpoints.

PoSplat [46], DepthSplat [41], STORM [42] and VGGT++ [34]; (3) Voxel-based approaches, represented by EVolSplat [21], UniSplat [27] .

4.2 Experimental Results and Analysis

Comparisons with SOTA methods. Tab. 1 and Fig. 3 present quantitative and qualitative comparisons. Following DGGT, we train on the full Waymo Open Dataset training split with 798 scenes and evaluate on the test split with 202 scenes. We use 4 input frames for reconstruction and assess novel view synthesis on interpolation-generated intermediate frames, with metrics computed over all frames. The experimental results reveal that existing pixel-level feed-forward methods tend to produce a large number of redundant Gaussians, where overlapping and mutually interfering primitives adversely affect rendering quality. Voxel-based approaches partially mitigate this issue by aggregating multi-view information within voxel grids. However, they remain fundamentally limited by grid resolution and do not fully resolve representation redundancy. In contrast, EAGS achieves state-of-the-art rendering quality while requiring substantially fewer Gaussians for scene representation. This improvement stems from our model leveraging previously generated Gaussians as prior information, enabling the sparse generation of new Gaussians on demand based on existing reconstruction results. This not only ensures the compactness of the scene representation but also guarantees that every newly generated Gaussian contributes effectively to the reconstruction quality.

Zero-shot Inference. To evaluate the generalizability, we evaluate various methods on the unseen dataset, with the results summarized in Tab. 2. All models are trained on the Waymo dataset without fine-tuning. Under this zero-

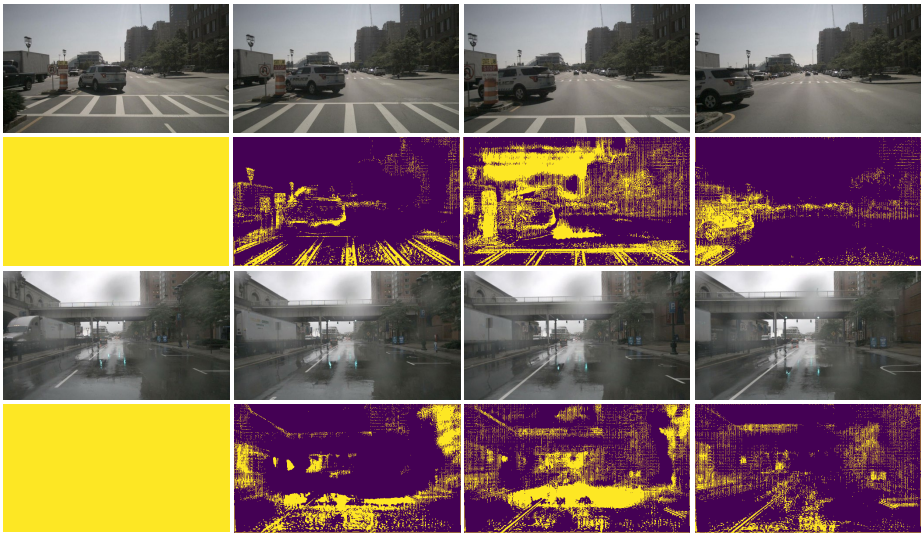

Fig. 4: Visualization of generation confidence maps. Frames are shown in chronological order from left to right. The Viridis colormap is used, where yellow indicates Gaussian generation and purple denotes regions without Gaussians.

shot setting, the proposed method achieves the best performance, demonstrating strong cross-dataset generalization. While DGGT also produces competitive results due to its strong backbone and pose-free design, our method shows superior generalization by effectively exploiting spatio-temporal correlations among Gaussians across consecutive frames.

Visualization of Generation Confidence. To validate the "on-demand generation" design philosophy of the proposed scheme, we visualize the binarized generation confidence maps across four consecutive frames, as illustrated in Fig. 4. For the initial frame of the sequence, the model is forced to generate all candidate Gaussians to initialize the scene, due to the absence of previously generated Gaussians as geometric priors. In subsequent frames, the model only performs supplementary generation based on the deficiencies of the established scene, resulting in highly sparse confidence maps. Notably, we observe that the third frame provides detailed supplements relative to the second frame. This can be attributed to the rapid decay mechanism for low-confidence Gaussians, as well as new generation requirements arising from the enhanced resolution of object details as the viewpoint changes.

Robustness Analysis. The results are shown in Fig. 5 and Fig. 6. These visualizations demonstrate the adaptability and robustness of our method in complex dynamic scenarios. When dynamic occlusions occur, background regions that are initially hidden become observable over time as foreground objects move. By aggregating these complementary temporal observations, our method progressively reconstructs previously occluded background areas. Moreover, when both

Table 2: Quantitative Comparison under Zero-Shot Settings on nuScenes [1] and Argoverse2 [38] datasets.

Method	nuScenes			Argoverse2		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MVSplat [5]	17.84	0.563	0.451	18.67	0.647	0.304
NoPoSplat [46]	19.75	0.545	0.394	22.00	0.646	0.237
DepthSplat [41]	19.52	0.601	0.376	22.05	0.636	0.280
STORM [42]	17.77	0.669	0.394	20.83	0.542	0.326
DGGT [4]	25.31	0.794	**0.152**	26.34	0.812	0.155
Ours	**26.14**	**0.813**	0.159	**27.41**	**0.825**	**0.150**

the ego vehicle and surrounding vehicles are moving, our method can faithfully reconstruct the scene and remove truly dynamic objects, even in the presence of some false-positive dynamic predictions where static regions are incorrectly labeled as dynamic and not retained as persistent anchors. In the FoV extrapolation setting, although slight viewpoint discrepancies may exist because both methods are pose-free, our method still preserves stable object structures in out-of-FoV regions, such as trees and vehicles. The black regions in the visualization mainly correspond to occluded or insufficiently observed areas, which can be further corrected by subsequent frames.

To further assess robustness, we introduce synthetic segmentation noise by randomly flipping dynamic masks with a probability of 20%. The results show that performance remains stable, with only a minor PSNR decrease of 0.12 dB. This result can be explained by the spatio-temporal aggregation and error-aware mechanism, which together reduces reliance on single-frame mask accuracy and enables recovery from noisy predictions using consistent observations across views and time. This experiment confirms the strong fault tolerance of the proposed approach with respect to segmentation errors.

4.3 Ablation Study

In this section, we analyze the effectiveness of our key components on the Waymo dataset. The ablation results are summarized in Tab. 3.

Ablation Study of Gaussian Pruning. First, we perform ablation studies on voxelization and the soft-gating mechanism to evaluate their roles in Gaussian pruning. Results show that voxelization effectively aggregates Gaussians with different confidence levels but is less effective at suppressing redundant regions composed of low-confidence primitives. Although soft gating is more effective at removing Gaussians, directly deleting them may disrupt potential spatio-temporal associations, leading the model to sacrifice sparsity to maintain PSNR. Furthermore, ablation on the depth-guided loss indicates that it encourages compact representations in geometrically certain regions while adopting

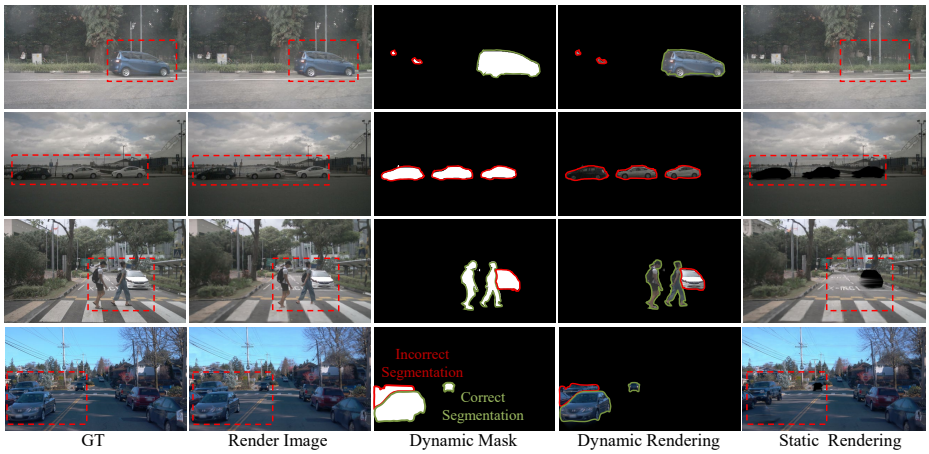

Fig. 5: Robustness analysis. Note that in the dynamic mask, red regions indicate false-positive segmentation, while green regions indicate correct segmentation. Consequently, even when false-positive dynamic segmentation induce localized holes, the overall rendering fidelity is virtually unaffected.

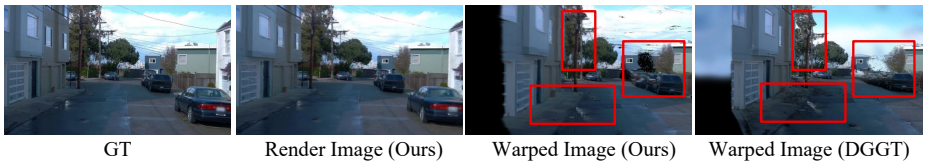

Fig. 6: Rendering results outside the current frame's field of view.

a conservative strategy in ambiguous areas, thereby balancing reconstruction quality and sparsity.

Ablation Study on Error Encoder. To evaluate the necessity of the Error Encoder, we perform an ablation by removing this module, which restricts the model to generating confidence scores for Gaussian primitives based solely on the 2D image features of the current frame. Experimental results indicate that, without the guidance of error information, the model tends to learn one-to-one correspondences between individual pixels and Gaussian primitives. This prevents the model from dynamically distributing Gaussians in accordance with the actual requirements of scene reconstruction. Consequently, redundant Gaussians suffer from mutual occlusion and spatial ambiguity, ultimately compromising reconstruction quality.

Ablation Study on Sequence Length. The results demonstrate that the proposed method not only supports an arbitrary number of input views but also maintains stable performance. Notably, the decrease in the average number of Gaussians as L increases is reasonable. Unlike pixel-aligned methods, where the number of primitives grows linearly with the number of frames, our approach

Table 3: Ablation studies. " w/o " indicates that the module is removed.

	PSNR↑	SSIM↑	LPIPS↓	PGS
w/o soft gating	27.43	0.812	0.169	89806
w/o voxelization	27.18	0.801	0.171	104434
w/o $\mathcal{L}_{gp}$	27.83	0.834	0.167	62121
w/o $\mathcal{L}_{distill}$	13.04	0.464	0.437	26201
w/o error encoder	26.95	0.787	0.174	80229
Complete model (L=8)	28.32	0.853	0.155	42836
Complete model (L=16)	28.11	0.845	0.157	38794
Complete model (L=4)	28.02	0.849	0.157	49556

leverages historical Gaussians as priors to generate new Gaussian primitives on demand, resulting in a slower growth in the number of Gaussians.

Ablation Study of Gaussian Distillation Loss. Experimental results indicate that removing the distillation consistency loss leads to a catastrophic collapse of scene reconstruction. This failure fundamentally stems from the fact that the model learns a joint representation of the scene as a collective ensemble of all Gaussian primitives, coupled with a joint distribution of confidence and opacity. Under this coupling mechanism, pruning Gaussians directly based on confidence scores inevitably results in the erroneous removal of critical backbone Gaussians, thereby triggering reconstruction failure. Additionally, we observe that directly optimizing the photometric error between I_{prune} and the ground truth, rather than using distillation learning, tends to indiscriminately increase the confidence scores of all Gaussian primitives to fit specific viewpoints. As a result, the network degenerates into a trivial pixel-wise generation scheme.

5 Conclusion

This paper introduces an efficient feed-forward scheme for 3D Gaussian Splatting, which aims to decouple and prune various Gaussians from the perspective of long-term reusability, addressing the inherent limitations of pixel-aligned and voxel-based paradigms. The proposed method incorporates error-awareness and a dual-confidence guidance mechanism to achieve on-demand Gaussian generation and differentiable pruning, demonstrating superior performance. Due to the stochastic and sparse nature of dynamic objects, we suppress redundancy only within a single frame to avoid cumulative drift and maintain real-time responsiveness. In future work, we will explore the spatio-temporal correlation of dynamic objects across multiple frames.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous

driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024)
3. Chen, S., Peng, P.: Freegen: Feed-forward reconstruction-generation co-training for free-viewpoint driving scene synthesis. arXiv preprint arXiv:2512.04830 (2025)
4. Chen, X., Xiong, Z., Chen, Y., Li, G., Wang, N., Luo, H., Chen, L., Sun, H., Wang, B., Chen, G., et al.: Dggt: Feedforward 4d reconstruction of dynamic driving scenes using unposed images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1265–1276 (2026)
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. arXiv preprint arXiv:2403.14627 (2024)
6. Chen, Y., Gu, C., Jiang, J., Zhu, X., Zhang, L.: Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. International Journal of Computer Vision **134**(3), 83 (2026)
7. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
8. Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., et al.: Instantsplat: Sparse-view gaussian splatting in seconds. arXiv preprint arXiv:2403.20309 (2024)
9. Fang, Z., Xie, R., Jin, X., Ye, Q., Chen, W., Zheng, W., Wang, R., Huo, Y.: A3gs: Arbitrary artistic style into arbitrary 3d gaussian splatting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17751–17760 (2025)
10. Feng, C., Yu, W., Cheng, X., Tang, Z., Zhang, J., Yuan, L., Tian, Y.: Ae-nerf: Augmenting event-based neural radiance fields for non-ideal conditions and larger scenes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2924–2932 (2025)
11. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 634–644 (2024)
12. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4220–4230 (2024)
13. 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. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
14. Jiang, Y., Yu, C., Xie, T., Li, X., Feng, Y., Wang, H., Li, M., Lau, H., Gao, F., Yang, Y., et al.: Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–1 (2024)
15. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
16. 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)

17. Li, D., Huang, S.S., Lu, Z., Duan, X., Huang, H.: St-4dgs: Spatial-temporally consistent 4d gaussian splatting for efficient dynamic scene rendering. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
18. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers.(2022). URL https://arxiv. org/abs/2203.17270 (2022)
19. Liu, D., Wang, Z., Chen, P.: Dsem-nerf: Multimodal feature fusion and global–local attention for enhanced 3d scene reconstruction. Information Fusion **115**, 102752 (2025)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Miao, S., Huang, J., Bai, D., Yan, X., Zhou, H., Wang, Y., Liu, B., Geiger, A., Liao, Y.: Evolsplat: Efficient volume-based gaussian splatting for urban view synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11286–11296 (2025)
22. 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)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
25. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
26. Shah, S., Dey, D., Lovett, C., Kapoor, A.: Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In: Field and service robotics: Results of the 11th international conference. pp. 621–635. Springer (2017)
27. Shi, C., Shi, S., Lyu, X., Liu, C., Sheng, K., Zhang, B., Jiang, L.: Unisplat: Unified spatio-temporal fusion via 3d latent scaffolds for dynamic driving scene reconstruction. arXiv preprint arXiv:2511.04595 (2025)
28. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
29. Szymanowicz, S., Insafutdinov, 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)
30. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
31. Tian, Q., Tan, X., Xie, Y., Ma, L.: Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7374–7382 (2025)

32. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12922–12931 (2022)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
35. Wang, W., Chen, Y., Zhang, Z., Liu, H., Wang, H., Feng, Z., Qin, W., Zhu, Z., Chen, D.Y., Zhuang, B.: Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. arXiv preprint arXiv:2509.19297 (2025)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
37. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21729–21740 (2023)
38. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
39. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., Zhang, J.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). pp. 19114–19138 (2024)
40. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
41. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: CVPR (2025)
42. Yang, J., Huang, J., Chen, Y., Wang, Y., Li, B., You, Y., Igl, M., Sharma, A., Karkus, P., Xu, D., Ivanovic, B., Wang, Y., Pavone, M.: Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. In: ICLR (2025)
43. Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M., et al.: Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. arXiv preprint arXiv:2311.02077 (2023)
44. Yang, Y., Shao, J., Li, X., Shen, Y., Geiger, A., Liao, Y.: Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2857–2869 (2025)
45. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
46. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207 (2024)
47. Yugay, V., Li, Y., Gevers, T., Oswald, M.R.: Gaussian-slam: Photo-realistic dense slam with gaussian splatting. arXiv preprint arXiv:2312.10070 (2023)

48. Zhan, Y.T., Yang, H.b., Ho, C.Y., Chiang, J.C., Peng, W.H.: Cat-3dgs pro: A new benchmark for efficient 3dgs compression. In: 2025 33rd European Signal Processing Conference (EUSIPCO). pp. 1367–1371. IEEE (2025)
49. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)