2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ADGaussian: Generalizable Gaussian Splatting for Autonomous Driving via Multi-modal Joint Learning

Qi Song[1], Chenghong Li[1], Haotong Lin[2], Sida Peng[2], Rui Huang[1†]

Abstract— **We present a novel approach, termed ADGaussian, for generalizable street scene reconstruction. The proposed method enables high-quality rendering from merely single-view input. Unlike prior Gaussian Splatting methods that primarily focus on geometry refinement, we emphasize the importance of joint optimization of image and depth features for accurate Gaussian prediction. To this end, we first incorporate sparse LiDAR depth as an additional input modality, formulating the Gaussian prediction process as a joint learning framework of visual information and geometric clue. Furthermore, we propose a Multi-modal Feature Matching strategy coupled with a Multi-scale Gaussian Decoding model to enhance the joint refinement of multi-modal features, thereby enabling efficient multi-modal Gaussian learning. Extensive experiments on Waymo and KITTI demonstrate that our ADGaussian achieves state-of-the-art performance and exhibits superior zero-shot generalization capabilities in novel-view shifting. Project page.**

I. INTRODUCTION

Recently, 3D Gaussian Splatting (3DGS) [1] has garnered significant attention in the fields of 3D scene reconstruction and novel view synthesis [2] due to its real-time rendering speed and high-quality output. One key application is the modeling of street scenes from image sequences, which plays a vital role in areas such as autonomous driving [3].

When modeling urban scenes, some methods follow per-scene optimization techniques [4]–[6], notably StreetGaussians [7] that represents dynamic urban street as a set of point clouds equipped with semantic logits and 3D Gaussians. While such an approach excels in high-quality reconstruction, it struggles with expensive training cost and large-range novel view synthesis, which motivates the exploration of generalizable models that avoid scene-specific fine-tuning.

To achieve generalizable street scene reconstruction, most existing methods build upon the architectures of PixelSplat [8] or MVSplat [9]. For instance, GGRt [10] introduces a pose-free architecture to iteratively update multi-view depth map and subsequently estimates Gaussian primitives based on PixelSplat. Similarly, GGS [11] enhances the depth estimations of MVSplat by integrating a multi-view depth refinement module. Nevertheless, multi-view feature matching-based depth estimation may fail in challenging conditions such as texture-less areas and reflective surfaces. To tackle this issue, the concurrent work DepthSplat [12] combines pre-trained depth features from Depth Anything V2 [13] with multi-view depth estimations for accurate depth regression, where the estimated depth features are then concatenated with image features for Gaussian prediction.

Given the great generalization capability of Depth Anything V2, it is reasonable to extend DepthSplat to urban street scenarios. However, DepthSplat faces specific limitations when applied to these environments. First, as highlighted in the dashed box of Fig. 1, the visual rendering quality is constrained by the performance of pre-trained depth models, which often exhibit inconsistent accuracy across diverse street datasets and scenarios. Second, simply concatenating image and depth features for the final Gaussian prediction without any information sharing or multi-modal feature fusion may lead to unexpected spatial misalignment, evidenced by the distorted car shape in the red box of Fig. 1.

To overcome these limitations, we present ADGaussian, a novel multi-modal framework that jointly optimizes visual rendering quality and geometric accuracy for street scenes. Instead of relying on pre-trained depth foundation models, we choose to integrate sparse LiDAR depth information as an additional input modality, which is more practical in real-world scenarios and provides precise metric scale priors for geometry reconstruction. Given two complementary modalities (i.e., image data and sparse depth map), the framework's core innovation lies in its synergistic processing of these two data streams, enabling effective information sharing and joint optimization between different modalities. Specifically, we introduce a Multi-modal Feature Matching strategy augmented by the Depth-guided Position Embedding, which contains a Siamese-style encoder paired with an information cross-attention decoder. This design ensures a cohesive fusion of geometric and appearance information, resulting in well-aligned multi-modal tokens. Subsequently, we employ a Multi-scale Gaussian Decoding model to aggregate multi-scale depth information into the resulting multi-modal tokens for the final 3D Gaussian predictions.

As presented in Fig. 1, ADGaussian excels in both visual rendering and geometric reconstruction. Notably, the bottom row demonstrates our model's robustness under extreme viewpoint variations, where competing methods often produce distorted geometry. This superior performance validates the efficacy of ADGaussian's synchronized cross-modal optimization paradigm.

Overall, this work makes the following contributions:

- We present ADGaussian, the first generalizable framework that formulates street scene Gaussian prediction as joint visual and geometric learning.

*This work was supported by Shenzhen Science and Technology Program under grant No. JCYJ20220818103006012 and ZDCY20250901103359008.

[1]School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Longgang, Shenzhen, Guangdong, 518172, P.R. China
[2]Zhejiang University, Zhejiang, 310058, P.R. China
[†]Corresponding author is also affiliated with School of Artificial Intelligence, ruihuang@cuhk.edu.cn

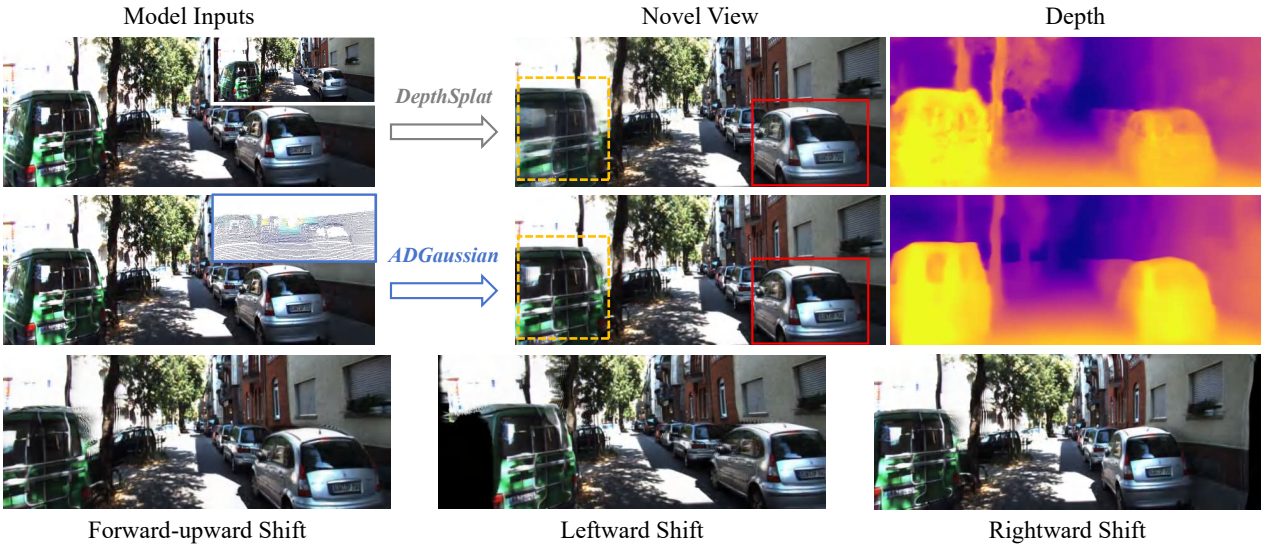

Fig. 1. We introduce **ADGaussian**, a generalizable Gaussian framework that achieves superior street scene reconstruction in both visual and geometric quality. The bottom row illustrates the results of viewpoint shifting, further demonstrating the robustness of our method under varying viewpoint changes.

- We develop a Multi-modal Feature Matching strategy along with a Multi-scale Gaussian Decoding model to facilitate effective multi-modal Gaussian learning.
- We conduct extensive comparisons on two datasets, verifying our approach's state-of-the-art performance and the effectiveness of the proposed components.

II. RELATED WORK

A. Generalizable 3D Gaussian Splatting

Generalizable Gaussian Splatting [14]–[16] aims to learn powerful priors that enable effective generalization across unseen scenes. Existing methods can be broadly categorized into two groups based on their handling of camera parameters. The first group, including approaches like MVSGaussian [17] and SplatterImage [18], predicts per-pixel 3D Gaussian primitives using known camera parameters. These works typically demonstrate superior reconstruction accuracy due to the precise geometric constraints provided by the camera poses. The second group of methods [19], [20] proposes to jointly predict camera parameters and 3D representations, eliminating the need for known camera poses. For instance, GGRt [10] employs an Iterative Pose Optimization Network to estimate and iteratively update the relative pose between target and reference images. DrivingForward [21] adopts pose network and depth network to determine the position of the Gaussian primitives in a self-supervised manner. In street scene modeling, however, camera poses provide critical constraints for determining scene scale and enhancing reconstruction accuracy from image sequences. Moreover, camera poses are readily accessible in street scenes, making them a practical and reliable data resource. Therefore, we choose to leverage posed images for our approach.

B. Depth and Gaussian Splatting

Depth quality has been demonstrated to play a pivotal role in Gaussian Splatting, serving as the foundation for accurate geometry reconstruction and realistic rendering. To ensure precise geometric fidelity, existing approaches [22]–[24] incorporate additional depth supervision into the optimization process. However, since dense Ground Truth depth data is often unavailable in practical applications, researchers have turned to pre-trained depth foundation models [13], [25] as an alternative source of reliable geometric cues. For example, Chung et al. [26] rescale pre-trained depth maps using sparse COLMAP points to provide depth constraints. DepthSplat [12] fuses pre-trained depth features with multi-view cost volume features to help depth refinement. While these methods have significantly improved geometric accuracy, we contend that their primary focus on geometry enhancement overlooks the crucial interplay between appearance and structure. In contrast to previous approaches, we argue that joint optimization of image and depth features is more critical for achieving high-quality reconstruction that excels in both visual fidelity and structural accuracy.

C. LiDAR-Integrated Gaussian Splatting

The integration of LiDAR data has emerged as a widely adopted approach in street scene reconstruction [27], [28], due to its effectiveness in facilitating geometry learning. The conventional methodologies typically involve two main steps: initializing Gaussians from LiDAR point clouds [29] to establish the basic scene structure, and further supervising predicted Gaussian positions with LiDAR priors [30], [31] to refine geometric details. Typically, TCLC-GS [32] constructs a hybrid 3D representation by combining LiDAR geometries with image colors, enabling simultaneous initialization of both geometric and appearance attributes of 3D Gaussians. Rather than directly using LiDAR point clouds, we propose leveraging sparse LiDAR depth to bridge the modality gap between LiDAR and camera data. Furthermore, we integrate depth priors as an additional input modality into a unified optimization framework, achieving joint optimization of depth geometries and image photometric attributes, as opposed to the common practice of initialization alone.

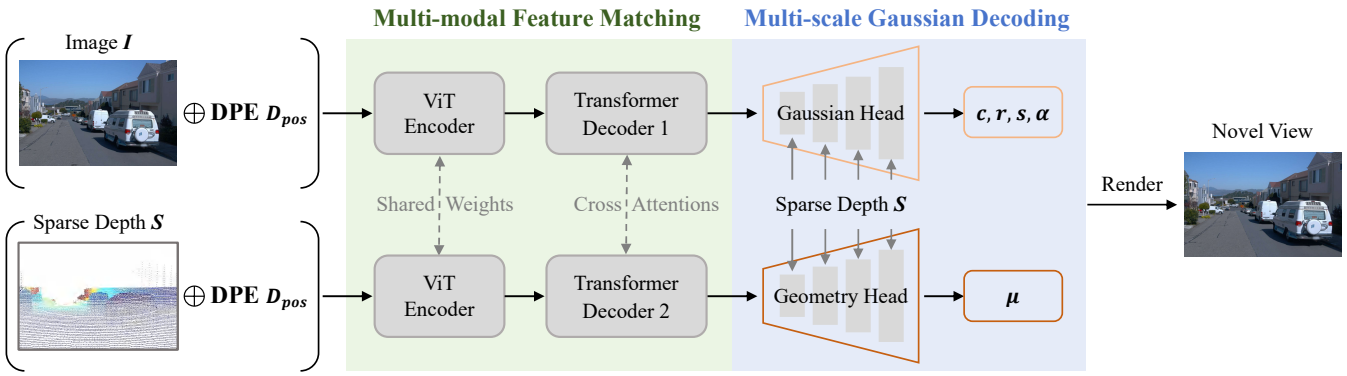

Fig. 2. **Overall framework of ADGaussian.** Given monocular posed image and sparse depth as inputs, we first extract fused multi-modal features through Multi-modal Feature Matching strategy enhanced by Depth-guided Positional Embedding (DPE). These aligned multi-modal tokens are then processed by our Multi-scale Gaussian Decoding module, which hierarchically integrates depth cues across scales to produce optimized 3D Gaussian outputs.

III. Method

Depth foundation models [13], [25] have been integrated into the Gaussian Splatting to improve geometry reconstruction. However, such a framework often suffers from suboptimal rendering quality due to the insufficient interactions between photometric and geometric clues. To address this, we propose ADGaussian, a synchronized multi-modal optimization architecture that combines sparse depth data with monocular images for enhanced street scene modeling.

A. Preliminary

Recently, some works have investigated the advantages of using a pre-trained depth foundation model for image-conditioned 3D Gaussian reconstruction. For instance, DepthSplat [12] processes multi-view images $\{I^i\}_{i=1}^N (I \in \mathbb{R}^{H\times W\times 3})$ using two parallel branches to extract dense per-pixel depth. One branch focuses on modeling cost volume features C^i from the multi-view input, while the other employs a pre-trained monocular depth backbone, specifically Depth Anything V2 [13], to obtain monocular depth features F^i_{mono}. These per-view cost volumes and monocular features are then concatenated for depth regression. Finally, DepthSplat predicts all remaining Gaussian parameters using the concatenated image, depth, and feature information.

Intuitively, such models can be easily adapted to urban scenes. Nonetheless, we observed that the effectiveness of reconstruction is heavily dependent on the performance of the pre-trained depth foundation models, resulting in inconsistent accuracy across different street datasets and scenarios. Furthermore, the processing of image and depth features always occurs in parallel for each view, without any information sharing or synchronized optimization, which constrains the model's learning capacity.

In this paper, instead of relying on pre-trained depth foundation models, we propose utilizing sparse LiDAR depth measurements as an additional input. This choice is particularly advantageous for street scene applications, as LiDAR data offers both greater practical accessibility in real-world autonomous driving systems and reliable metric-scale depth priors crucial for accurate geometric reconstruction.

B. Multi-modal Feature Matching

In this subsection, we seek to find an effective way to integrate sparse LiDAR depth into Gaussian Splatting. To this end, we propose a *Multi-modal Feature Matching* architecture tailored for urban scenarios to enable the synchronous integration of sparse depth information and color image data. Throughout this process, *Depth-guided Position Embedding* incorporates depth cues into the position embedding, enhancing 3D spatial awareness and improving multi-modal contextual comprehension.

a) Multi-modal Feature Matching: As illustrated in Fig. 2, the first part of our model is the *Multi-modal Feature Matching* of photometric features from the image and geometric cues from depth data. This is achieved through a Siamese-style encoder and an information cross-attention decoder, inspired by the DUSt3R series [33], [34].

Specifically, given a monocular image $I \in \mathbb{R}^{H\times W\times 3}$ and synchronized sparse depth map $S \in \mathbb{R}^{H\times W\times 1}$, we first replicate the depth map across channels to match the image's dimensional structure. These multi-modal inputs are then fed into a weight-sharing ViT encoder, resulting in two token representations F_I and F_S:

$$F_I = \text{Encoder}(I), F_S = \text{Encoder}(S) \tag{1}$$

The two identical encoders collaboratively process multi-modal features in a weight-sharing manner, allowing for the automatic learning of similarity characteristics.

After that, the transformer decoders equipped with cross attentions are employed to enhance information sharing and synchronized optimization between the two multi-modal branches. This step is crucial for producing well-fused multi-modal feature maps:

$$\begin{aligned} G_I &= \text{Decoder}_1(F_I, F_S), \\ G_S &= \text{Decoder}_2(F_S, F_I) \end{aligned} \tag{2}$$

b) Depth-guided Positional Embedding (DPE): The conventional positional embedding in Vision Transformers encodes either relative or absolute spatial positions on a 2D image plane to ensure spatial awareness within the image. However, relying solely on the geometric properties of a

TABLE I

QUANTITATIVE COMPARISONS WITH STATE OF THE ART ON WAYMO DATASET. OUR ADGAUSSIAN OUTPERFORMS EXISTING METHODS IN NEARLY ALL SCENARIOS. CELLS HIGHLIGHTED IN DENOTES THE BEST AND SECOND-BEST PERFORMANCES.

	Scene	PSNR↑			SSIM↑			LPIPS↓		
		MVSplat [9]	DepthSplat [12]	**Ours**	MVSplat [9]	DepthSplat [12]	**Ours**	MVSplat [9]	DepthSplat [12]	**Ours**
Static	003	21.79	19.99	31.09	0.679	0.627	0.931	0.143	0.192	0.059
	069	24.79	25.67	31.17	0.729	0.748	0.923	0.143	0.136	0.073
	232	28.79	26.76	30.52	0.873	0.819	0.904	0.077	0.094	0.083
	495	28.09	26.49	31.21	0.884	0.819	0.929	0.086	0.106	0.056
Dynamic	016	24.16	24.30	27.16	0.678	0.746	0.875	0.137	0.173	0.092
	021	19.58	18.42	19.61	0.636	0.619	0.659	0.243	0.316	0.273
	080	25.37	24.19	27.18	0.765	0.759	0.873	0.116	0.169	0.085
	096	21.55	21.67	21.46	0.684	0.680	0.691	0.250	0.264	0.263

2D image plane is insufficient for our synchronized multi-modal design. To this end, we propose a straightforward Depth-guided Positional Embedding (DPE) to integrate depth positions with image-based spatial positions

In particular, given the downsampled image size $H_L \times W_L$ and the sparse depth map, we first flatten the 2D grid of spatial positions into a 1D vector through row-major ordering, where each spatial position (i, j) in the 2D grid is mapped to linear index $k = i \times W_L + j$ in the 1D vector. Meanwhile, the sparse depth map is downsampled to match the image resolution, generating an independent set of depth indices that complement the spatial positions. The final positional embedding D_{pos} is then formed by concatenating the flattened spatial coordinates with their corresponding depth values, establishing an integrated xy-z coordinate representation that encodes both planar and depth-wise positional information. By integrating both spatial and depth geometry, this module provides a comprehensive positional prior for effective multi-modal feature fusion.

C. Multi-scale Gaussian Decoding

Given the multi-modal tokens G_I and G_S, our objective is to predict pixel-aligned Gaussian parameters $\{(\mu, \alpha, \Sigma, c)\}^{H \times W}$, where μ, α, Σ, and c are the 3D Gaussian's center position, opacity, covariance, and color information. To fully leverage appearance cues and the geometry priors provided by image token G_I and depth token G_S, we implement two separate regression heads with the same architecture, namely Gaussian Head and Geometry Head, to generate different Gaussian parameters.

The two regression heads adhere to the DPT [35] architecture, enhanced with an additional multi-scale depth encoding that delivers precise scale priors for Gaussian prediction. In particular, at each scale within the DPT Decoder, we initially resize the input depth map to align with the spatial size of the current feature scale. After that, the resized depth map is processed through a shallow network comprising two convolutional layers to extract depth features, which are then added to the DPT intermediate features. Finally, the input image and depth map, each processed by a single convolutional layer, are individually incorporated into the final features of the Gaussian Head and Geometry Head to facilitate either appearance-based or geometry-based Gaussian decoding.

D. Training Loss

Our model is trained using a combination of novel view synthesis loss and depth loss:

$$\mathcal{L} = \mathcal{L}_{\text{nvs}} + \mathcal{L}_{\text{depth}} \tag{3}$$

a) Novel view synthesis loss: We train our full model with a combination of mean squared error (MSE) and LPIPS losses between rendered and Ground Truth image colors:

$$\mathcal{L}_{\text{nvs}} = \text{MSE}(I_{\text{pred}}, I_{\text{gt}}) + \lambda \cdot \text{LPIPS}(I_{\text{pred}}, I_{\text{gt}}) \tag{4}$$

where the LPIPS loss weight λ is set to 0.05.

b) Depth loss: We leverage depth loss to smooth the depth values of neighboring pixels, thereby minimizing abrupt changes over small regions:

$$\mathcal{L}_{\text{depth}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{dD_i}{dx}e^{-\frac{dI_i}{dx}} + \frac{dD_i}{dy}e^{-\frac{dI_i}{dy}}\right) \tag{5}$$

where $\frac{dD_i}{dx}$, $\frac{dD_i}{dy}$, $\frac{dI_i}{dx}$, and $\frac{dI_i}{dy}$ denote the first derivatives of depth and image in the x and y-axis directions, respectively.

IV. EXPERIMENTS

A. Implementation Details

a) Datasets and metrics: We evaluate our proposed approach on two widely used autonomous driving datasets: the Waymo Open Dataset [36] and the KITTI Tracking benchmarks [37]. For both datasets, we adopt a train-test split ratio of approximately 1:7. Specifically, on the Waymo dataset, our focus primarily lies on static and dynamic scenes, where each scene type is divided into 4 test scenes and 28 training scenes. Similarly, for the KITTI dataset, the split consists of 5 test scenes and 37 training scenes. For render quality evaluation, we employ the standard image quality metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [38], and the Learned Perceptual Image Patch Similarity (LPIPS) [39].

b) Training details: We employ the Adam optimizer and cosine learning rate schedule, with an initial learning rate of $1e$-4. We train our model on a single 3090 Ti GPU, running for 150k iterations on both Waymo and KITTI datasets, with a batch size of 1. To ensure a fair comparison, all experiments are carried out at resolutions of 320×480 for the Waymo dataset and 256×608 for the KITTI dataset.

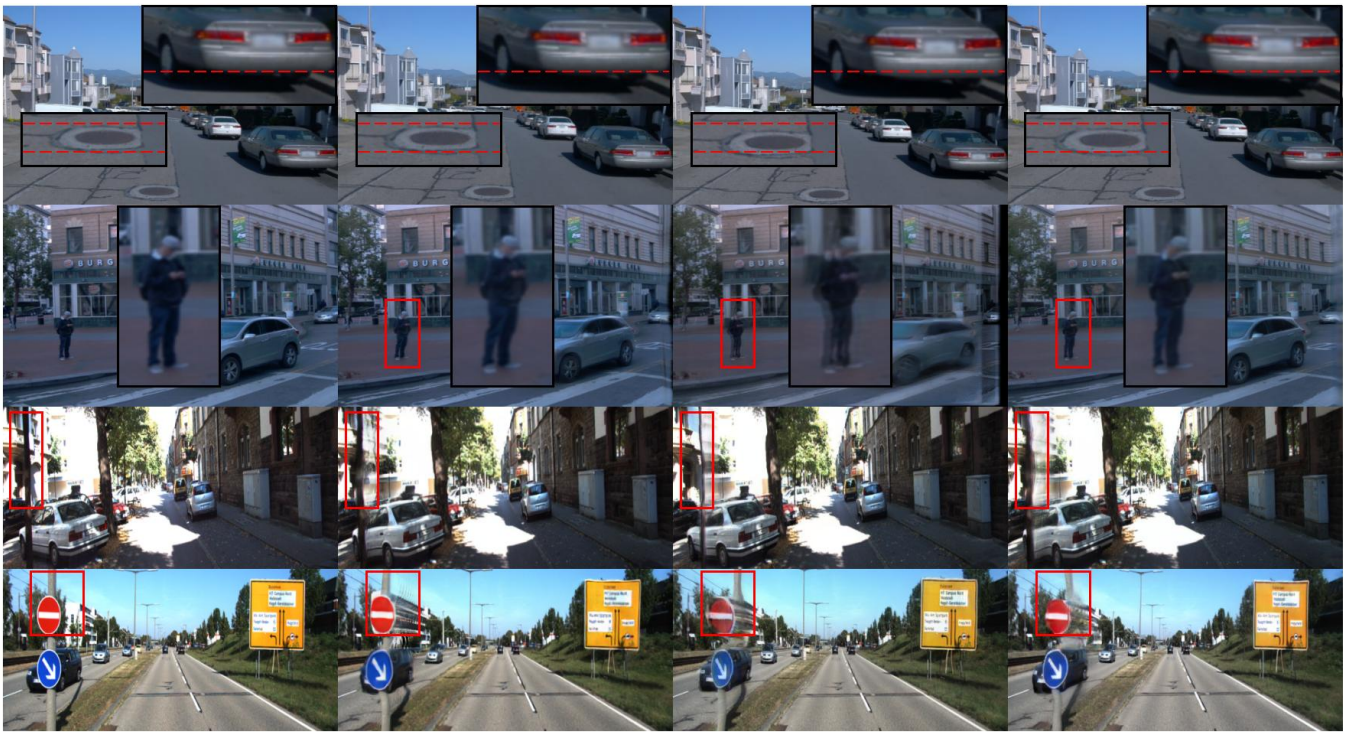

Fig. 3. **Qualitative comparisons with state of the art.** Our ADGaussian surpasses all other competitive models in rendering quality within urban scenarios. The zoom-in comparisons in the first row further reveal our enhanced spatial alignment capabilities.

TABLE II

QUANTITATIVE COMPARISONS WITH STATE OF THE ART ON KITTI DATASET. KITTI[1] REFERS TO SUBSEQUENT TEMPORAL FRAME RENDERING, WHILE KITTI[2] INDICATES SPATIAL LEFT-TO-RIGHT CAMERA RENDERING. CELLS HIGHLIGHTED IN DENOTES THE BEST AND SECOND-BEST PERFORMANCES.

Method	KITTI[1]			KITTI[2]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MVSplat [9]	23.52	0.760	0.152	24.65	0.833	0.141
DepthSplat [12]	21.99	0.715	0.173	25.36	0.838	0.135
Ours	23.60	0.776	0.164	26.10	0.853	0.122

B. Comparisons with the State of the Art

When comparing our work with current state-of-the-art Gaussian Splatting methods, we selected MVSplat [9] (a multi-view cost volume-based approach) and DepthSplat [12] (a depth foundation model-based approach) as primary baselines. Since our method focuses on pose-aware generalizable Gaussian Splatting, we excluded pose-free methods and per-scene optimized baselines from the main comparisons. To ensure fair and consistent comparisons, all baseline methods were re-implemented and trained on both datasets using identical experimental settings to ours. Specifically, for each scenario, both MVSplat and DepthSplat utilize consecutive frame pairs as input, with the subsequent immediate frame serving as the target novel view for evaluation.

The quantitative comparisons on the Waymo and KITTI benchmarks are presented in Table I and Table II, respectively. On the Waymo dataset, our ADGaussian surpasses previous state-of-the-art models on almost all visual metrics, with particularly significant gains in static scenes and consistent performance across diverse scenarios. While competing methods suffer from pervasive spatial misalignment issues that severely degrade their metric accuracy (as evidenced in Fig. 3), our model effectively resolves this fundamental challenge through cross-modal joint learning, which synchronizes geometric and appearance optimization to achieve pixel-perfect alignment. On the KITTI dataset, we introduce an additional left-to-right view synthesis setting, leveraging its stereo camera data. As shown in Table II, our method achieves superior performance in both settings compared to prior works. However, the performance gain on KITTI is less pronounced compared to that on Waymo. This is primarily attributed to the overall lower image quality and poor color reproduction of the KITTI dataset. Since our method relies solely on a single image as input, it retains fewer image details compared to previous works, which further constrains its performance on datasets with inferior image quality.

We provide qualitative results of the two datasets in Fig. 3. As can be observed, the zoomed regions in the first row reveal pervasive spatial misalignment in previous methods, which leads to significant degradation in evaluation metrics. The last two rows further demonstrate that our model achieves superior rendering quality, particularly in occluded regions and fine details such as slender signal poles.

Besides, the comparison between DepthSplat and MVSplat shows that DepthSplat exhibits stronger depth inference capabilities, attributed to its enhanced geometry reconstruction

TABLE III

ABLATION STUDIES ON THE WAYMO DATASET. WE REPORT THE AVERAGED SCORES ACROSS ALL VALIDATION SCENES FOR A MORE INTUITIVE REFLECTION OF MODEL PERFORMANCE.

Setup	PSNR↑	SSIM↑	LPIPS↓
Full Model	31.00	0.921	0.068
w/o DPE	30.31	0.908	0.078
w/o Multi-scale	28.73	0.868	0.100
w/o DPE & Multi-scale	27.81	0.846	0.114
w/o Matching	26.68	0.814	0.106

TABLE IV

ANALYSES ON FORWARD VIEW SHIFTING AND RUNNING EFFICIENCY.

Method	PSNR↑	SSIM↑	LPIPS↓	Time↓	Memory↓
MVSplat [9]	24.84	0.777	0.133	0.22/0.14	11.11
DepthSplat [12]	23.40	0.726	0.190	0.37/0.28	21.17
Ours	27.68	0.877	0.101	0.29/0.18	17.52

facilitated by pre-trained depth models. However, DepthSplat falls short in overall visual reconstruction quality due to its insufficient integration of appearance attributes, which is consistent with our earlier analysis in the preceding sections.

C. Ablations and Analyses

a) Ablations on proposed components: The ablation studies are detailed in Table III to confirm the efficacy of proposed components. First, it can be seen that the full model achieves the highest performance, boasting a PSNR, SSIM, and LPIPS score of 31.0, 0.921, and 0.068, respectively. Notably, the removal of the Depth-guided Positional Embedding (DPE) resulted in a decrease across all metrics (0.69, 1.3%, and 1%, respectively), emphasizing the significance of depth positions in facilitating the joint optimization of multi-modal features. Furthermore, the model lacking Multi-scale Gaussian Decoding (w/o Multi-scale) exhibited reduced performance, achieving a PSNR of 28.73, an SSIM of 0.868, and an LPIPS of 0.100, underscoring the effectiveness of multi-level depth decoding and independent Gaussian inference. Removing both DPE and Multi-scale led to a more substantial drop in performance, notably a 4.6% decrease in the LPIPS score. This quantitative degradation aligns with the qualitative results in Fig. 4, which validates the importance of our DPE and Multi-scale Gaussian Decoding in addressing spatial misalignment for street scene reconstruction.

Finally, to showcase the effectiveness of our synchronized multi-modal optimization formulation, we present the results without Multi-modal Feature Matching (w/o Matching) by substituting the sparse depth input with a color image from the subsequent frame. It is evident that Multi-modal Feature Matching brought about significant enhancement in PSNR, SSIM, and LPIPS (4.32, 10.7%, and 3.8%, respectively), highlighting the importance of information exchange and synchronized optimization of image-related appearance features and depth-related geometric features.

b) Robustness across frames: In the previous experiments, the next frame (*t+1* frame) was used as the target for novel view synthesis. To further evaluate the robustness under more significant viewpoint changes, we extend the prediction to the *t+2* frame. As reported in Table IV, our ADGaussian outperforms in handling larger temporal and spatial shifts, even with only a single input frame.

Fig. 4. **Ablations on the Waymo dataset.** We present a visual comparisons between our baseline model (w/o DPE & Multi-scale) and the full model. Our experiments demonstrate that the full model can effectively address the spatial misalignment issues prevalent in street scene reconstruction.

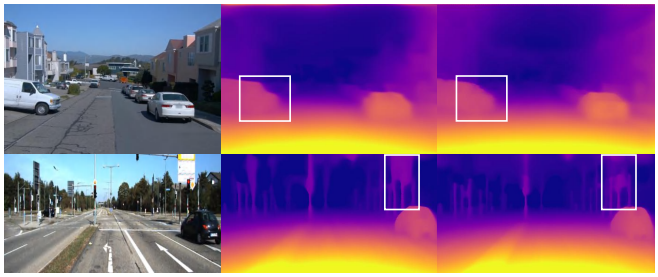

Fig. 5. **Depth comparisons with depth completion networks.** Our method demonstrates superior depth estimation performance in certain challenging regions, even without depth pre-training.

c) Analysis on running efficiency: The right side of table IV shows training/inference runtime and memory consumption of ADGaussian, MVSplat and DepthSplat at 320×480 resolution to validate ADGaussian's practicality. ADGaussian uses 3.65 GB less training memory than DepthSplat for higher efficiency, and its inference speed is merely 0.04 s slower than MVSplat with comparable overall runtime.

D. Analyses on Multi-modal Inputs

To ensure a fair comparison, we further constructed baseline networks with identical multi-modal inputs using state-of-the-art depth completion methods. Specifically, we re-implemented CFormer [40] and BPNet [41] as the comparison targets, which take both image data and sparse depth as inputs and predict Gaussian parameters using multi-modal fused features. Also, we initialized these models with weights pre-trained on the KITTI depth completion dataset. By maintaining identical input modalities and training conditions, these baselines allow us to evaluate the performance gains attributable to the inclusion of depth data, independent of the architectural advancements in our framework.

a) Comparison Results: As shown in Table V, it is evident that the inclusion of additional depth input alone does not significantly enhance the quality of novel view

TABLE V

PERFORMANCE ANALYSIS ON MULTI-MODAL INPUTS. MODELS MARKED WITH "∗" ARE MODIFIED WITH A GAUSSIAN HEAD FOR 3DGS PREDICTION. THE TERM "WAYMO + DEPTH DROP" REFERS TO OUR ROBUSTNESS EVALUATION ON DEPTH QUALITY.

Method	Waymo			KITTI			Waymo + Depth Drop		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
CFormer* [40]	25.71	0.796	0.126	21.35	0.761	0.192	23.67	0.780	0.141
BPNet* [41]	26.10	0.802	0.144	19.68	0.626	0.336	24.89	0.783	0.152
Ours	31.00	0.921	0.068	23.60	0.776	0.164	30.56	0.912	0.074

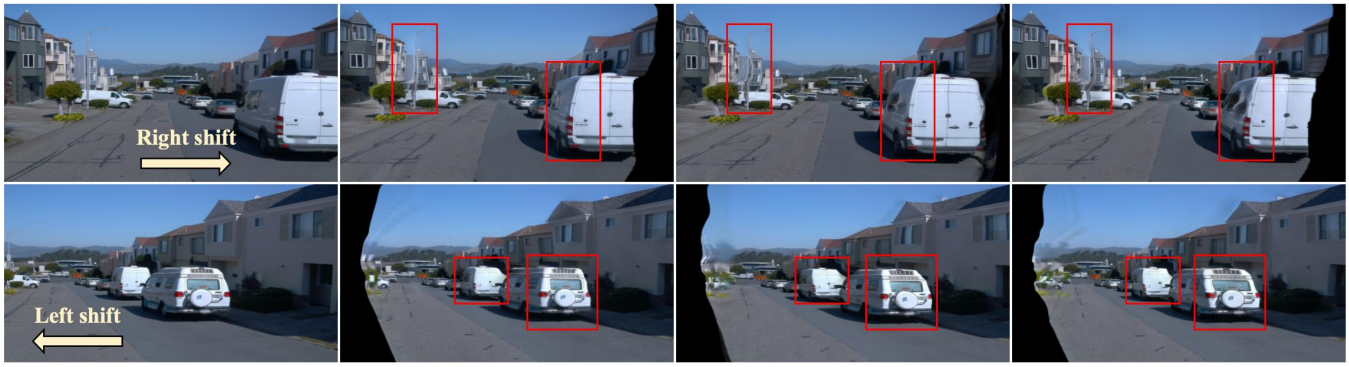

Fig. 6. **Visual comparisons on view shifting.** The figure displays the performance of right and left shifting of the given images on the Waymo dataset. The challenging areas are marked with red rectangles. As observed, our model exhibits superior robustness under large viewpoint changes.

rendering. This demonstrates the limitations of relying solely on accurate depth prediction for high-quality rendering, further highlighting the critical need for our joint optimization. Additionally, as demonstrated in Fig. 5, our model effectively preserves fine structural details (e.g., car contours and pole geometries) without pre-training on depth completion tasks.

b) Robustness to depth quality: We evaluate our model's robustness to depth quality degradation by randomly discarding approximately 50% of LiDAR points during testing. As presented in the right part of Table V, ADGaussian remains competitive with full-depth-input performance under such conditions. The performance gap compared to baselines validates our cross-modal optimization framework's ability to compensate for missing depth information through effective appearance-geometry integration.

E. Application: Novel-view Shifting

The concept of novel-view shifting involves generating images from significantly varied perspectives compared to the original viewpoints present in the training data. This task is particularly demanding as it usually necessitates reliable depth estimations to handle substantial changes in viewpoint and scale. In this study, we further investigate the model's robustness in view shifting. Firstly, the Ground Truth right camera images provided in the KITTI dataset are used to evaluate the quantitative performance of view shifting. As depicted in Table VI, our model significantly outperforms both MVSplat and DepthSplat in zero-shot view shifting from left to right cameras. It is noteworthy that our zero-shot view shifting results are only slightly lower than our normally trained model (PSNR: 23.60, SSIM: 0.776, LPIPS: 0.164). Moreover, visual comparisons on the Waymo dataset

TABLE VI

ROBUST ANALYSES ON NOVEL-VIEW SHIFTING ON THE KITTI DATASET. MODELS TRAINED ON MULTI-FRAME IMAGES ARE UTILIZED DIRECTLY TO EVALUATE THEIR CAPABILITY FOR NOVEL-VIEW SHIFTING, TRANSITIONING FROM THE LEFT CAMERA TO THE RIGHT CAMERA, ALL WITHOUT ADDITIONAL FINE-TUNING.

Method	PSNR↑	SSIM↑	LPIPS↓
MVSplat [9]	14.39	0.474	0.382
DepthSplat [12]	15.07	0.452	0.377
Ours	21.81	0.770	0.184

are presented in Fig. 6. ADGaussian exhibits exceptional view shifting quality, accurately maintaining object shapes and preserving intricate details even in texture-less regions during both leftward and rightward viewpoint changes.

V. CONCLUSION

This paper presents ADGaussian, a novel multimodal framework that advances generalizable street scene reconstruction via synergistic learning of visual and geometric features. We validate that jointly optimizing RGB imagery and sparse depth inputs significantly improves both geometric quality and visual fidelity. Comprehensive experiments on Waymo and KITTI datasets establish state-of-the-art performance in complex urban scenarios. The framework also shows strong zero-shot generalization for large viewpoint shifts while maintaining metric scale consistency. Current limitations include the sparsity of single-view information and the challenge of modeling dynamic objects. Extending the approach to multi-frame fusion could help address these gaps.

REFERENCES

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[3] Q. Song, Q. Hu, C. Zhang, Y. Chen, and R. Huang, "Divide and conquer: Improving multi-camera 3d perception with 2d semantic-depth priors and input-dependent queries," *IEEE Transactions on Image Processing*, vol. 33, pp. 897–909, 2024.

[4] P.-C. Kung, X. Zhang, K. A. Skinner, and N. Jaipuria, "Lihi-gs: Lidar-supervised gaussian splatting for highway driving scene reconstruction," *arXiv preprint arXiv:2412.15447*, 2024.

[5] Y. Yan, Z. Xu, H. Lin, H. Jin, H. Guo, Y. Wang, K. Zhan, X. Lang, H. Bao, X. Zhou *et al.*, "Streetcrafter: Street view synthesis with controllable video diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 822–832.

[6] G. Zhao, X. Wang, C. Ni, Z. Zhu, W. Qin, G. Huang, and X. Wang, "Recondreamer++: Harmonizing generative and reconstructive models for driving scene representation," *arXiv preprint arXiv:2503.18438*.

[7] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians for modeling dynamic urban scenes.(2023)," 2023.

[8] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 457–19 467.

[9] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–386.

[10] H. Li, Y. Gao, C. Wu, D. Zhang, Y. Dai, C. Zhao, H. Feng, E. Ding, J. Wang, and J. Han, "Ggrt: Towards pose-free generalizable 3d gaussian splatting in real-time," in *European Conference on Computer Vision*. Springer, 2024, pp. 325–341.

[11] H. Han, K. Zhou, X. Long, Y. Wang, and C. Xiao, "Ggs: Generalizable gaussian splatting for lane switching in autonomous driving," *arXiv preprint arXiv:2409.02382*, 2024.

[12] H. Xu, S. Peng, F. Wang, H. Blum, D. Barath, A. Geiger, and M. Pollefeys, "Depthsplat: Connecting gaussian splatting and depth," *arXiv preprint arXiv:2410.13862*, 2024.

[13] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[14] Y. Wang, T. Huang, H. Chen, and G. H. Lee, "Freesplat: Generalizable 3d gaussian splatting towards free-view synthesis of indoor scenes," *arXiv preprint arXiv:2405.17958*, 2024.

[15] C. Wewer, K. Raj, E. Ilg, B. Schiele, and J. E. Lenssen, "latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction," in *European Conference on Computer Vision*. Springer, 2024, pp. 456–473.

[16] Y. Chen, J. Wang, Z. Yang, S. Manivasagam, and R. Urtasun, "G3r: Gradient guided generalizable reconstruction," in *European Conference on Computer Vision*. Springer, 2024, pp. 305–323.

[17] T. Liu, G. Wang, S. Hu, L. Shen, X. Ye, Y. Zang, Z. Cao, W. Li, and Z. Liu, "Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo," in *European Conference on Computer Vision*. Springer, 2024, pp. 37–53.

[18] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, "Splatter image: Ultra-fast single-view 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 208–10 217.

[19] B. Smart, C. Zheng, I. Laina, and V. A. Prisacariu, "Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs," *arXiv preprint arXiv:2408.13912*, 2024.

[20] B. Ye, S. Liu, H. Xu, X. Li, M. Pollefeys, M.-H. Yang, and S. Peng, "No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images," *arXiv preprint arXiv:2410.24207*, 2024.

[21] Q. Tian, X. Tan, Y. Xie, and L. Ma, "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, no. 7, 2025, pp. 7374–7382.

[22] H. Lu, T. Xu, W. Zheng, Y. Zhang, W. Zhan, D. Du, M. Tomizuka, K. Keutzer, and Y. Chen, "Drivingrecon: Large 4d gaussian reconstruction model for autonomous driving," *arXiv preprint arXiv:2412.09043*.

[23] S. Zheng, B. Zhou, R. Shao, B. Liu, S. Zhang, L. Nie, and Y. Liu, "Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 680–19 690.

[24] M. Turkulainen, X. Ren, I. Melekhov, O. Seiskari, E. Rahtu, and J. Kannala, "Dn-splatter: Depth and normal priors for gaussian splatting and meshing," *arXiv preprint arXiv:2403.17822*, 2024.

[25] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 106–10 116.

[26] J. Chung, J. Oh, and K. M. Lee, "Depth-regularized optimization for 3d gaussian splatting in few-shot images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 811–820.

[27] J. Wang, S. Manivasagam, Y. Chen, Z. Yang, I. A. Bârsan, A. J. Yang, W.-C. Ma, and R. Urtasun, "Cadsim: Robust and scalable in-the-wild 3d reconstruction for controllable sensor simulation," *arXiv preprint arXiv:2311.01447*, 2023.

[28] Z. Yang, S. Manivasagam, Y. Chen, J. Wang, R. Hu, and R. Urtasun, "Reconstructing objects in-the-wild for realistic sensor simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 661–11 668.

[29] M. Khan, H. Fazlali, D. Sharma, T. Cao, D. Bai, Y. Ren, and B. Liu, "Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction," *arXiv preprint arXiv:2407.02598*, 2024.

[30] C. Jiang, R. Gao, K. Shao, Y. Wang, R. Xiong, and Y. Zhang, "Li-gs: Gaussian splatting with lidar incorporated for accurate large-scale reconstruction," *IEEE Robotics and Automation Letters*, 2024.

[31] N. Huang, X. Wei, W. Zheng, P. An, M. Lu, W. Zhan, M. Tomizuka, K. Keutzer, and S. Zhang, "S3gaussian: Self-supervised street gaussians for autonomous driving," *arXiv preprint arXiv:2405.20323*, 2024.

[32] C. Zhao, S. Sun, R. Wang, Y. Guo, J.-J. Wan, Z. Huang, X. Huang, Y. V. Chen, and L. Ren, "Tclc-gs: Tightly coupled lidar-camera gaussian splatting for autonomous driving: Supplementary materials," in *European Conference on Computer Vision*. Springer, 2024, pp. 91–106.

[33] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[34] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.

[35] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[36] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[37] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[40] Y. Zhang, X. Guo, M. Poggi, Z. Zhu, G. Huang, and S. Mattoccia, "Completionformer: Depth completion with convolutions and vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 527–18 536.

[41] J. Tang, F.-P. Tian, B. An, J. Li, and P. Tan, "Bilateral propagation network for depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9763–9772.