applied sciences

MDPI

Article

Dynamic Street-Scene Reconstruction with Semantic Priors and Temporal Constraints

Qingwu Duan [1,2], Kaichen Ren [1,2], Mingsheng Huang [2,3], Jie Liu [2,3], Siyu Li [2,3] and Sili Gao [2,3,*]

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China; duanqw2023@shanghaitech.edu.cn (Q.D.)
[2] Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: gaosili@mail.sitp.ac.cn

Featured Application

High-fidelity and temporally stable reconstruction of dynamic road scenes for autonomous-driving simulation, rare-event replay, and closed-loop evaluation.

Abstract

Dynamic street-scene reconstruction from sparse viewpoints over long temporal spans is challenged by temporal instability, ghosting near occlusions, and background drift. This paper presents SPT-Gauss (Semantic Prior and Temporal constraint-enhanced Gaussian splatting), a Gaussian-splatting framework that improves dynamic reconstruction without object-level annotations by combining dense semantic priors with lightweight, parameter-level temporal regularization. SPT-Gauss distills per-pixel semantic features from a frozen 2D foundation model into 4D Gaussian primitives, estimates static and dynamic regions via a dual-evidence motion mask, and regularizes temporal parameters through a semantic-guided velocity constraint and a static-lifetime prior to suppress spurious background motion. Experiments on the Waymo Open Dataset and KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute) show consistent improvements over representative baselines in both 4D reconstruction and novel-view synthesis, with reduced temporal artifacts and improved fidelity in motion-challenging regions.

Keywords: dynamic scene reconstruction; 4D Gaussian splatting; temporal consistency; semantic distillation; novel view synthesis; autonomous driving

Academic Editor: Pedro Couto

Received: 2 February 2026
Revised: 25 February 2026
Accepted: 26 February 2026
Published: 2 March 2026

1. Introduction

Autonomous driving in open-road environments increasingly benefits from dynamic 3D representations that are measurable, renderable, and editable throughout the perception–prediction–planning pipeline [1]. High-fidelity reconstruction provides geometric priors and occlusion completion for downstream tasks such as detection, segmentation, and tracking, while photorealistic neural rendering enables the replay of rare events for adversarial testing and closed-loop evaluation. From a systems perspective, incremental updates and long-term maintenance of 3D/4D assets across edge, roadside, and cloud deployments can reduce operational costs and improve robustness in complex traffic scenarios.

Novel-view synthesis (NVS) and neural rendering have progressed from implicit volumetric radiance fields such as Neural Radiance Fields (NeRF) [1] to a variety of acceleration and sparsification techniques, including multi-resolution hash grids, explicit radiance

tensors, and factorized plane representations [2–5]. More recently, explicit point-based formulations have drawn increasing attention due to their efficiency and editability. In particular, 3D Gaussian Splatting (3DGS) enables real-time rendering and fast convergence for static scenes via differentiable rasterization and depth-ordered compositing [6]. Extending 3DGS to dynamic scenes typically requires introducing time-varying parameters (e.g., positions, rotations, opacity, or learned deformations) and enforcing temporal regularization to reduce drift, ghosting, and flickering across frames [7–16].

Among dynamic Gaussian formulations, periodic-vibration models provide a compact way to incorporate time with minimal changes to the explicit representation. They parameterize each primitive with differentiable temporal oscillations and lifetime decay, allowing a static background and moving agents to be represented under a unified set of temporal parameters while preserving efficiency and editability [12,17]. Related directions jointly estimate scene motion and appearance using neural flow or spatiotemporal Gaussian coupling, enabling label-efficient dynamic reconstruction and multimodal synthesis (e.g., RGB, depth, and optical flow) [13,18]. Instruction- or constraint-driven editing has also been explored on Gaussian backbones [19].

Dynamic reconstruction in urban driving remains challenging due to a large-scale structure, persistent motion, frequent occlusions, and sparse viewpoints. Methods with object-level priors decompose the background and agents using 3D boxes, masks, or tracking, offering controllable rendering and editing [20–23]. However, such pipelines depend on substantial annotations and engineering effort. Weakly supervised and self-supervised approaches reduce reliance on object-level labels via scene decomposition, canonicalization, and deformation modeling, but they often suffer from background drift and temporal instability over long sequences or under heavy occlusions. Recent work mitigates these issues with staged training, temporal consistency losses, and geometry-aware regularization [7,24–27]. In parallel, feature distillation injects high-level semantics from 2D foundation models into explicit 3D/4D representations, supporting retrieval, editing, and weak-label propagation [28,29]. Large-scale encoders such as Contrastive Language–Image Pretraining (CLIP) [30], DINOv2 (self-DIstillation with NO labels, version 2) [31], the Segment Anything Model (SAM) [32], Language-driven Segmentation (LSeg) [33], and Mask2Former [34,35] provide strong teacher signals for dense semantics.

These observations motivate a practical question: how can one improve temporal stability and dynamic reconstruction quality in driving scenes without relying on object-level annotations while preserving the efficiency and editability of Gaussian splatting? In this work, we propose SPT-Gauss, a dynamic Gaussian framework that integrates semantic priors with lightweight parameter-level temporal constraints under a periodic-vibration model. Figure 1 illustrates the multimodal outputs of SPT-Gauss on a representative driving scene. The rendered RGB (c) faithfully reconstructs the ground-truth view (a) while the semantic maps (b, d) validate successful 2D-to-4D distillation from the frozen teacher to 4D Gaussians. The dual-evidence motion mask (g) reliably separates dynamic agents from the static background, enabling clean background rendering (e) with moving objects removed. Per-pixel depth (f) and velocity magnitude (h) provide complementary geometric and motion cues for downstream tasks. The framework consists of three components. First, it performs 2D-to-4D semantic distillation by transferring pixel-aligned features from a frozen 2D foundation model to 4D Gaussians, equipping each primitive with a compact semantic vector. Second, it constructs a dual-evidence motion mask by combining teacher–student feature discrepancy with semantic priors, and stabilizes the separation with temporal voting. Third, it introduces two parameter-level temporal constraints, including a semantic-guided velocity constraint and a static-lifetime prior, which regularize temporal parameters to suppress background drift and reduce long-sequence jitter.

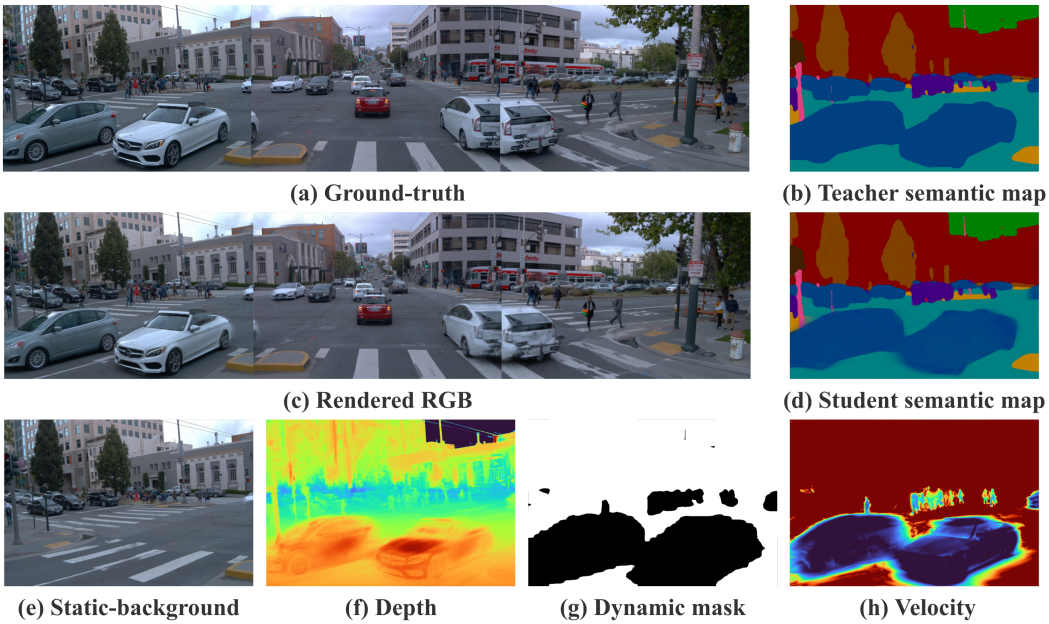

Figure 1. Overview of SPT-Gauss. (**a**) Ground-truth view; (**b**) teacher semantic map (distinct colors represent different semantic classes); (**c**) rendered RGB; (**d**) student semantic map after 2D-to-4D distillation (same color coding as (**b**)); (**e**) stabilized static-background rendering enabled by temporal constraints; (**f**) pseudo-colored depth (warm colors indicate closer regions); (**g**) motion mask from dual-evidence fusion (bright regions indicate detected motion); (**h**) per-pixel velocity magnitude (brighter values indicate higher velocity).

Contributions are summarized as follows: (1) We present a 2D-to-4D semantic-distillation scheme that transfers dense semantics from 2D foundation models to 4D Gaussians, yielding compact per-primitive semantic vectors for analysis and editing. (2) We propose a dual-evidence motion mask that fuses teacher–student feature discrepancy with semantic priors, and apply temporal voting to obtain robust static/dynamic separation for supervision routing and temporal regularization. (3) We introduce parameter-level temporal constraints, including a semantic-guided velocity constraint and a static-lifetime prior, to reduce background drift and temporal jitter and to improve long-sequence stability and rendering quality.

The remainder of the paper is organized as follows. Section 2 describes the proposed framework, including semantic distillation, dual-evidence motion masking, temporal constraints, and optimization. Section 3 presents quantitative and qualitative evaluations on Waymo Open and KITTI. Section 4 provides ablation studies, mechanism analysis, and a discussion of limitations. Section 5 concludes the paper.

2. Materials and Methods

This section describes the proposed semantic- and temporal-prior-driven dynamic Gaussian framework, SPT-Gauss. As shown in Figure 2, the pipeline consists of five parts: preliminaries, 2D-to-4D semantic distillation, dual-evidence motion masking, semantics-driven temporal constraints, and optimization. The framework integrates dense 2D semantic priors with lightweight temporal modeling under a Periodic-Vibration Gaussian representation, and it is trained without object-level annotations.

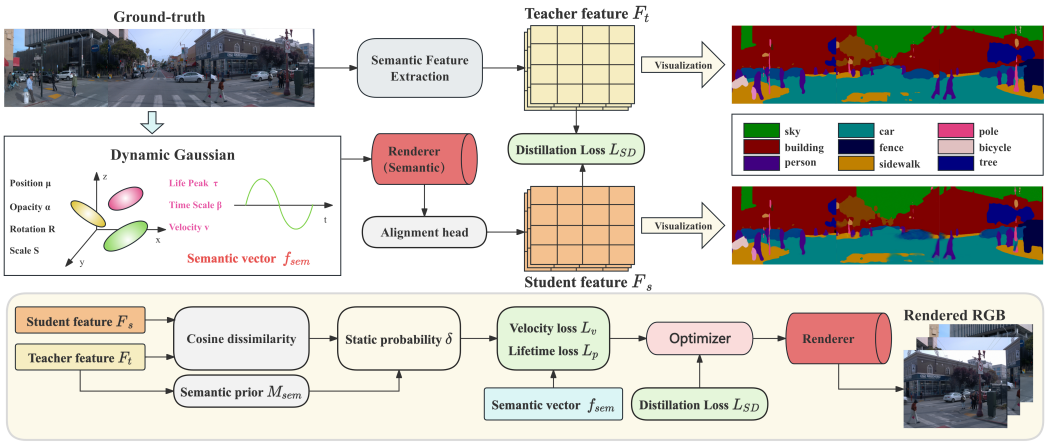

Figure 2. Overall pipeline of SPT-Gauss. The method distills dense 2D semantics into 4D Gaussians, estimates motion masks via dual-evidence fusion, and applies parameter-level temporal constraints for improved temporal stability in dynamic street scenes. Different colors represent different pipeline modules.

2.1. Preliminaries: 3DGS and PVG

The 3D Gaussian Splatting (3DGS) represents a scene using a set of anisotropic Gaussian primitives. Each primitive stores a 3D center, anisotropic scale and rotation, opacity, and appearance parameters. Rendering is performed by differentiable rasterization and depth-ordered alpha compositing [6,36]. A single 3D Gaussian is defined as

$$G_i(\boldsymbol{x}) = \exp\Big(-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\Big), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^3$ is an arbitrary 3D query point, $\boldsymbol{\mu}_i$ is the 3D center, and $\boldsymbol{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}_i^\top$ is the covariance parameterized by rotation $\mathbf{R}_i$ and scale $\mathbf{S}_i$. After projection to the image plane, the 2D covariance is approximated by

$$\boldsymbol{\Sigma}_i' = \mathbf{J}\,\mathbf{W}\,\boldsymbol{\Sigma}_i\,\mathbf{W}^\top\mathbf{J}^\top, \tag{2}$$

where $\mathbf{W}$ is the world-to-camera extrinsic transform and $\mathbf{J}$ is the Jacobian approximation of perspective projection. For a pixel, the rendered color is computed by depth-sorted α compositing:

$$C = \sum_{i=1}^{N} T_i\,\alpha_i\,c_i, \qquad T_i = \prod_{j<i}(1-\alpha_j), \tag{3}$$

where N is the number of depth-sorted Gaussians overlapping the pixel, T_i is the accumulated transmittance (i.e., the product of $(1-\alpha_j)$ over all preceding primitives), α_i depends on the primitive opacity and its projected footprint at the pixel, and c_i denotes the appearance.

A standard 3DGS is time-invariant, which limits its applicability to road scenes with persistent motion (e.g., vehicles and pedestrians). Periodic-Vibration Gaussians (PVGs) introduce a compact temporal parameterization, in which each primitive follows a differentiable oscillatory trajectory and a lifetime-controlled visibility decay around a peak time τ [17]. For each primitive, PVG defines a temporal period l, a velocity vector v, a lifetime scale β, and a base opacity o:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}(t) &= \boldsymbol{\mu} + \frac{l}{2\pi}\sin\Big(\frac{2\pi(t-\tau)}{l}\Big)\,\boldsymbol{v}, \\ \tilde{o}(t) &= o\,\exp\Big(-\tfrac{1}{2}(t-\tau)^2\,\beta^{-2}\Big). \end{aligned} \tag{4}$$

The primitive state at time t is $H(t) = \{\tilde{\boldsymbol{\mu}}(t), \mathbf{q}, \mathbf{s}, \tilde{o}(t), \mathbf{c}\}$, and the rendered image is

$$\hat{I}_t = \mathrm{Render}\big(\{H_i(t)\}_{i=1}^{N}; \mathbf{K}_t, \mathbf{E}_t\big), \tag{5}$$

with intrinsics $\mathbf{K}_t$ and extrinsics $\mathbf{E}_t$. We define the staticness ratio as

$$\rho = \beta / l, \tag{6}$$

where a larger ρ indicates longer visibility relative to the oscillation period. When $\boldsymbol{v} = \mathbf{0}$ and $\rho \to \infty$, PVG reduces to standard 3DGS. In this way, static and dynamic content share the same rendering backbone and are differentiated only by temporal parameters $\{\boldsymbol{v}, \beta, l, \tau\}$.

2.2. 2D-to-4D Semantic Distillation (SD)

Reconstruction losses alone may not reliably disentangle camera-induced appearance changes from real-world motion. We therefore distill dense semantic features from a frozen 2D foundation model into the 4D Gaussian representation so that each primitive carries a compact semantic vector used for prior injection and motion masking.

We adopt Language-driven Segmentation (LSeg) as the teacher [33], whose pixel features are aligned to the CLIP text space [30]. Given an RGB frame I_t at time t, the teacher feature map is

$$F_t = \mathrm{LSeg}(I_t). \tag{7}$$

On the student side, each primitive is assigned a learnable semantic vector $f_{\mathrm{sem},i}$. Analogous to RGB rendering, the student semantic feature at pixel $\mathbf{p}$ is computed by alpha-composited aggregation over the depth-sorted visible set $\mathcal{V}(\mathbf{p}, t)$:

$$F_s(\mathbf{p}, t) = \sum_{i \in \mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p}, t)\, f_{\mathrm{sem},i}, \tag{8}$$

where $w_i(\mathbf{p}, t)$ are the standard compositing weights induced by α-blending (i.e., $w_i = T_i \alpha_i$). To match the teacher feature dimensionality, we apply a lightweight linear projection head $U(\cdot)$:

$$\tilde{F}_s(\mathbf{p}, t) = U\big(F_s(\mathbf{p}, t)\big). \tag{9}$$

We minimize a pixel-wise L_1 distillation loss:

$$\mathcal{L}_{\mathrm{SD}} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \big\| \tilde{F}_s(\mathbf{p}, t) - F_t(\mathbf{p}) \big\|_1, \tag{10}$$

where Ω is the pixel set at the current resolution. After optimization, semantics are embedded into per-primitive vectors $f_{\mathrm{sem},i}$ and propagated over time through PVG.

2.3. Dual-Evidence Motion Mask (DEMM)

Using the teacher features F_t and the rendered student semantics $\tilde{F}_s$, we estimate a motion mask from two complementary cues: (i) teacher–student feature discrepancy and (ii) a semantic prior indicating regions that are likely static. The fusion yields a soft static probability map $\delta(\mathbf{p}, t) \in (0, 1)$, which is used as a differentiable weight in temporal constraints; binary masks can be obtained for visualization or evaluation.

2.3.1. Teacher–Student Feature Discrepancy

For static surfaces, multi-frame observations correspond to the same world points and the student features are expected to match the teacher features. Pixels on moving

objects or near occlusion boundaries tend to show discrepancies. We define a pixel-wise cosine dissimilarity:

$$D(\mathbf{p}, t) = 1 - \cos\big(\tilde{F}_s(\mathbf{p}, t),\, F_t(\mathbf{p})\big). \tag{11}$$

A larger $D(\mathbf{p}, t)$ indicates a higher likelihood of motion or inconsistent alignment.

2.3.2. Semantic Prior and Fusion

Using the teacher model, we obtain class scores $\{S_k(\mathbf{p}, t)\}$ and form a soft static prior by summing scores over a set of static-leaning categories $\mathcal{C}_{\text{stat}}$. In all experiments, $\mathcal{C}_{\text{stat}}$ consists of nine categories: *road*, *sidewalk*, *building*, *wall*, *fence*, *vegetation*, *sky*, *pole*, and *traffic sign*. These categories correspond to scene elements that are geometrically stationary in world coordinates and are consistently present across urban driving sequences. Categories associated with potentially movable objects (e.g., *car*, *person*, *bicycle*) are excluded. The soft static prior is computed as:

$$M_{\text{sem}}(\mathbf{p}, t) = \sum_{k \in \mathcal{C}_{\text{stat}}} S_k(\mathbf{p}, t). \tag{12}$$

We fuse the two cues with a logistic regressor to produce the static probability:

$$\delta(\mathbf{p}, t) = \sigma\big(a \cdot (1 - D(\mathbf{p}, t)) + b \cdot M_{\text{sem}}(\mathbf{p}, t) + c\big), \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid and $a, b, c \in \mathbb{R}$ are learnable scalars. For binary masks, we threshold δ:

$$M^0_{\text{stat}}(\mathbf{p}, t) = \mathbf{1}\big(\delta(\mathbf{p}, t) > \tau_s\big), \qquad M^0_{\text{dyn}}(\mathbf{p}, t) = 1 - M^0_{\text{stat}}(\mathbf{p}, t), \tag{14}$$

with $\tau_s = 0.5$ by default.

Temporal Voting (Conservative Merge)

Single-frame motion masks can be noisy due to transient occlusions, shadow boundaries, or teacher-feature inconsistencies. To improve robustness, we perform temporal voting over a window $\mathcal{T}(t)$ centered at t with half-width r (i.e., $\mathcal{T}(t) = \{t-r, \ldots, t, \ldots, t+r\}$; we use $r = 2$ by default, covering five consecutive frames). We adopt a conservative merge strategy: a pixel is marked static only if it is consistently static across the window (intersection), while it is marked dynamic if it is predicted dynamic in any frame (union). This asymmetric design is deliberate: for temporal regularization, incorrectly labeling a dynamic pixel as static (false negative) is more harmful than the reverse, because it would suppress legitimate motion via the velocity constraint. The intersection rule for static labels and the union rule for dynamic labels thus minimize false negatives in dynamic detection at the cost of slightly over-segmenting dynamic regions, which is a safer trade-off for reconstruction quality.

$$M_{\text{stat}}(\mathbf{p}, t) = \bigwedge_{t' \in \mathcal{T}(t)} M^0_{\text{stat}}(\mathbf{p}, t'), \qquad M_{\text{dyn}}(\mathbf{p}, t) = \bigvee_{t' \in \mathcal{T}(t)} M^0_{\text{dyn}}(\mathbf{p}, t'). \tag{15}$$

In training, we use the soft weight $\delta(\mathbf{p}, t)$ (not binarized) for differentiability; the temporally voted binary masks are used for qualitative visualization and optional evaluation.

2.4. Semantics-Driven Temporal Constraints

We impose temporal constraints to suppress spurious motion on static regions while maintaining temporal coherence for dynamic targets. The constraints are applied at the parameter level of PVG by using (i) the pixel-level static probability $\delta(\mathbf{p}, t)$ and (ii) a

back-projected per-primitive static weight w_i^{stat}, which measures how strongly primitive i contributes to pixels with high static probability.

2.4.1. Back-Projected Static Weight

For a frame at time t, we define

$$w_i^{\text{stat}}(t) = \frac{\sum_{\mathbf{p}\in\Omega} w_i(\mathbf{p},t)\,\delta(\mathbf{p},t)}{\sum_{\mathbf{p}\in\Omega} w_i(\mathbf{p},t) + \varepsilon}, \tag{16}$$

where $w_i(\mathbf{p},t)$ are compositing weights, Ω is the pixel domain, and ε is a small constant. When summing losses across a mini-batch, we use w_i^{stat} averaged over frames in the batch.

2.4.2. Semantic Velocity Constraint (SVC)

We apply a semantic gate to modulate the PVG velocity magnitude. For primitive i, we compute a gate from its semantic vector:

$$g_i = \sigma(\mathbf{w}_g^\top f_{\text{sem},i} + b_g) \in (0,1), \qquad \boldsymbol{v}_i^{\text{eff}} = g_i\,\boldsymbol{v}_i, \tag{17}$$

where $\mathbf{w}_g$ and b_g are learnable parameters. The PVG trajectory uses $\boldsymbol{v}_i^{\text{eff}}$ (all other rendering components remain unchanged). To measure projected motion, we compute a symmetric-step displacement in the image plane using a fixed Δ:

$$V(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,\big\|\Pi\big(\tilde{\boldsymbol{\mu}}_i(t+\Delta)\big) - \Pi\big(\tilde{\boldsymbol{\mu}}_i(t-\Delta)\big)\big\|_1, \tag{18}$$

where $\Pi(\cdot)$ denotes camera projection. We penalize projected motion on pixels with high static probability:

$$\mathcal{L}_v = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega} \delta(\mathbf{p},t)\cdot V(\mathbf{p},t). \tag{19}$$

2.4.3. Static-Lifetime Prior (SLP)

Constraining instantaneous speed may still allow slow drift over long sequences. We therefore regularize the PVG staticness ratio for primitives that contribute to static regions. For each primitive, define

$$\rho_i = \beta_i / l_i, \tag{20}$$

and impose a lower bound $\rho^\star$ weighted by the back-projected static weight:

$$\mathcal{L}_\rho = \sum_i w_i^{\text{stat}}\,\max\big(0,\,\rho^\star - \rho_i\big), \tag{21}$$

where $\rho^\star > 0$ controls the preference for a long lifetime relative to the oscillation period.

2.5. Optimization

We optimize the Gaussian parameters using a reconstruction loss combined with semantic distillation and temporal constraints. The photometric term is a weighted sum of pixel-wise L_1 and SSIM:

$$\mathcal{L}_{\text{rgb}} = \lambda_1\,\|I_t - \hat{I}_t\|_1 + \lambda_{\text{ssim}}\,(1 - \text{SSIM}(I_t, \hat{I}_t)). \tag{22}$$

In addition, LiDAR point clouds are projected onto the camera plane to form sparse inverse-depth maps [37,38]. Let $Z_t(\mathbf{p})$ denote the sparse inverse depth and $M_{\text{dep}}(\mathbf{p})$ the validity mask. We use a masked L_1 depth loss:

$$\mathcal{L}_{\text{dep}} = \frac{1}{\sum_{\mathbf{p}} M_{\text{dep}}(\mathbf{p}) + \varepsilon} \sum_{\mathbf{p}\in\Omega} M_{\text{dep}}(\mathbf{p}) \left|\hat{Z}_t(\mathbf{p}) - Z_t(\mathbf{p})\right|, \tag{23}$$

where $\hat{Z}_t(\mathbf{p})$ is the rendered inverse depth.

We regularize semantic vectors and PVG velocities with

$$\mathcal{L}_{\text{reg}} = \sum_i \|f_{\text{sem},i}\|_2^2 + \sum_i \|\boldsymbol{v}_i\|_2^2. \tag{24}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{dep}}\mathcal{L}_{\text{dep}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{25}$$

To avoid over-penalizing motion before geometry and appearance stabilize, we apply a warm-start schedule. Specifically, because temporal constraints applied too early can interfere with coarse geometry convergence, λ_v increases linearly from 0 to 0.5 over the first 5k iterations, allowing photometric and depth losses to establish scene structure first. $\lambda_{\text{SD}} = 1.0$ is kept constant throughout training. $\lambda_\rho \equiv 0.15$ is chosen as a moderate value to balance between suppressing background drift and preserving the adaptability of dynamic primitives. The staticness lower bound $\rho^\star$ increases linearly from 1.0 to 1.5 over the first 15k iterations, reflecting that the distinction between truly static and slowly drifting primitives becomes more reliable as training proceeds. Unless otherwise specified, we use $\lambda_{\text{reg}} = 1 \times 10^{-4}$ for $\sum_i \|f_{\text{sem},i}\|_2^2$ and $\sum_i \|\boldsymbol{v}_i\|_2^2$ (implemented by scaling the corresponding terms), and set λ_{dep} in $[0.1, 0.3]$ depending on scene sparsity. All loss weights were calibrated on held-out Waymo sequences to maintain balanced gradient magnitudes across terms. We further verified robustness by sweeping $\lambda_v \in \{0.1, 0.3, 0.5, 0.7\}$ and $\lambda_\rho \in \{0.05, 0.10, 0.15, 0.20\}$; performance remained within ± 0.3 dB PSNR across these ranges.

3. Results

3.1. Experimental Setup

3.1.1. Datasets

We evaluate SPT-Gauss on two widely used large-scale road-scene benchmarks, the Waymo Open Dataset and KITTI [37,38]. Waymo Open provides synchronized multi-camera and multi-LiDAR sequences with accurate timestamps and calibration. Following PVG [17], we select four challenging urban sequences. Three forward-facing cameras are used for training at 960×640, and a fourth camera is held out for novel-view synthesis (NVS) evaluation. Each selected Waymo sequence spans approximately 200 frames, yielding roughly 600 training images from the three cameras per sequence. KITTI provides multiview camera streams and vehicle poses. Following the SUDS protocol [39], we select motion-rich sequences and use the left–right stereo pair (1242×375) for training and evaluation. Each KITTI sequence contains approximately 60–120 stereo pairs, resulting in roughly 120–240 training images per sequence.

3.1.2. Evaluation Protocols

We use two evaluation protocols on Waymo (v1.4.2) for clarity and reproducibility. (i) *Main comparison protocol:* All methods in Tables 1 and 2 are evaluated on the same set of four sequences following PVG [17], with identical camera splits for reconstruction and NVS. (ii) *Ablation protocol:* The ablation study in Table 3 is conducted on a reduced subset of Waymo sequences to enable faster and controlled analysis of module behaviors. Therefore, absolute values in Table 3 are intended for relative comparison among ablated variants only and are not directly comparable to the main comparison results.

Table 1. Quantitative comparison on Waymo Open and KITTI under the main comparison protocol. We report 4D reconstruction and NVS performance. A higher result is better for PSNR/SSIM (↑); a lower one is better for LPIPS (↓). Bold values indicate the best result in each column.

Method	Waymo Open 4D Reconstruction			Waymo Open NVS			KITTI 4D Reconstruction			KITTI NVS		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
3DGS [6]	27.99	0.866	0.293	25.08	0.822	0.319	21.02	0.811	0.202	19.54	0.776	0.224
StreetSurf [40]	26.70	0.846	0.372	23.78	0.822	0.401	24.14	0.819	0.257	22.48	0.763	0.304
EmerNeRF [27]	28.11	0.786	0.373	25.92	0.763	0.384	26.95	0.828	0.218	25.24	0.801	0.237
SUDS [39]	28.83	0.805	0.317	25.36	0.783	0.384	28.83	0.917	0.147	26.07	0.797	0.131
MARS [41]	21.81	0.681	0.430	20.69	0.636	0.453	27.96	0.900	0.185	24.31	0.845	0.160
CoDa-4DGS [26]	30.16	0.898	0.240	26.04	0.857	0.269	30.53	0.926	0.095	25.48	0.871	0.142
PVG [17]	32.46	0.910	0.229	28.11	0.849	0.279	32.83	0.937	0.070	27.43	0.879	0.114
SPT-Gauss (Ours)	**34.12**	**0.926**	**0.189**	**30.23**	**0.905**	**0.197**	**34.50**	**0.955**	**0.057**	**29.80**	**0.903**	**0.108**

Table 2. Performance on dynamic regions of Waymo Open under the main comparison protocol. Region-wise scores are computed using ground-truth camera segmentation masks provided by Waymo (evaluated on labeled frames only). Bold values indicate the best result in each column.

Method	D-PSNR (↑)	D-SSIM (↑)
3DGS [6]	18.65	0.803
EmerNeRF [27]	24.56	0.819
CoDa-4DGS [26]	26.08	0.871
PVG [17]	27.60	0.862
SPT-Gauss (Ours)	**30.82**	**0.921**

Table 3. Ablation on Waymo Open under the ablation protocol (reduced subset). Absolute values are not directly comparable to Table 1. Bold values indicate the best result in each column.

Setting	PSNR	SSIM	LPIPS	D-PSNR	D-SSIM
w/o SD	33.28	0.956	0.072	32.89	0.951
w/o DEMM	34.20	0.965	0.066	33.80	0.958
w/o SVC	35.02	0.969	0.066	34.22	0.966
w/o SLP	35.28	0.969	0.062	34.45	0.967
Full	**35.54**	**0.971**	**0.060**	**34.63**	**0.970**

3.1.3. Metrics

We report PSNR (↑), SSIM (↑), and LPIPS (↓) [42,43] for 4D reconstruction and NVS. PSNR (Peak Signal-to-Noise Ratio) measures pixel-level fidelity by computing the ratio between the maximum possible signal power and the mean squared error; it is sensitive to overall luminance and color accuracy but may not fully capture perceptual quality. SSIM (Structural Similarity Index Measure) [42] evaluates structural consistency by jointly considering luminance, contrast, and structural correlations within local patches, thereby better reflecting human perception of image degradation than purely pixel-wise metrics. LPIPS (Learned Perceptual Image Patch Similarity) [43] leverages deep features extracted from a pretrained network to quantify perceptual distance, capturing high-level texture and semantic discrepancies that neither PSNR nor SSIM can fully account for. Together, these three metrics provide complementary views of rendering quality: PSNR captures signal-level accuracy, SSIM reflects structural fidelity, and LPIPS assesses perceptual realism, which is the standard evaluation protocol adopted by the novel-view synthesis community. To analyze reconstruction quality across different regions, we additionally report static-region PSNR (S-PSNR) and dynamic-region PSNR (D-PSNR) computed using estimated motion masks. These region-wise metrics provide complementary insights into static/dynamic behavior; they are used consistently across methods under the same protocol. We compare against representative NeRF-based and Gaussian-splatting-based

baselines, including 3DGS [6], SUDS [39], StreetSurf [40], EmerNeRF [27], MARS [41], PVG [17], and CoDa-4DGS [26].

3.1.4. Implementation Details

For baseline methods, we use the officially released code and pretrained configurations provided by the respective authors. All baselines are trained on the same data splits and evaluated under the same protocol as SPT-Gauss to ensure a fair comparison. For methods that do not provide default settings for Waymo or KITTI, we adopt the hyperparameters recommended in their original papers and, where necessary, tune the learning rate and batch size on a held-out sequence to obtain competitive performance. All reported baseline numbers are obtained from our own re-training runs rather than copied from prior publications because differences in data preprocessing, resolution, and evaluation code can introduce non-negligible discrepancies. The only exception is SUDS on KITTI, for which we use the authors' released checkpoints since re-training under the same protocol produced comparable results (within 0.1 dB PSNR).

All experiments are performed on one NVIDIA vGPU with 48 GB memory using Python 3.9, PyTorch 2.0, and CUDA 11.8. Gaussians are initialized from the ego-LiDAR point cloud (instead of SfM). Time-related parameters are initialized neutrally ($v_i = \mathbf{0}$, $l_i = 1$, $\beta_i = 1$). For stable early optimization, the linear layers used for semantic gating and evidence fusion are initialized to produce near-neutral outputs (so that $g_i \approx 0.5$ and $\delta \approx 0.5$ at initialization). The LSeg teacher is frozen; its per-pixel features are cached offline (optionally quantized to 8-bit) to reduce memory usage during training. We use Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of 2×10^{-3}, cosine-decayed to 1×10^{-4} [44,45]. The batch size is 2, gradients are clipped at 1.0, and we adopt a two-stage resolution curriculum: a pre-warm phase at one-quarter resolution followed by a ramp to full resolution to jointly optimize distillation and temporal constraints. Loss weights follow the warm-start schedule in Equation (25).

3.2. Quantitative Evaluation

Table 1 summarizes the main results on Waymo Open and KITTI under the main comparison protocol. Across both benchmarks, SPT-Gauss improves PSNR/SSIM and reduces LPIPS compared with recent Gaussian-based baselines such as PVG and CoDa-4DGS. Table 2 reports performance on dynamic regions of Waymo. SPT-Gauss improves D-PSNR by $+3.22$ dB and D-SSIM by $+0.059$ over PVG under the same evaluation protocol, suggesting that incorporating semantic priors and parameter-level temporal constraints can improve reconstruction quality in motion-dominant areas.

3.3. Qualitative Evaluation

Figure 3 presents qualitative comparisons on Waymo Open and KITTI. Gaussian-based dynamic representations such as PVG and CoDa-4DGS can exhibit temporal artifacts in challenging regions, including ghosting near occlusion boundaries and background drift around moving objects. SPT-Gauss reduces these artifacts in many cases and produces visually more stable renderings, especially in static structures while retaining motion details.

Figure 4 further compares dynamic–static decomposition between PVG and SPT-Gauss. PVG may show motion leakage into the static layer and aliasing artifacts in the dynamic layer. In contrast, SPT-Gauss yields cleaner separation in the shown examples, with dynamic components concentrating on moving objects and static components maintaining sharper textures.

Figure 3. Qualitative comparison results. The first two rows are from Waymo Open, and the last two rows are from KITTI. From left to right: Ground Truth, 3DGS, PVG, CoDa-4DGS, and SPT-Gauss.

Figure 4. Comparison of dynamic and static decomposition. The first column shows the ground truth (**top**) and the full reconstruction of SPT-Gauss (**bottom**). The remaining columns show decomposition results of PVG and SPT-Gauss, where the top and bottom rows represent dynamic and static components, respectively.

4. Discussion

4.1. Ablation Study and Mechanism Analysis

We conduct an ablation study on Waymo under the ablation protocol (reduced subset) to analyze the contribution of each component. All training configurations are kept identical across variants, except for disabling the corresponding module. Because the submodules in SPT-Gauss are coupled (e.g., motion masking and temporal constraints rely on distilled semantics), removing an upstream component can change the behavior of downstream

modules. Therefore, the ablation results are primarily used to compare variants within the same protocol and to interpret module interactions.

Table 3 summarizes the ablation results on the reduced subset. Disabling any component degrades performance, indicating complementary contributions of semantic distillation, motion masking, and parameter-level constraints. Removing SD reduces reconstruction metrics, consistent with the role of dense semantic guidance. Removing DEMM or SVC mainly affects dynamic-region performance, suggesting that reliable motion separation and velocity regularization are important for motion-dominant areas. Removing SLP tends to reduce long-term stability, consistent with its role in discouraging slow drift.

To assess the sensitivity of the motion mask and reconstruction quality to the choice of $\mathcal{C}_{\text{stat}}$, we evaluate three configurations under the ablation protocol: the default nine-class set, a seven-class subset that excludes *pole* and *traffic sign*, and a five-class subset that further excludes *vegetation* and *sky*. Table 4 reports the results. The main performance gap arises when removing vegetation and sky (from seven to five classes), which weakens the static prior on large homogeneous regions that occupy a substantial fraction of driving-scene images. In contrast, removing pole and traffic sign (from nine to seven classes) has a negligible effect, as these thin structures contribute limited pixel area. Overall, PSNR varies by less than 0.4 dB across configurations, indicating moderate robustness to the exact class composition.

Table 4. Sensitivity to the static class set $\mathcal{C}_{\text{stat}}$ (ablation protocol). Bold values indicate the best result in each column.

$\mathcal{C}_{\text{stat}}$	# Classes	PSNR	SSIM	D-PSNR
w/o veg., sky, pole, sign	5	35.18	0.968	34.25
w/o pole & sign	7	35.52	0.971	34.61
Default (all static)	9	**35.54**	**0.971**	**34.63**

We further examine the effect of the voting half-width r and the static threshold τ_s. Table 5 reports results under the ablation protocol with varying $r \in \{1, 2, 3\}$ (at fixed $\tau_s = 0.5$) and $\tau_s \in \{0.3, 0.5, 0.7\}$ (at fixed $r = 2$). Increasing r from 1 to 2 yields a noticeable improvement by suppressing single-frame noise, while $r = 3$ offers diminishing returns and risks over-smoothing near motion boundaries. The threshold τ_s has a mild effect: lower values classify more pixels as static, which strengthens temporal regularization but risks suppressing slow-moving objects, while higher values are more permissive toward motion. Performance remains within 0.3 dB PSNR across the tested range, and the default settings ($r = 2$, $\tau_s = 0.5$) offer a balanced trade-off.

Table 5. Sensitivity to temporal voting half-width r and static threshold τ_s (ablation protocol). Bold values indicate the default setting and the best result in each column.

r	τ_s	PSNR	SSIM	D-PSNR
1	0.5	35.32	0.969	34.38
2	**0.5**	**35.54**	**0.971**	**34.63**
3	0.5	35.48	0.970	34.55
2	0.3	35.41	0.970	34.42
2	**0.5**	**35.54**	**0.971**	**34.63**
2	0.7	35.38	0.969	34.51

To illustrate the role of semantic distillation in motion estimation, Figure 5 compares motion masks with and without SD. Without semantic guidance, masks can be noisier

and may activate in background regions, while enabling SD typically yields more compact dynamic regions and cleaner static areas in the shown examples.

Figure 5. Motion-mask comparison with and without semantic distillation (SD). The first column shows the ground truth; the next two columns correspond to results without SD (w/o SD) and with SD (w/ SD), respectively.

To assess the robustness of the reported improvements, Table 6 presents per-sequence PSNR on the four Waymo sequences used in the main comparison, together with the mean and standard deviation across sequences. SPT-Gauss consistently outperforms PVG on every individual sequence, and the standard deviation of the per-sequence PSNR gain is moderate, indicating that the improvement is not driven by a single outlier sequence.

Table 6. Per-sequence PSNR (4D reconstruction) on Waymo Open. Δ denotes the per-sequence gain of SPT-Gauss over PVG.

Method	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Mean	Std.
PVG	31.25	32.66	32.55	33.37	32.46	0.88
SPT-Gauss	32.73	34.13	34.08	35.54	34.12	1.15
Δ	+1.48	+1.47	+1.53	+2.17	+1.66	–

4.2. Limitations and Future Work

SPT-Gauss relies on the quality and domain coverage of the teacher semantic model, and performance can degrade under conditions that are under-represented by the teacher or training data (e.g., extreme lighting changes, adverse weather, or sensor noise). In addition, periodic temporal parameterization may be less expressive for non-periodic, abrupt motions.

We identify three challenging scenarios that merit explicit discussion. (i) *Adverse lighting.* Under low-light conditions, glare, or rapid illumination transitions (e.g., tunnel entry/exit), both the LSeg teacher features and the photometric reconstruction loss become less reliable, which can degrade semantic distillation quality and weaken the motion mask. Incorporating illumination-invariant feature encoders or appearance embedding normalization may alleviate this issue. (ii) *Adverse weather.* Rain, fog, and snow introduce semi-transparent particles, specular reflections on wet surfaces, and reduced visibility, all of which violate the Lambertian assumption implicit in the photometric loss and may confuse the teacher–student discrepancy signal. Dedicated weather-aware data augmentation or domain-adaptive distillation would be needed to handle such conditions reliably. (iii) *Abrupt non-periodic motion.* The sinusoidal trajectory model in PVG is inherently smooth

and periodic, making it ill-suited for sudden lane changes, emergency braking, or pedestrians darting into the road. Such motions may produce ghosting or temporal smearing that the current velocity constraint cannot fully suppress. More expressive temporal bases (e.g., piecewise-linear trajectories or neural motion fields) could better accommodate these dynamics. We did not include targeted stress tests for the above scenarios in the current study, because the standard Waymo Open and KITTI benchmarks used in this work predominantly contain daytime, clear-weather sequences; constructing controlled experiments for adverse conditions would require specialized datasets (e.g., nuScenes-Night or the Boreas dataset) and is beyond the scope of this paper. Nevertheless, we consider robustness under these conditions a critical requirement for practical deployment and an important direction for future work. When a trained SPT-Gauss model is rendered from a viewpoint that deviates substantially from the training camera distribution, several types of artifacts may arise. First, regions that are occluded or unobserved during training lack sufficient multiview constraints, leading to incomplete geometry and floater artifacts in the rendered image. Second, the view-dependent appearance modeled by the Gaussian primitives may extrapolate poorly under large angular deviations, producing color shifts or specular hallucinations. Third, the 2D covariance approximation used in Gaussian splatting assumes a local affine projection, which can introduce shape distortion for primitives that are near the image boundary or at extreme off-axis angles. These errors are common to splatting-based representations and become more pronounced as the rendering viewpoint diverges from the training trajectory. The current single-stage joint optimization with warm-start scheduling is effective but may not fully exploit opportunities for faster convergence; strategies such as coarse-to-fine primitive activation, decoupled static/dynamic optimization, or adaptive loss weighting via uncertainty estimation merit further investigation. Future work will explore these directions alongside domain-robust semantic distillation, more expressive temporal parameterizations, and stronger cross-sensor constraints to improve both efficiency and robustness in diverse driving conditions.

5. Conclusions

This paper presents SPT-Gauss, a dynamic Gaussian framework that integrates dense 2D semantic priors with parameter-level temporal constraints under a periodic-vibration representation, without requiring object-level annotations. The method distills per-primitive semantic vectors from a frozen 2D foundation model and uses a dual-evidence motion mask to support static/dynamic separation. By regularizing temporal parameters via a semantic-guided velocity constraint and a static-lifetime prior, SPT-Gauss reduces background drift and improves temporal stability in long driving sequences while preserving the efficiency and editability of Gaussian splatting. Experiments on Waymo Open and KITTI demonstrate consistent improvements over representative baselines in both reconstruction and novel-view synthesis, with additional gains in motion-dominant regions under the same evaluation protocol.

Author Contributions: Conceptualization, Q.D. and S.G.; methodology, Q.D.; software, Q.D. and K.R.; validation, M.H., J.L. and S.L.; formal analysis, Q.D.; investigation, K.R.; resources, S.G.; data curation, K.R.; writing—original draft preparation, Q.D.; writing—review and editing, S.G.; visualization, Q.D.; supervision, S.G.; project administration, S.G.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available dataset Waymo Open Dataset was analyzed in this study and can be found here: https://waymo.com/open/, accessed on 31 January 2026. The publicly available dataset KITTI Vision Benchmark Suite was analyzed in this study and can be found here: https://www.cvlibs.net/datasets/kitti/, accessed on 31 January 2026.

Acknowledgments: We thank the anonymous reviewers for their constructive comments. The authors used generative AI tools (e.g., ChatGPT, powered by GPT-4, OpenAI, 2024) for language polishing and improving clarity only. No AI was used to generate or analyze data or to create/modify figures, results, or conclusions. The authors reviewed and take full responsibility for the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3DGS	3D Gaussian Splatting
PVG	Periodic-Vibration Gaussian
NVS	Novel-View Synthesis
LiDAR	Light Detection and Ranging
SPT-Gauss	Semantic Prior and Temporal Constraint-Enhanced Gaussian Splatting
SD	Semantic Distillation
DEMM	Dual-Evidence Motion Mask
SVC	Semantic Velocity Constraint
SLP	Static-Lifetime Prior
SSIM	Structural Similarity Index Measure
LPIPS	Learned Perceptual Image Patch Similarity

References

1. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 405–421. [CrossRef]
2. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* **2022**, *41*, 102. [CrossRef]
3. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: Radiance Fields without Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 5501–5510. [CrossRef]
4. Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; Su, H. TensoRF: Tensorial Radiance Fields. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 333–350. [CrossRef]
5. Fridovich-Keil, S.; Meanti, G.; Warburg, F.R.; Recht, B.; Kanazawa, A. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 12479–12488. [CrossRef]
6. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 139. [CrossRef]
7. Luiten, J.; Kopanas, G.; Leibe, B.; Ramanan, D. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In Proceedings of the International Conference on 3D Vision (3DV), Davos, Switzerland, 18–21 March 2024; pp. 800–809.
8. 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 (CVPR), Seattle, WA, USA, 17–18 June 2024; pp. 20331–20341. [CrossRef]
9. Bae, J.; Kim, S.; Yun, Y.; Lee, H.; Bang, G.; Uh, Y. Per-Gaussian Embedding-Based Deformation for Deformable 3D Gaussian Splatting. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 325–343. [CrossRef]
10. Wan, D.; Lu, R.; Zeng, G. Superpoint Gaussian Splatting for Real-Time High-Fidelity Dynamic Scene Reconstruction. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; Volume 235, pp. 49957–49972.

11. Zhu, R.; Liang, Y.; Chang, H.; Deng, J.; Lu, J.; Yang, W.; Zhang, T.; Zhang, Y. MotionGS: Exploring Explicit Motion Guidance for Deformable 3D Gaussian Splatting. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 37, pp. 101790–101817.
12. Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; Wang, X. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 20310–20320. [CrossRef]
13. Lin, Y.; Dai, Z.; Zhu, S.; Yao, Y. Gaussian-Flow: 4D Reconstruction with Dynamic 3D Gaussian Particle. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 21264–21274. [CrossRef]
14. Duan, Y.; Wei, F.; Dai, Q.; He, Y.; Chen, W.; Chen, B. 4D-Rotor Gaussian Splatting: Towards Efficient Novel View Synthesis for Dynamic Scenes. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference (ACM SIGGRAPH 2024 Conference Papers), Denver, CO, USA, 27 July–1 August 2024; pp. 1–11. [CrossRef]
15. Huang, N.; Wei, X.; Zheng, W.; An, P.; Lu, M.; Zhan, W.; Tomizuka, M.; Keutzer, K.; Zhang, S. S3Gaussian: Self-Supervised Street Gaussians for Autonomous Driving. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Vienna, Austria, 1–5 June 2026; (accepted, to appear).
16. Lan, L.; Shao, T.; Lu, Z.; Zhang, Y.; Jiang, C.; Yang, Y. 3DGS2: Near Second-order Converging 3D Gaussian Splatting. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference (ACM SIGGRAPH Conference Papers'25), Vancouver, BC, Canada, 10–14 August 2025; pp. 1–10. [CrossRef]
17. Chen, Y.; Gu, C.; Jiang, J.; Zhu, X.; Zhang, L. Periodic Vibration Gaussian: Dynamic Urban Scene Reconstruction and Real-Time Rendering. *arXiv* **2023**, arXiv:2311.18561. [CrossRef]
18. Wang, B.; Zhang, Y.; Li, J.; Yu, Y.; Sun, Z.; Liu, L.; Hu, D. SplatFlow: Learning Multi-frame Optical Flow via Splatting. *Int. J. Comput. Vis.* **2024**, *132*, 3023–3045. [CrossRef]
19. Mou, L.; Chen, J.K.; Wang, Y.X. Instruct 4D-to-4D: Editing 4D Scenes as Pseudo-3D Scenes Using 2D Diffusion. *arXiv* **2024**, arXiv:2406.09402.
20. Zhou, X.; Lin, Z.; Shan, X.; Wang, Y.; Sun, D.; Yang, M.H. DrivingGaussian: Composite Gaussian Splatting for Surrounding Dynamic Autonomous Driving Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 21634–21643. [CrossRef]
21. Yan, Y.; Lin, H.; Zhou, C.; Wang, W.; Sun, H.; Zhan, K.; Lang, X.; Zhou, X.; Peng, S. Street Gaussians for Modeling Dynamic Urban Scenes. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 156–173. [CrossRef]
22. Chen, Z.; Yang, J.; Huang, J.; de Lutio, R.; Esturo, J.M.; Ivanovic, B.; Litany, O.; Gojcic, Z.; Fidler, S.; Pavone, M.; et al. OmniRe: Omni Urban Scene Reconstruction. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025; pp. 1486–1505.
23. Liu, Y.; Luo, C.; Fan, L.; Wang, N.; Peng, J.; Zhang, Z. CityGaussian: Real-Time High-Quality Large-Scale Scene Rendering with Gaussians. In *Computer Vision—ECCV 2024, Milan, Italy, 29 September–4 October*; Springer Nature: Cham, Switzerland, 2024; pp. 265–282. [CrossRef]
24. Kim, H.; Cho, J.; Kang, S.J. Guess The Unseen: Dynamic 3D Scene Reconstruction from Partial 2D Glimpses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–18 June 2024; pp. 20018–20028. [CrossRef]
25. Peng, C.; Zhang, C.; Wang, Y.; Xu, C.; Xie, Y.; Zheng, W.; Keutzer, K.; Tomizuka, M.; Zhan, W. DeSiRe-GS: 4D Street Gaussians for Static-Dynamic Decomposition and Surface Reconstruction for Urban Driving Scenes. *arXiv* **2024**, arXiv:2411.11921.
26. Song, R.; Liang, C.; Xia, Y.; Zimmer, W.; Cao, H.; Caesar, H.; Festag, A.; Knoll, A. Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025.
27. 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. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
28. Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; Kadambi, A. Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–18 June 2024; pp. 21676–21685. [CrossRef]
29. Chen, G.; Wang, W. A Survey on 3D Gaussian Splatting. *arXiv* **2024**, arXiv:2401.03890. [CrossRef]
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning (ICML), Online, 18–24 July 2021; pp. 8748–8763.

31. 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* **2023**, arXiv:2304.07193.
32. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 4015–4026. [CrossRef]
33. Li, B.; Weinberger, K.Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven Semantic Segmentation. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022.
34. Cheng, B.; Schwing, A.G.; Kirillov, A. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021; Volume 34, pp. 17864–17875.
35. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299. [CrossRef]
36. Porter, T.; Duff, T. Compositing Digital Images. In *Proceedings of the ACM SIGGRAPH Computer Graphics*; Association for Computing Machinery: New York, NY, USA, 1984; Volume 18, pp. 253–259. [CrossRef]
37. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361. [CrossRef]
38. 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 (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2446–2454. [CrossRef]
39. Turki, H.; Zhang, J.Y.; Ferroni, F.; Ramanan, D. SUDS: Scalable Urban Dynamic Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12375–12385. [CrossRef]
40. Guo, J.; Deng, N.; Li, X.; Bai, Y.; Shi, B.; Wang, C.; Ding, C.; Wang, D.; Li, Y. StreetSurf: Extending Multi-view Implicit Surface Reconstruction to Street Views. *arXiv* **2023**, arXiv:2306.04988.
41. Wu, Z.; Liu, T.; Luo, L.; Zhong, Z.; Chen, J.; Xiao, H.; Hou, C.; Lou, H.; Chen, Y.; Yang, R.; et al. MARS: An Instance-aware, Modular and Realistic Simulator for Autonomous Driving. In Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI), Fuzhou, China, 22–23 July 2023; pp. 3–15. [CrossRef]
42. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
43. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595. [CrossRef]
44. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
45. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.